UNIVERSITY OF WATERLOO

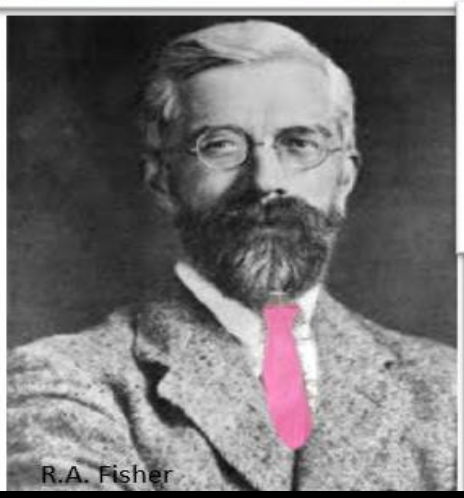
WATERLOO MATHEMATICS

Contact SAS

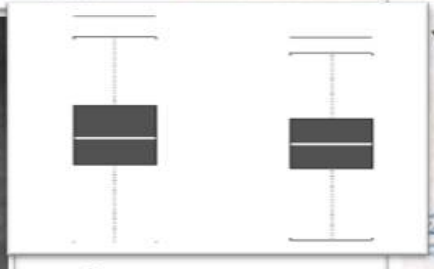
Statistics and Actuarial Science
Mathematics 3 Building
University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada
N2L 3G1
(519) 888-4567 x33550

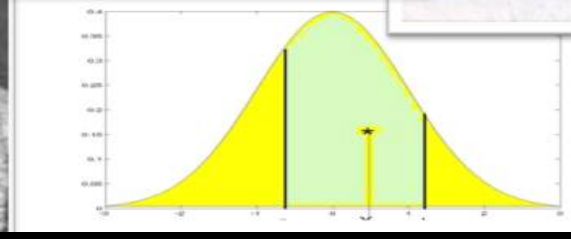
STAT 221/231/241 COURSE NOTES FALL 2016 EDITION

STATISTICS AND ACTUARIAL SCIENCE



R.A. Fisher





Carl Friedrick Gauss

STATISTICS 231 COURSE NOTES

Department of Statistics and Actuarial Science, University of Waterloo

Fall 2016 Edition

Contents

[1] Optional
[2] Optional
[3] Optional

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 Don McLeish, Cyntha Struthers, 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 lists of errors, or sections that are confusing, or additional remarks/suggestions to your instructor or castruth@uwaterloo.ca.

Specific topics in these notes also have associated video files or Powerpoint shows that can be accessed at www.watstat.ca. Where possible we reference these videos in the text.

1. INTRODUCTION TO STATISTICAL SCIENCES

1.1 Statistical Sciences

Statistical Sciences are concerned with all aspects of *empirical studies* including problem formulation, planning of an experiment, data collection, analysis of the data, and the conclusions that can be made. An empirical study is one in which we learn by observation or experiment. A key feature of such studies is that there is usually uncertainty in the conclusions. An important task in empirical studies is to quantify this uncertainty. 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. The uncertainty as to whether a policy holder will have a claim over the next year, or whether the price of a stock will rise or fall, is the basis of financial risk for the insurer and the investor. In medical research, decisions must be made about the safety and efficacy of new treatments for diseases such as cancer and HIV. In developing speech recognition software, computer scientist deal must deal with the uncertainty that arises because spoken language is so complex.

Empirical studies deal with *populations* and *processes*; both of which are collections of individual *units*. In order to increase our knowledge about a population we examine a *sample* of units carefully selected from that population. To study a process of units we examine a sample of units generated by the process. Two challenges arise since we only see a sample from the population or process and not all of the units are the same. 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) because they do not know how the drug will perform on different types of people, what its side effects will be, and so on. For cost and ethical reasons, they can involve only a relatively small sample of subjects in the study. Variability in human populations is ever-present; people have varying degrees of hypertension, they react differently to the drug, they have different side effects. One might similarly want to study variability in currency or stock values, variability in sales for a company over time, or variability in the number of hits and response times for a commercial web site. Statistical Sciences deal both with the study of variability

in processes and populations, and with good (that is, informative, cost-effective) ways to collect and analyze data about such processes.

We can have various objectives when we collect and analyze data from a population or process. In addition to furthering knowledge, these objectives may include decision-making and the improvement of processes or systems. Many problems involve a combination of objectives. For example, government scientists collect data on fish stocks in order to further scientific knowledge and also to provide information to policy makers who must set quotas or limits on commercial fishing.

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 the Statistical Sciences to provide and analyze data so that the maximum amount of information or knowledge may be obtained[4]. Poor or improperly analyzed data may be useless or misleading. The same could be said about poorly collected data.

We use probability models to represent many phenomena, populations, or processes and to deal with problems that involve variability. You studied these models in your first probability course and you have seen how they describe variability. This course will focus on the collection, analysis and interpretation of data and the probability models studied earlier 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. You should review this material.

Statistical Sciences is a large discipline and this course is only an introduction. Our broad objective 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 the 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[5] 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

[4] 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

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

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 complemintary (sic) scarce factor is the ability to understand that data and extract value from it.

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 population of all persons aged 18-25 living in Ontario on January 1, 2016; and the population of all car insurance policies issued by a particular insurance company in the year 2016. A *process* is a system by which units are produced. For example, the hits on a website could be considered as units in a process. Of course this process would be quite complex and difficult to describe. Students taking STAT 231 now and into the future or claims made by car insurance policy holders 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).

We pose questions about populations or processes by defining *variates* for the units which are characteristics of the units. Variates can be of different types. Variates such as height and weight of a person, lifetime of an electrical component, and time until recurrence of disease after medical treatment are all examples of *continuous* or *measured* variates. Variates such as the number of defective smartphones sold by a particular company in a week, the number of deaths in a year on a dangerous highway or the number of damaged pixels in a monitor are all examples of *discrete* variates.

Variates such as hair colour, university program or marital status 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. 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. In such a case 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 variate classifies a unit by size, for example, large, medium or small, then the variate is a categorical variate. However since this categorical variate has a natural ordering, it is also called an *ordinal* variate. Another example of an ordinal variate would be opinion on a given statement for which the categories might be: strongly agree, agree, neutral, disagree, strongly disagree.

Variates can also be *complex* such as an image or an open ended response to a survey

question.

We are interested in functions of the variates over the whole population; for example the *average* drop in blood pressure due to a treatment for individuals with hypertension or the *proportion* of a population having a certain characteristic. We call these functions *attributes* of the population or process.

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 time between claims. The values of y 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.

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 12 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) **Experiments or Experimental Studies.** An experiment is a study 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[6]. (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

[6] One of the most important studies was conducted in the Waterloo school board; see for example "Six-year follow-up of the first Waterloo school smoking prevention trial" at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1350177/

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

With products or services such as credit cards it is common to conduct direct marketing campaigns in which large numbers of individuals are contacted by mail and invited to acquire a product or service. Such individuals are usually picked from a much larger number of persons on whom the company has information. For example, in a credit card marketing campaign a company might have data on several million persons, pertaining to demographic (e.g. sex, age, place of residence), financial (e.g. salary, other credit cards held, spending patterns) and other variates. Based on the data, the company wishes to select persons whom it considers have a good chance of responding positively to the mail-out. The challenge is to use data from previous mail campaigns, along with the current data, to achieve as high a response rate as possible.

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: graphical and numerical.

The basic set-up is as follows. 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 real 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 when the variate is either continuous or discrete. 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).

Measures of location:

- The *(sample) mean* also called the sample average: $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.
- 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.

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

- 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.

Measures of dispersion or variability:

- The *(sample) variance*:

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

  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 which is described below.

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.

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.

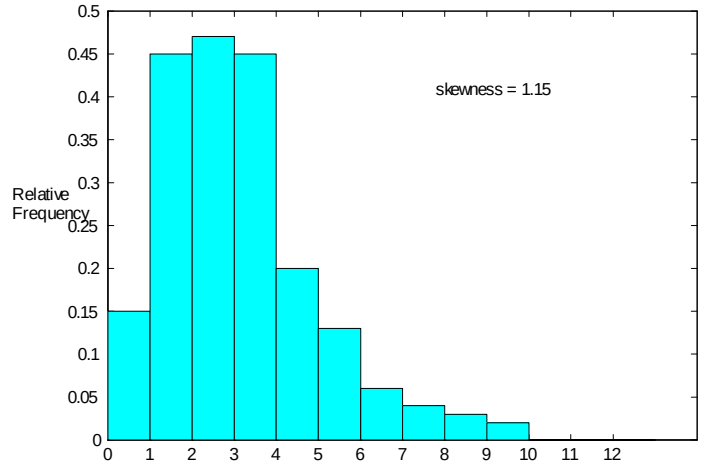


Figure 1.1: **Relative frequency histogram for data with positive skewness**

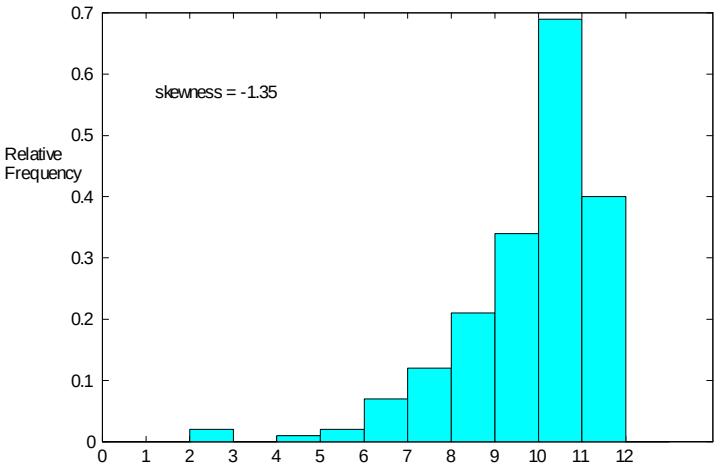


Figure 1.2: **Relative frequency histogram for data with negative skewness**

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 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 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.

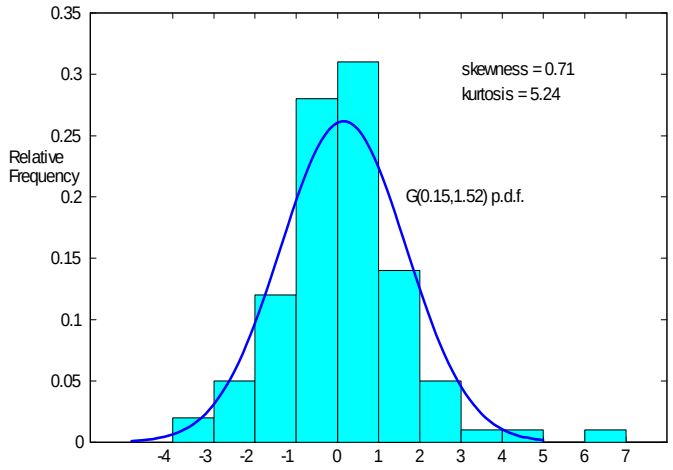


Figure 1.3: **Relative frequency histogram for data with kurtosis > 3**

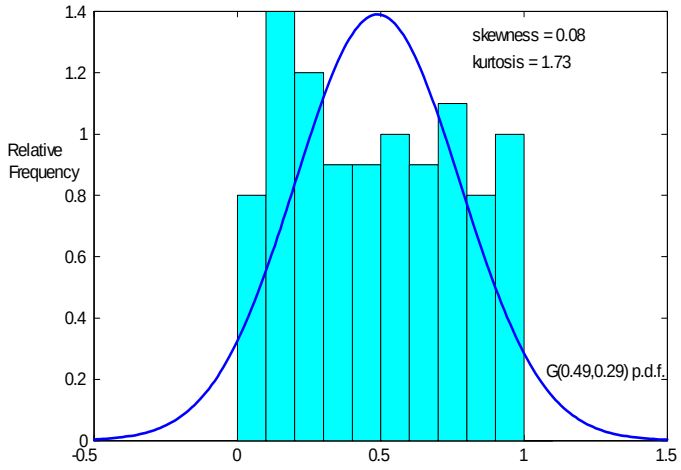


Figure 1.4: **Relative frequency histogram for data with kurtosis < 3**

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 heaver tails (and a more peaked center) than data that are Normally distributed. 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 roughly $1-p$ are greater. 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 determined quantiles.

Definition 1 *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* $\{y_1, y_2, \ldots, y_n\}$. *The pth (sample) quantile (also called the* $100p$*th (sample) percentile) is a value, call it* $q(p)$*, determined as follows:*

- *Let* $m = (n+1)p$ *where* n *is the sample size.*
- *If* m *is an integer between* 1 *and* m *then* $q(p) = y_{(m)}$ *which is the* m*'th largest value in the data set.*
- *If* m *is not an integer but* $1 < m < n$ *then determine the closest integer* j *such that* $j < m < j+1$ *and take* $q(p) = \frac{1}{2}\left[y_{(j)} + y_{(j+1)}\right]$.

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 2 *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 $p = 0.25$, $m = (12+1)(0.25) = 3.25$ so

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

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

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

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

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

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

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

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

Since the interquartile range is less affected by a few extreme observations (see Problem 2) it is a more robust measure of variability as compared to the sample standard deviation.

Definition 4 *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:* $y_{(1)}$, $q(0.25)$, $q(0.5)$, $q(0.75)$, $y_{(n)}$.

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).

Example 1.3.2 Comparison of Body Mass Index

In a longitudinal study (that is, the people in the sample were followed over time) of obesity in New Zealand, a sample of 150 men and 150 women were selected from workers aged 18 to 60. Many variates were measured for each subject (unit), including their height (m) and weight (kg) at the start of the study. These variates are both continuous variates. Their initial Body Mass Index (BMI) was also calculated. BMI is used to measure obesity or severely low weight. It is defined as:

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

There is some variation in what different guidelines refer to as "overweight", "underweight", etc. We present one such classification in Table 1.1. The BMI obesity classification is an example of a ordinal variate.

Table 1.1: BMI Obesity Classification

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

The data are available in the file *bmidata.txt* posted on the course website. For statistical analysis of 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, sex (M = male or F = female), height, weight and BMI. Each subsequent row gives the variate values for a particular subject.

Table 1.2: First Five Rows of the File bmidata.txt

subject	sex	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

The five number summaries for the BMI data for each sex are given in Table 1.3 along with the sample mean and standard deviation.

Table 1.3: Summary of BMI by Sex

Sex	$y_{(1)}$	$q\,(0.25)$	$q\,(0.5)$	$q\,(0.75)$	$y_{(n)}$	$\bar{y}$	s
Female	16.4	23.4	26.8	29.75	38.8	26.92	4.60
Male	18.3	24.6	26.75	29.15	37.5	27.08	3.56

From the table, we see that there are only small differences in the median and the mean. For the 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.

We can also construct a *relative frequency table* that gives the proportion of subjects that fall within each obesity class by sex.

Table 1.4: BMI Relative Frequency Table by Sex

Obesity Classification	Males	Females
Underweight	0.01	0.02
Normal	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

From Table 1.4, we see that the reason for the larger variability for females is that there is a greater proportion of females in the extreme classes.

Sample Correlation

So far we have looked only at graphical 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 5 *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

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2, \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\ \text{and } S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2. \end{aligned}$$

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 $r = 0.55$ and for the females $r = 0.31$ which indicates that there is a positive relationship between height and weight which is exactly what we would expect.

Relative Risk

Recall that categorical variates consist of group or category names that do not necessarily have any ordering. If two variates of interest in a study are categorical variates then it does not make sense to use sample correlation 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). One set of data collected in the study are given in Table 1.5.

Table 1.5: Physicians' Health Study

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

What measure can be used to summarize the relationship between taking daily aspirin and the occurrence of CHD?

One measure which is used to summarize the relationship between two categorical variates is *relative risk*. To define relative risk consider a generalized version of Table 1.5 given by

Table 1.6: General Two-way Table

	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

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 PHS if we let $A =$ takes daily aspirin and $B =$ CHD then we can estimate this ratio using the ratio of the sample proportions.

Definition 6 *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 PHS the relative risk of CHD in the placebo group as compared to the aspirin group is

$$\text{relative risk} = \frac{189/(189+10845)}{104/(104+10933)} = 1.82.$$

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

We consider several types of plots for a data set $\{y_1, y_2, \ldots, y_n\}$ and one type of plot for a data set $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$.

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 be always be 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. Here the *shapes* of the two distributions are somewhat bell-shaped. In each case the skewness is positive but close to zero while the kurtosis is close to three.

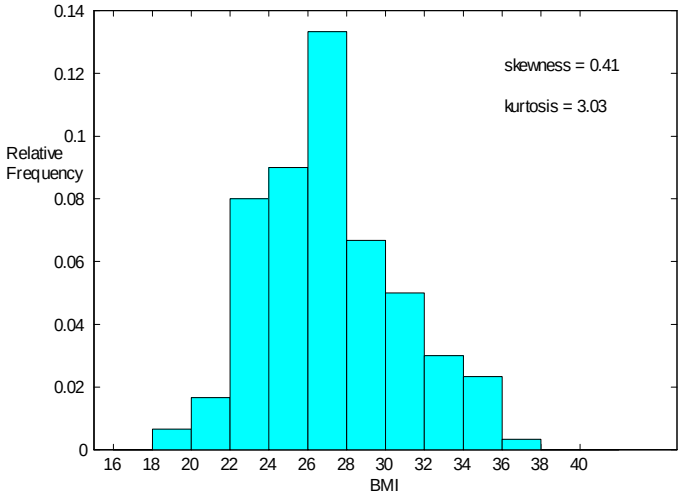


Figure 1.5: **Relative frequency histogram for male BMI data**

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 data are available in the file *brakepaddata.txt* posted on the course website. Notice that the distribution of the brake pad lifetimes has a very different shape compared to the BMI histograms. The shape does not resemble a bell-shaped curve. The distribution is

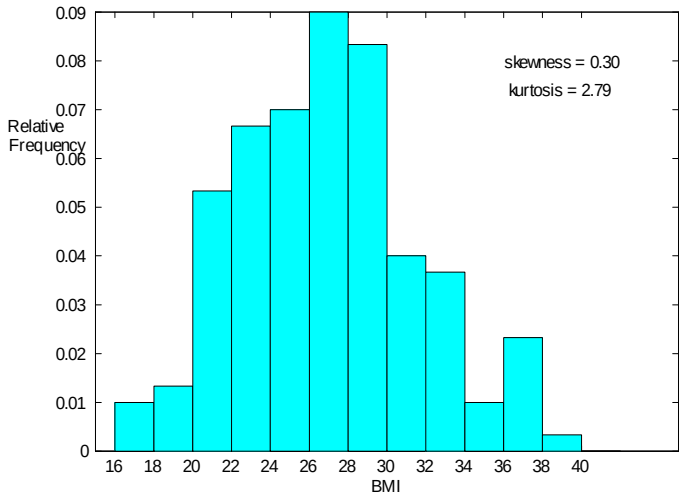


Figure 1.6: **Relative frequency histogram for female BMI data**

not symmetric and has a long right tail which is consistent with a skewness value equal to 1.28 which is positive and not close to zero. The sample mean is $\bar{y} = 49.03$ thousand km and the sample variance 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.

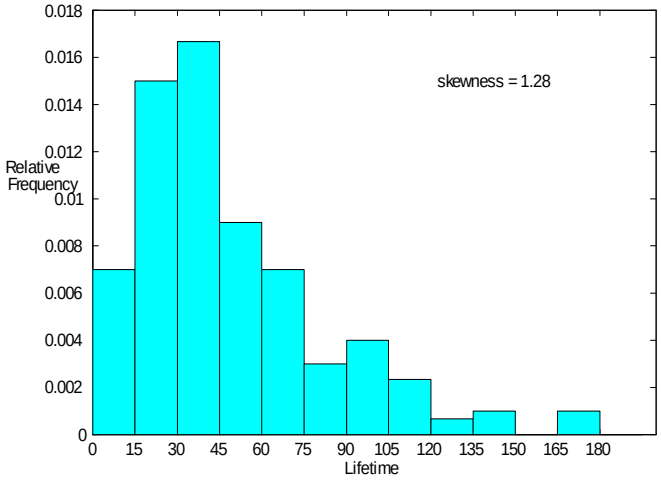


Figure 1.7: **Relative frequency histogram of brake pad lifetime data**

Empirical Cumulative Distribution Functions

Another way to portray the values of a variate $\{y_1, y_2, \ldots, y_n\}$ is to determine the proportion of values in the set which are smaller than any given value. This is called the *empirical cumulative distribution function* or *e.c.d.f.* and is defined by

$$\hat{F}(y) = \frac{\text{number of values in } \{y_1, y_2, \ldots, y_n\} \text{ which are} \leq y}{n}.$$

To construct $\hat{F}(y)$, it is convenient to first order the y_i's $(i = 1, 2, \ldots, n)$ to give the ordered values $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$. Then, we note that $\hat{F}(y)$ is a step function with a jump at each of the ordered observed values $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$. More details on constructing the empirical cumulative distribution function and its close relative the *qqplot* are provided in Section 2.4 but for the moment consider the empirical cumulative distribution function as an estimate, based on the data, of the population cumulative distribution function.

Example 1.3.2 Continued

Figure 1.8 shows the empirical cumulative distribution function for male and female heights on the same plot. The plot of the empirical cumulative distribution function does not show the shape of the distribution as clearly as a histogram does. However, it does show the proportion of y-values in any given interval; the proportion in the interval $(a, b]$ is just $\hat{F}(b) - \hat{F}(a)$. In addition, this plot allows us to determine the pth quantile or $100p$th percentile (the left-most value on the horizontal axis y_p where $\hat{F}(y_p) = p$), and in particular the median (the left-most value $\hat{m}$ on the horizontal axis where $\hat{F}(\hat{m}) = 0.5$). For example, we see from Figure 1.8 that the median height for females is about 1.60m and for males the median height is about 1.73m.

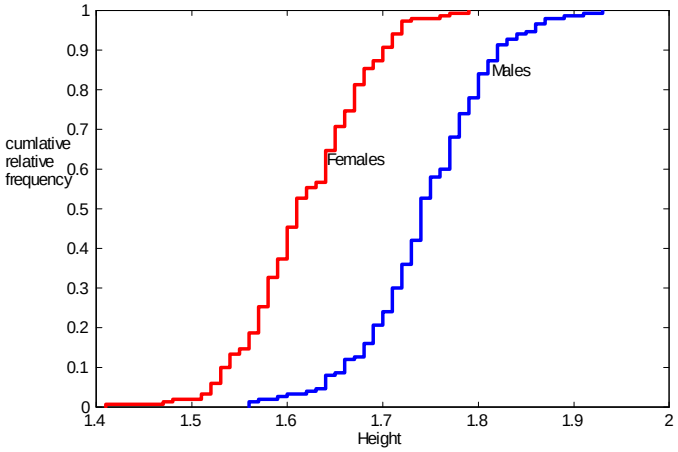


Figure 1.8: **Empirical cumulative distribution function of heights for males and for females**

Boxplots

In many situations, we want to compare the values of a variate for two or more groups, as in Example 1.3.2 where we compared BMI values and heights for males versus females. When the number of groups is large or the sample sizes within groups are small, side-by-side *boxplots* are a convenient way to display the data. Boxplots are also called *box and whisker plots*.

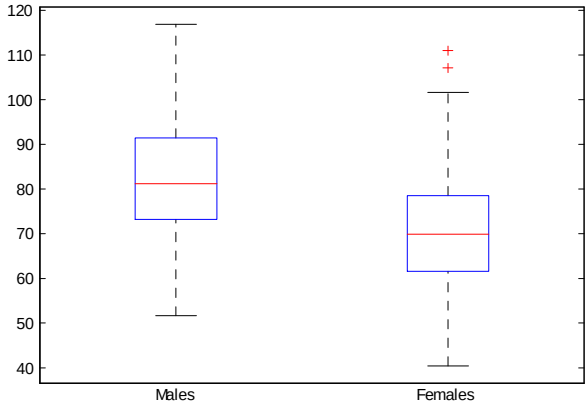


Figure 1.9: **Boxplots of weights for males and females**

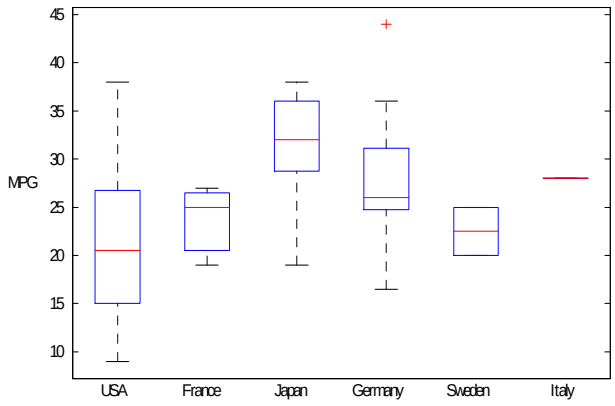


Figure 1.10: **Boxplots for miles per gallon for 100 cars from six different countries**

The boxplot is usually displayed vertically. The line inside the box corresponds to the median and the bottom and top sides of the box correspond to the lower quartile $q(0.25)$ and the upper quartile $q(0.75)$ respectively. 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 1.9 displays side-by-side boxplots of male and female weights from Example 1.3.2. We can see that males are generally heavier than females but that the spread of the two distributions is about the same. For the males and the females, the center line in the box, which corresponds to the median, divides the box and whiskers approximately in half which indicates that both distributions are roughly symmetric about the median. For the females there are two large outliers.

Boxplots are particularly usually for comparing several groups. Figure 1.10 shows a comparison of the miles per gallon (MPG) for 100 cars by country of origin. The boxplot makes it easy to see the differences and similarities between the cars from different countries.

The graphical summaries we have just discussed are most useful for summarizing variates which are either continuous or discrete with many possible values. For categorical variates the data can be best summarized using bar graphs and pie charts. Such graphs are often used incorrectly. See Problems 17-20.

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 then 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 the two variates are related in some way.

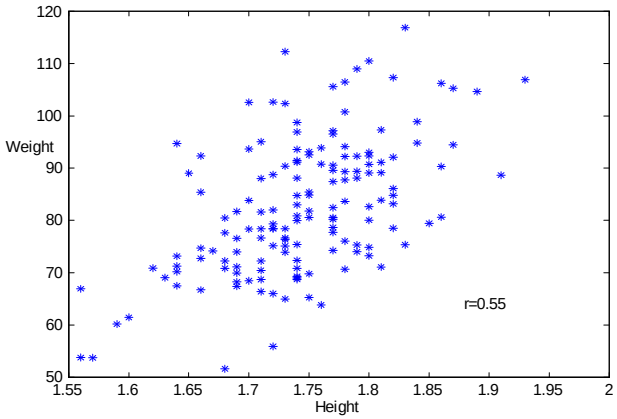


Figure 1.11: **Scatterplot of weight versus height for males**

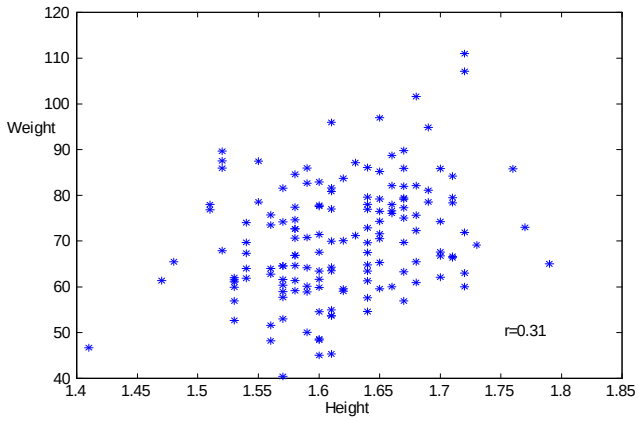


Figure 1.12: **Scatterplot of weight versus height for females**

Figures 1.11 and 1.12 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.

1.4 Probability Distributions and Statistical Models

Statistical models are used 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. With 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 (that is, the population histogram) 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

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 (teenagers living in Ontario), suppose we had a list of most of the units in the population. 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.)

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

The parameter θ represents the unknown proportion of smokers in the population, 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

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

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 as in the above brake pad example, we can draw strong parallels between summaries of the data set described in Section 1.3 and properties of the corresponding probability model Y. For example,

- The sample mean $\bar{y}$ corresponds to the population mean $E(Y) = \mu$.
- 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 sample standard deviation s corresponds to σ, the population standard deviation of Y, where $\sigma^2 = E[(Y - \mu)^2]$.
- 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 1.4.3 A Gaussian Distribution Example

Earlier, we described an experiment where the goal was to see if there is a relationship 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 $Y_i \backsim G(\beta_0 + \beta_1 x_i, \sigma)$, $i = 1, 2, \ldots, 40$ 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 types: 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 partially explain or determine the distribution of the response variate.

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, the response variable Y might be the weight (in kg) of a randomly selected female in the age range 16-25, in some population. A person's weight is related to their height. We might want to study this relationship by considering females with a given height x (say in meters), and proposing that the distribution of Y, given x is Gaussian, $G(\alpha + \beta x, \sigma)$. That is, we propose that the average (expected) weight of a female depends linearly on her height x and we write this as $E(Y|x) = \alpha + \beta x$. It would be possible to reverse the roles of the two variates and consider weight to be the explanatory variate and height to be the response variate, if for example we wished to predict height using data on individuals' weights.

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 problems will be discussed, loosely referred to as *estimation problems, hypothesis testing problems* and *prediction problems*. In the first type, the problem is to estimate one or more attributes of a process or population. 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.

Hypothesis testing problems involve using the data to assess the truth of some question or hypothesis. 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.

In prediction problems, we use the data to predict a future value for a process variate or a unit to be selected from the 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 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 A 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 teenage girls higher than the rate among teenage boys?" From the data, we see that the sample proportion of girls 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 teenage girls and boys.

Example 1.5.2 A 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 data are available in the file *canfillingdata.txt* posted on the course website.

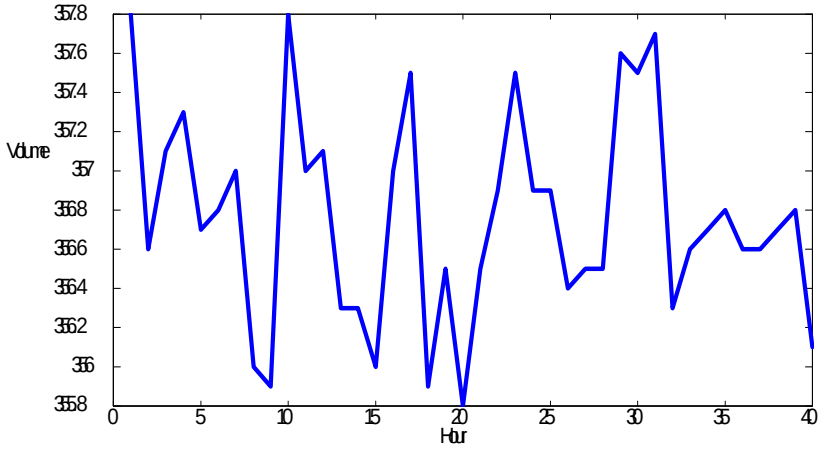


Figure 1.13: **Run chart of the volume for the new machine over time**

First we examine if the behaviour of the two machines is stable over time. In Figures 1.13 and 1.14, a *run chart* of the volumes over time for each machine is given. There is no 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.

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 1.15 and 1.16 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 reasonable given that the sample size for both data sets is only forty. Note that it only makes sense to compare density functions and relative frequency histograms (not standard) since the areas both equal one.

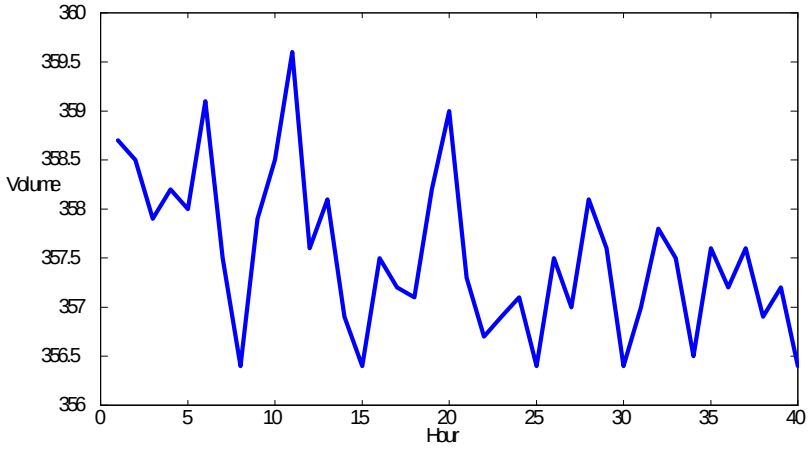


Figure 1.14: **Run chart of the volume for old machine over time**

None of the 80 cans had volume less than the required 355ml. However, we examined only 40 cans per machine. We can use the Gaussian model to estimate the long term proportion of cans that fall below the required volume. For the new machine, we find that if $Y \sim G(356.8, 0.54)$ then $P(Y \leq 355) = 0.0005$ so about 5 in $10,000$ cans will be underfilled. The corresponding rate for the old machine is about 8 in $10,000$ cans. Of course these estimates are subject to a high degree of uncertainty because they are based on a small sample and we have no way to test that the models are appropriate so far into the tails of the distribution.

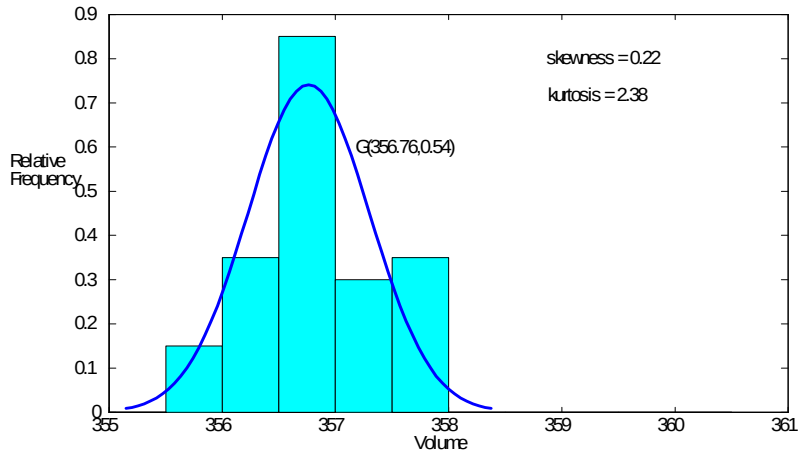


Figure 1.15: **Relative frequency histogram of volumes for the new machine**

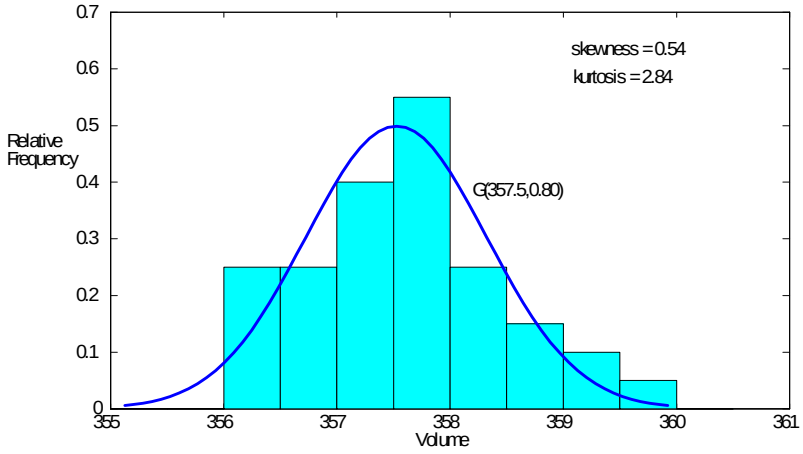


Figure 1.16: **Relative frequency histogram of volumes for the old machine**

We can also 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. It is an open source package that has extensive statistical capabilities and very good graphics procedures. 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 and the sample median are two different ways to measure the location of a data set $\{y_1, y_2, \ldots, y_n\}$. Let $\bar{y}$ be the sample mean and $\hat{m}$ be the sample median of the data set.

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

   (b) Suppose we transform the data by squaring so that $v_i = {y_i}^2$, $i = 1, 2, \ldots, n$. How are the sample mean and sample median of $v_1, v_2, \ldots, v_n$ related to $\bar{y}$ and $\hat{m}$?

   (c) Consider the quantities $r_i = y_i - \bar{y}$, $i = 1, 2, \ldots, n$. Show that $\sum_{i=1}^{n} r_i = 0$.

   (d) Suppose we include an extra observation y_0 to the data set and define $a(y_0)$ to be the mean of the augmented data set. Express $a(y_0)$ in terms of $\bar{y}$ and y_0. What happens to the sample mean as y_0 gets large (or small)?

   (e) Repeat the previous question for the sample median. Hint: Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ be the original data set with the observations arranged in increasing order.

   (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) Show that $W(\mu) = \sum_{i=1}^{n} |y_i - \mu|$ is minimized when $\mu = \hat{m}$. Hint: Calculate the derivative of $W(\mu)$ when $\mu < y(1)$, $y(1) < \mu < y(2)$ and so on. The minimum occurs where the derivative changes sign.

2. The sample standard deviation and the interquartile range are two different measures of the variability of a data set $(y_1, y_2, \ldots, y_n)$. Let s be the sample standard deviation and let IQR be the interquartile range of the data set.

   (a) 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 standard deviation and interquartile range of $u_1, u_2, \ldots, u_n$ related to s and IQR?

   (b) Show that $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\,(\bar{y})^2$.

   (c) Suppose we include an extra observation y_0 to the data set. Use the result in (b) to write the sample standard deviation of the augmented data set in terms of y_0 and the original sample standard deviation. What happens when y_0 gets large (or small)?

   (d) How does the IQR change as y_0 gets large?

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 $u_1, u_2, \ldots, u_n$ 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 2013 to December 2014. For this data set the sample mean was \$2500, the sample 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 ($10^{-6}m$). 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 in folder S231Datasets 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 boxplot and the empirical cumulative distribution function 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. Run the following code on the can filling data and compare with the summaries given in Example 1.5.2.

```
#Import dataset canfillingdata.txt in folder S231Datasets using RStudio
attach(canfillingdata)
#Separate the volumes by machine into separate vectors v1 and v2
v1<-volume[seq(1,79,2)]   # Puts New Machine values in vector v1
v2<-volume[seq(2,80,2)]   # Puts Old Machine values in vector v2
#
#Calculate summary statistics by machine
#Install moments package for skewness and kurtosis
install.packages("moments")
library(moments)
c(mean(v1),sd(v1),skewness(v1),kurtosis(v1))
fivenum(v1)  # Gives the 5 number summary
#R defines the 1st and 3rd quartiles slightly different than Def'n 1
c(mean(v2),sd(v2),skewness(v2),kurtosis(v2))
fivenum(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 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 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")
```

9. The data below show the lengths (in cm) of 43 male coyotes and 40 female coyotes captured in Nova Scotia. (Based on Table 2.3.2 in Wild and Seber 1999.) The data are available in the file *coyotedata.txt* posted on the course website.

Females x

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								

$$\sum_{i=1}^{40} x_i = 3569.6 \quad \sum_{i=1}^{40} x_i^2 = 320223.38$$

Males y

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					

$$\sum_{i=1}^{43} y_i = 3958.4 \quad \sum_{i=1}^{43} y_i^2 = 366276.84$$

(a) Plot relative frequency histograms of the lengths for females and males separately. Be sure to use the same intervals.

(b) Determine the five number summary for each data set.

(c) Plot side by side boxplots for the females and males. What do you notice?

(d) Compute the sample mean and sample standard deviation for the lengths of the female and male coyotes separately. Assuming μ = sample mean and σ = sample standard deviation, overlay the corresponding Gaussian probability density function on the histograms for the females and males separately. Comment on how well the Gaussian model fits each data set.

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

10. 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 20 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

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

(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 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 in folder S231Datasets using RStudio
attach(actordata)
cor(Value,Gross)   # Calculates sample correlation
plot(Value,Gross,main ="Actor Data",col="blue")  # scatterplot
# 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
```

11. 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.

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

12. 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) What is the probability function for Y? What assumptions have you made?

    (b) 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.
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 in a sample of size n.
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) If in a sample of n people the observed data were y_1, y_2, y_3, y_4 what would be reasonable estimates of θ_1, θ_2, θ_3, θ_4?

13. The IQ's of students of UWaterloo Math students have a Gaussian distribution with mean μ and standard standard deviation σ. Define the random variable Y = IQ of UWaterloo Math student.

(a) What is the probability density function of Y?

(b) What are $E(Y)$ and $Var(Y)$?

(c) Suppose that the IQ's for 16 students were:

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

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

What is a reasonable estimate of μ based on these data? What is a reasonable estimate of σ^2 based on these data? Estimate the probability that a random chosen UWaterloo Math student will have an IQ greater than 120.

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

(i) What is the distribution of

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

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

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

(iii) If $\sigma = 12$, find the smallest value of n such that $P\left(\left|\bar{Y} - \mu\right| \le 1.0\right) \ge 0.95$.

14. The lifetimes of a certain type of battery are Exponentially distributed with parameter θ. Define the random variable Y = lifetime of a battery.

    (a) What is the probability density function of Y?

    (b) What are $E(Y)$ and $Var(Y)$?

    (c) Suppose the lifetimes (in hours) for 20 batteries were:

    $$\begin{array}{cccccccccc} 20.5 & 9.9 & 206.4 & 9.1 & 45.8 & 232.7 & 127.8 & 60.4 & 4.3 & 3.6 \\ 184.8 & 3.0 & 4.4 & 72.3 & 22.3 & 195.3 & 86.3 & 8.8 & 23.3 & 4.1 \end{array} \qquad \sum_{i=1}^{20} y_i = 1325.1$$

    What is a reasonable estimate of θ based on these data? Estimate $P(Y > 100)$ using these data.

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

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

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

15. Accidents occur on Wednesday's at a particular intersection at random at the average rate of θ accidents per Wednesday according to a Poisson process. Define the random variable Y = number of accidents on Wednesday at this intersection.

    (a) What is the probability function for Y? How well do you think the assumptions of a Poisson process might hold in this case?

    (b) What are $E(Y)$ and $Var(Y)$?

    (c) Suppose on 6 consecutive Wednesday's the number of accidents observed was 0, 2, 0, 1, 3, 1. What is the probability of observing these data as a function of θ? What is a reasonable estimate of θ based on these data? Estimate the probability that there is at least one accident at this intersection next Wednesday.

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

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

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

16. 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 $E(\bar{Y})$, $Var(\bar{Y})$ and $E\left[(\bar{Y})^2\right]$.

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

    $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Y_i^2 - n\left(\bar{Y}\right)^2\right].$$

17. The pie chart in Figure 1.17, from Fox News, shows the support for various Republican Presidential candidates in 2012. What do you notice about this pie chart? Comment on how effective pie charts are in general at conveying information.

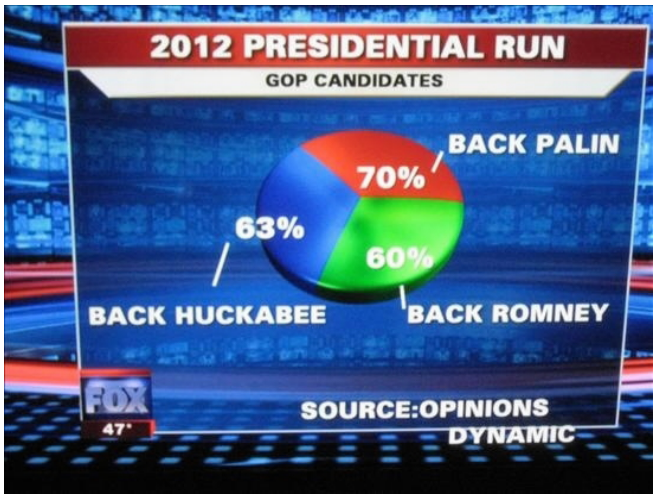


Figure 1.17: **Pie chart for support for Republican Presidental candidates**

18. For the graph in Figure 1.18 indicate whether you believe the graph is effective in conveying information by giving at least one feature of the graph which is either good or bad.

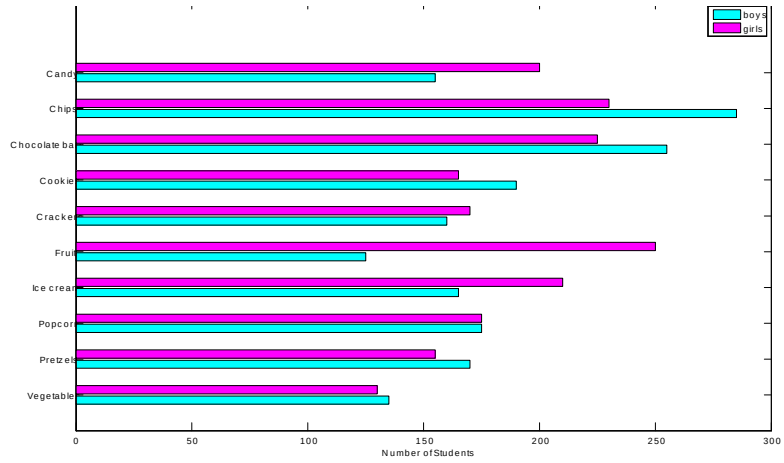


Figure 1.18: **Preferred snack choices of students at Ridgemont High School**

19. The graphs in Figures 1.19 and 1.20 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?

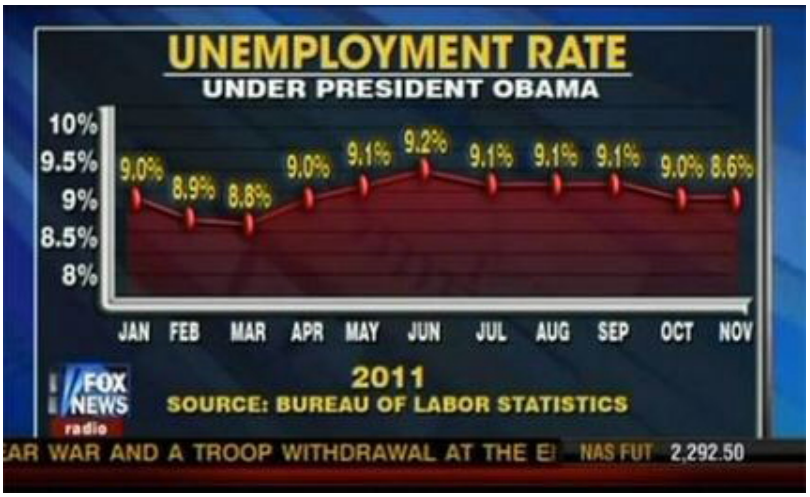


Figure 1.19: **Unemployment Rate under President Obama**

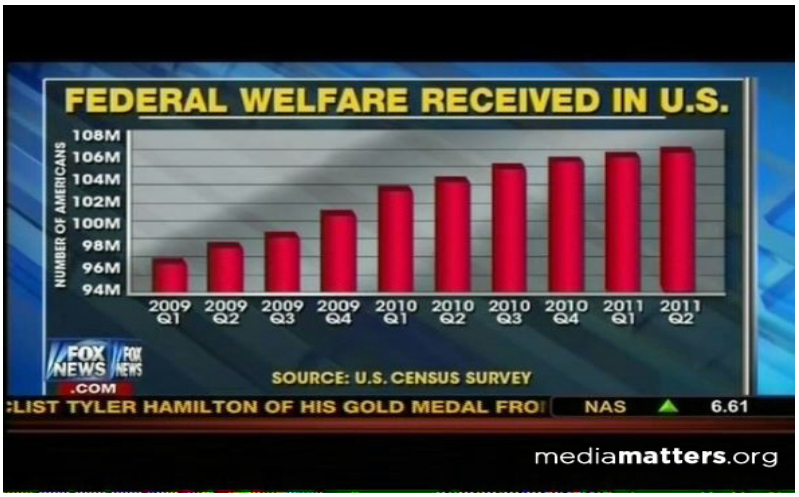


Figure 1.20: **Federal Welfare in the US**

20. Information about the mortality from malignant neoplasms (cancer) for females living in Ontario is given in figures 1.21 and 1.22 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 1.23. Which display seems to carry the most information?

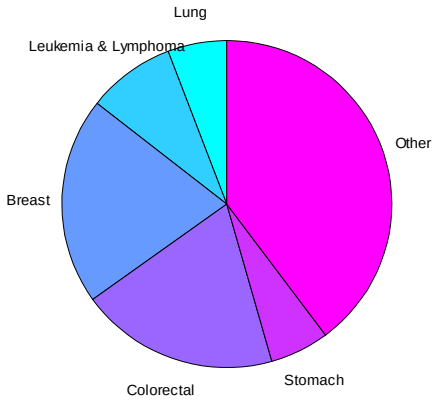


Figure 1.21: **Mortality from malignant neoplasms for females in Ontario 1970**

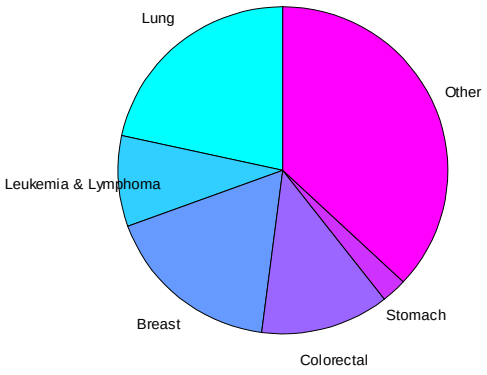


Figure 1.22: **Mortality from malignant neoplasms for females in Ontario in 2000**

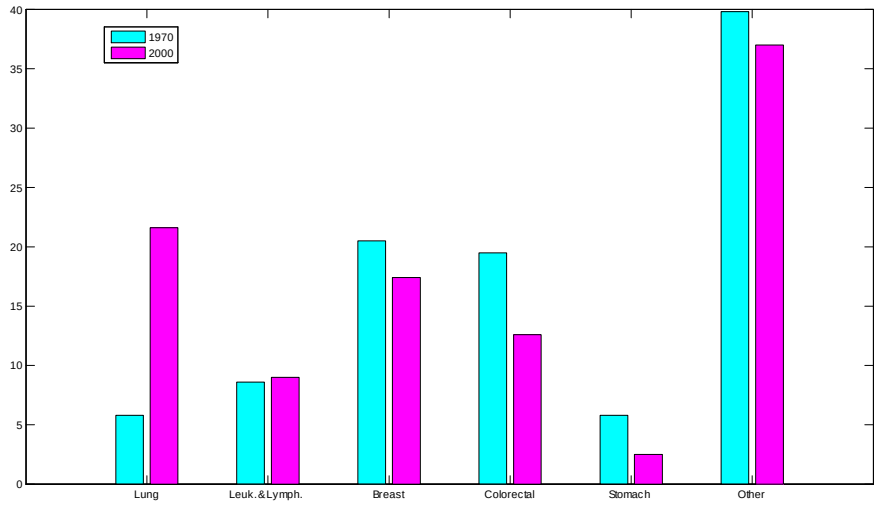


Figure 1.23: **Mortality from malignant neoplasms for females living in Ontario, 1970 and 2000**

2. STATISTICAL MODELS AND MAXIMUM LIKELIHOOD ESTIMATION

2.1 Choosing a Statistical Model

A statistical model is a mathematical model that incorporates probability[7] 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 considering random variables that represent characteristics of randomly selected units or individuals in the population or process, and by studying the probability distributions of these random variables. It is very important to be clear about what the "target" population or process is, and exactly how the variables being considered are defined and measured. These issues are discussed in Chapter 3.

A preliminary step in probability and statistics is the choice of a statistical model[8] to suit a given application. The choice of a model is usually driven 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.

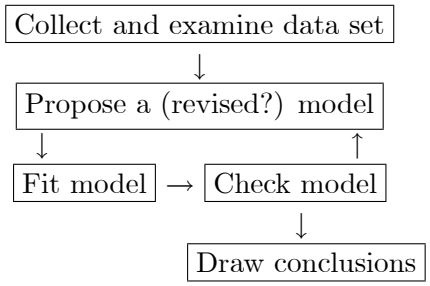
[7] The material in this section is largely a review of material you have seen in a previous probability course. This material is available in the STAT 230 Notes which are posted on the course website.

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

In probability theory, 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 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 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
c.d.f.	$F(x) = P(X \leq x) = \sum_{t \leq x} P(X = t)$ 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 $x \in \Re$
p.f./p.d.f.	$f(x) = P(X = x)$	$f(x) = \frac{d}{dx}F(x) \neq P(X = x) = 0$
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 $\theta > 0$. 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

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

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 $0 < \theta_i < 1$ 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 Estimation of Parameters and the Method of Maximum Likelihood

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,

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

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 7 *An 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 8 *The* likelihood function *for θ is defined as*

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

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 9 *The value of θ which maximizes $L(\theta)$ for given data $\mathbf{y}$ is called the* maximum likelihood estimate [9] *(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 Harris/Decima public opinion poll[10]

The article on the next page which was published in the CAUT (Canadian Association of University Teachers) Bulletin describes a poll conducted by the Harris/Decima company. Harris/Decima conducts polls for CAUT to learn about Canadian public opinion about post-secondary education in Canada.

[9] We will often distinguish between the random variable, the *maximum likelihood estimator*, which is the function of the data in general, and its numerical value for the data at hand, referred to as the *maximum likelihood estimate*.

[10] See the corresponding video "Harris/Decima poll and introduction to likelihoods" at www.watstat.ca

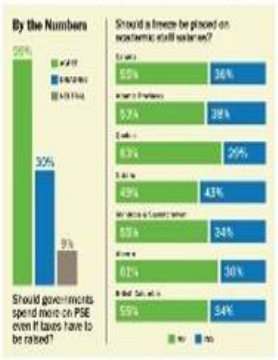
CAUT ACPPU BULLETIN

CAUT | AcademicWork.ca | Share

CANADA'S VOICE FOR ACADEMICS
Vol 59 | No 6 | June 2012

Harris / Decima Poll: Fund PSE Even If It Means Tax Hikes

BACK PRINT



A nationwide poll revealed last month that almost six out of 10 Canadians said spending on post-secondary education should be increased even if it means paying higher taxes.

The semi-annual CAUT survey by Harris/Decima also found a majority of Canadians believe the quality of post-secondary education is suffering because of underfunding.

Canadians are also concerned governments aren't doing enough to ensure access, with nearly 50 per cent of respondents saying the most important thing governments can do for post-secondary education is to lower tuition fees.

An equal number don't believe university and college teachers earn too much, with 55 per cent of those surveyed saying they were opposed to a freeze on academic staff compensation; the same number think freezing salaries would compromise the educational environment.

"As the poll demonstrated, Canadians are concerned about the quality of post-secondary education and want government to do more to improve access," said CAUT executive director James Turk.

"The majority also understand that imposing wage freezes on academic staff, as some provinces are trying to do, will just further erode quality," he added.

The Harris/Decima poll is based on a telephone survey of 2,000 randomly selected Canadian adults conducted between Nov. 11 and 21, 2010. The margin of error — which measures sampling variability — is ± 2.2 per cent, 19 times out of 20.

Home
Bookshelf
Commentary
News
President's Column
Classifieds
Careers
Subscriptions
Archives
Advertising
Masthead
Search
Français

The poll described in the article was conducted in November 2010. Harris/Decima uses a telephone poll of 2000 "representative" adults. Figure 2.1 shows the results for the polls conducted in fall 2009 and 2010. In 2009 and 2010, 26% of respondents agreed and 48% disagreed with the statement: "University and college teachers earn too much".Harris/Decima

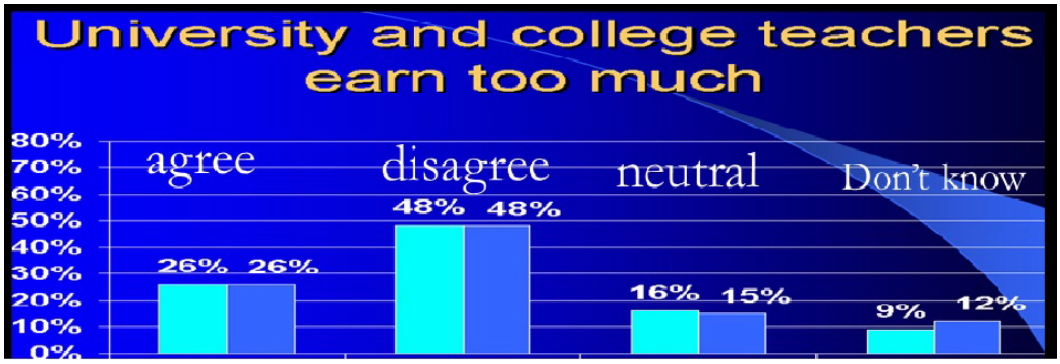


Figure 2.1: **Harris/Decima poll. The two bars are from polls conducted in Nov. 9, 2009 (left bar) and Nov 10, 2010 (right bar)**

declared their result to be accurate within ±2.2%, 19 times out of 20 (the margin of error for regional, demographic or other subgroups is larger). What does this mean and how were these estimates and intervals obtained?

Suppose that the random variable Y represents the number of individuals who, in a randomly selected group of n persons, agreed with the statement. Suppose we assume that Y is closely modelled by a Binomial distribution with probability function

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

where θ represents the proportion of the Canadian adult population that agree with the statement. If n people are selected and y people agree with the statement then the likelihood function is given by

$$\begin{aligned} L(\theta) &= P\left(y \text{ people agree with the statement } ; \theta\right) \\ &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } 0 < \theta < 1. \end{aligned} \tag{2.3}$$

It is easy to see that (2.3) is maximized by the value $\theta = \hat{\theta} = y/n$. (You should show this.) The estimate $\hat{\theta} = y/n$ is called the sample proportion. For the Harris/Decima poll conducted in 2010, $y = 520$ people out of $n = 2000$ people agreed with the statement so the likelihood function is

$$L\left(\theta\right) = \binom{2000}{520} \theta^{520} (1-\theta)^{1480} \quad \text{for } 0 < \theta < 1 \tag{2.4}$$

and the maximum likelihood estimate is $520/2000 = 0.26$ or 26%. This is also easily seen from a graph of the likelihood function (2.4) given in Figure 2.2.

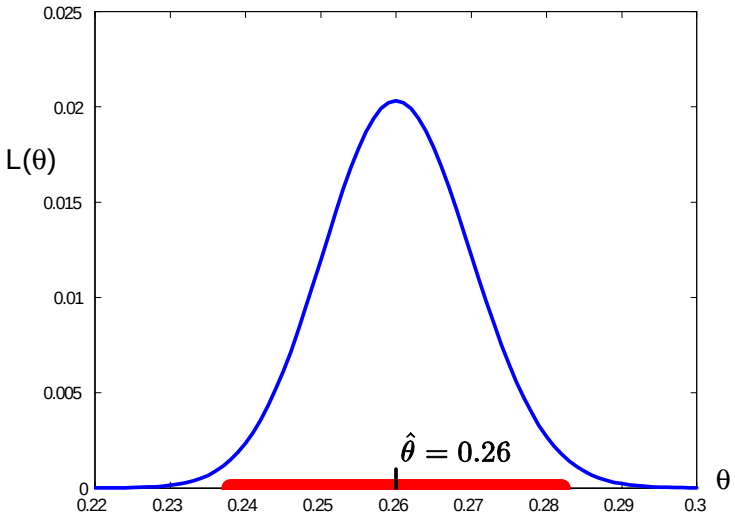


Figure 2.2: **Likelihood function for the Harris/Decima poll and corresponding interval estimate for θ**

The interval suggested by the pollsters was $26 \pm 2.2\%$ or $[23.8,\ 28.2]$. Looking at Figure 2.2 we see that the interval $[0.238, 0.282]$ 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.

Note that 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 $\hat{\theta} = y/n$ 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 10 *The **relative likelihood function** is defined as*

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

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

Sometimes it is easier to work with the *log* (*log* = *ln* in these course notes) of the likelihood function.

Definition 11 *The **log likelihood function** is defined as*

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

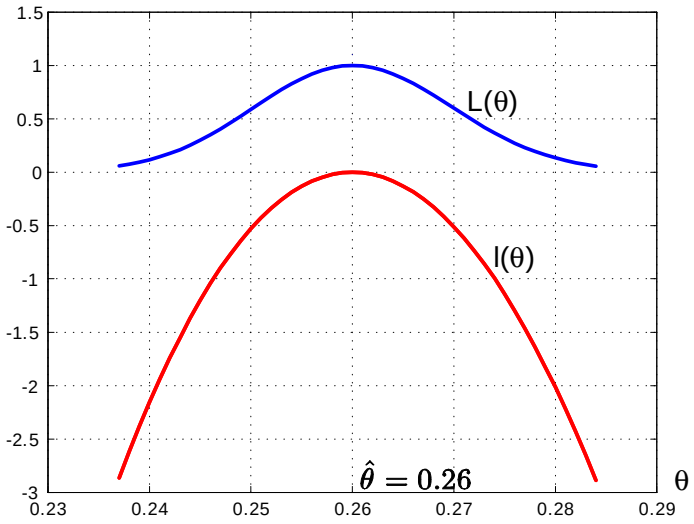


Figure 2.3: **The functions $L(\theta)$ (upper graph) and $l(\theta)$ (lower graph) are both maximized at the same value $\theta = \hat{\theta}$**

Note that $\hat{\theta}$ also maximizes $l(\theta)$. In fact in Figure 2.3 we see that $l(\theta)$, the lower of the two curves, is a monotone function of $L(\theta)$ so they increase together and decrease together. This implies that 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[11], 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

$$\frac{d}{d\theta} l(\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta}.$$

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 $\hat{\theta} = y/n$.

[11] Can you think of an example of a continuous function $f(x)$ defined on the interval $[0,1]$ for which the maximum $\max_{0 \le x \le 1} f(x)$ is NOT found by solving $f'(x) = 0$?

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

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

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

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

or more simply

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

The log likelihood is

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

with derivative

$$\frac{d}{d\theta} l\left(\theta\right) = n\left(\frac{\bar{y}}{\theta} - 1\right) = \frac{n}{\theta}\left(\bar{y} - \theta\right).$$

A first derivative test easily verifies 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

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

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 2.2.1 Continued

Harris/Decima also conducted a poll for CAUT in 2011 in which they asked respondents whether they agreed with the statement: "University and college teachers earn too much". In 2011, $y_2 = 540$ people agreed with the statement as compared to $y_1 = 520$ people in 2010. 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 would be

$$\begin{aligned} L(\theta) &= \binom{2000}{520}\theta^{520}(1-\theta)^{1480}\binom{2000}{540}\theta^{540}(1-\theta)^{1460} \\ &= \binom{2000}{520}\binom{2000}{540}\theta^{1060}(1-\theta)^{2940} \quad \text{for } 0<\theta<1 \end{aligned}$$

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

$$L(\theta) = \theta^{1060}(1-\theta)^{2940} \quad \text{for } 0<\theta<1.$$

The maximum likelihood estimate of θ based on the two independent experiments is $\hat{\theta} = 1060/4000 = 0.265$.

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

Example 2.2.3

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 $Y \backsim \text{Binomial}(n, \theta)$ 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

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

or more simply

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

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 $Y_i \backsim$ Binomial$(1; \theta)$ with probability function

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

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

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

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 2.2.4 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:

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

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 what we do not observe Y, but rather the "presence" indicator $I(Y > 0)$, or

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

Note that from (2.6),

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

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. The likelihood function is then

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

and the log likelihood function is

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

We cannot maximize $l(\theta)$ mathematically by solving $dl/d\theta = 0$, so we will use *numerical methods*. Suppose for example that $n = 40$ samples gave data as follows:

v_i (ml)	8	4	2	1
no. of samples	10	10	10	10
no. with $z_i = 1$	10	8	7	3

This gives

$$\begin{aligned} l(\theta) = {} & 10 \log(1 - e^{-8\theta}) + 8 \log(1 - e^{-4\theta}) + 7 \log(1 - e^{-2\theta}) \\ & + 3 \log(1 - e^{-\theta}) - 21\theta \quad \text{for } \theta > 0. \end{aligned}$$

Either by maximizing $l(\theta)$ numerically for $\theta > 0$, 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.

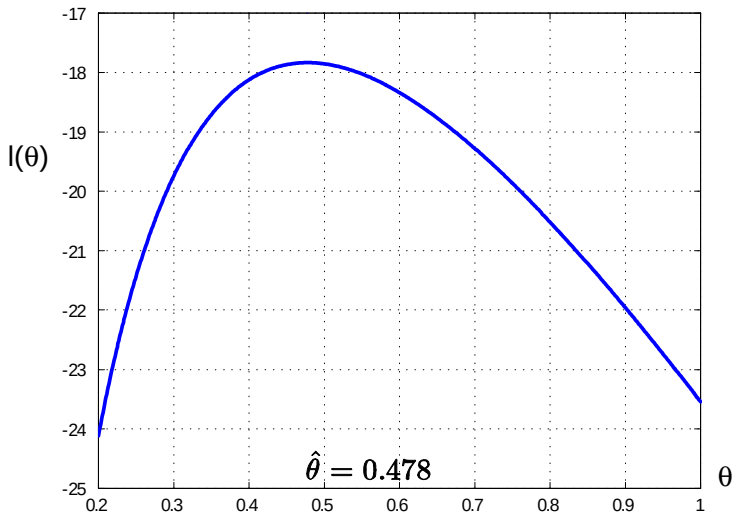


Figure 2.4: **The log likelihood function $l(\theta)$ for Example 2.2.4**

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 continuous distributions, $P(\mathbf{Y} = \mathbf{y}; \theta)$ is unsuitable as a definition of the likelihood since it always equals zero. In the continuous case, we define the likelihood function similarly to the discrete case but with the probability function $P(\mathbf{Y} = \mathbf{y}; \theta)$ replaced by the joint probability density function evaluated at the observed values. If $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables each with probability density function $f(y; \theta)$ then the joint probability density function of $(Y_1, Y_2, \ldots, Y_n)$ is

$$\prod_{i=1}^{n} f(y_i; \theta)$$

and we use this to construct the likelihood function.

Definition 12 *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*

$$L(\theta) = L(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta) \quad \textit{for } \theta \in \Omega. \tag{2.7}$$

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 $Y \backsim$ Exponential(θ) 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, from (2.7),

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

The log likelihood function is

$$l(\theta) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} y_i = -n\left(\log \theta + \frac{\bar{y}}{\theta}\right) \quad \text{for } \theta > 0$$

with derivative

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

A first derivative test easily verifies that the value $\theta = \bar{y}$ maximizes $l(\theta)$ and so $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate of θ.

Example 2.3.2 Likelihood function for Gaussian distribution

As an example involving more than one parameter, suppose that the random variable Y has a $G(\mu, \sigma)$ with probability density function

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

The likelihood function for $\boldsymbol{\theta} = (\mu, \sigma)$ based on the observed random sample $y_1, y_2, \ldots, y_n$ 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

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

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

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

To maximize $l(\mu, \sigma)$ with respect to both parameters μ and σ we solve [12] the two equations[13]

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

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

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

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

[12] 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.

[13] In case you have not met partial derivatives, the notation $\frac{\partial}{\partial \mu}$ means we are taking the derivative with respect to μ while holding the other parameter σ constant. Similarly $\frac{\partial}{\partial \sigma}$ is the derivative with respect to σ while 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 $\theta_1, \theta_2, \ldots, \theta_k$.[14]

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 $\theta_1 = 0.448$, $\theta_2 = 0.083$, $\theta_3 = 0.034$, $\theta_4 = 0.436$.)

[14] $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{k} y_i \log \theta_i$ 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

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

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 $0 < \alpha < 1$. In this case

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

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

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

or more simply

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

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 Estimates

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 13 *If $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ 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.

For data that have arisen from a discrete probability model, a straightforward way to check the fit of the model is to compare observed frequencies with the expected frequencies calculated using the assumed model as illustrated in the example below.

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 omitted 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 2.1.

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

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

The expected frequencies are also given in Table 2.1.

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.

Table 2.1: Frequency Table for Rutherford/Geiger Data

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

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 2.2.

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 2.2. 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.

Table 2.2: Frequency Table for Brake Pad 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

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 [15]

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

The first graphical method is to superimpose the probability density function on the relative frequency histogram of the data as we did in Figures 1.15 and 1.16 for the data from the can filler study.

Empirical Cumulative Distribution Functions

A second graphical procedure is to plot the empirical cumulative distribution function $\hat{F}(y)$ and then to superimpose on this a plot of the model-based cumulative distribution function, $P(Y \leq y; \theta) = F(y; \theta)$. We saw an example of such a plot in Chapter 1 but we provide more detail here. The objective is to compare two cumulative distribution functions, one that we hypothesized is the cumulative distribution function for the population, and the other obtained from the sample. If they differ a great deal, this would suggest that the hypothesized distribution is a poor fit.

[15] See the video at www.watstat.ca called "The empirical c.d.f. and the qqplot".

Example 2.6.3 Checking a Uniform$(0,1)$ model

Suppose, for example, we have 10 observations which we think might come from the Uniform$(0,1)$ distribution. The observations are as follows:

$$0.76 \quad 0.43 \quad 0.52 \quad 0.45 \quad 0.01 \quad 0.85 \quad 0.63 \quad 0.39 \quad 0.72 \quad 0.88.$$

The first step in constructing the empirical cumulative distribution function is to order the observations from smallest to largest[16] obtaining

$$0.01 \quad 0.39 \quad 0.43 \quad 0.45 \quad 0.52 \quad 0.63 \quad 0.72 \quad 0.76 \quad 0.85 \quad 0.88$$

If you were then asked, purely on the basis of this data, what you thought the probability is that a random value in the population falls below a given value y, you would probably respond with the proportion in the sample that falls below y. For example, since four of the values 0.01 0.39 0.43 0.45 are less than 0.5, we would estimate the cumulative distribution function at 0.5 using 4/10. Thus, we define the *empirical cumulative distribution function* for all real numbers y by the proportion of the sample less than or equal to y or:

$$\hat{F}(y) = \frac{\text{number of values in } \{y_1, y_2, \ldots, y_n\} \text{ which are } \leq y}{n}.$$

More generally for a sample of size n we first order the y_i's, $i = 1, 2, \ldots, n$ to obtain the ordered values $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$. $\hat{F}(y)$ is a step function with a jump at each of the ordered observed values $y_{(i)}$. If $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ are all different values, then $\hat{F}(y_{(j)}) = j/n$ and the jumps are all of size $1/n$. In general the size of a jump at a particular point y is the number of values in the sample that are equal to y, divided by n:

$$\text{size of jump in } \hat{F}(y) \text{ at } y = \frac{\text{number of values in } \{y_1, y_2, \ldots, y_n\} \text{ equal to } y}{n}.$$

Why is this a step function? In the data above there were no observations at all between the smallest number 0.01 and the second smallest 0.39. So for all $y \in [0.01, 0.39)$, the proportion of the sample which is less than or equal to y is the same, namely 1/10.

Having obtained this estimate of the population cumulative distribution function, it is natural to ask how close it is to a given cumulative distribution function, say the Uniform$(0,1)$ cumulative distribution function. We can do this with a graph of the empirical cumulative distribution function or more simply on a graph that just shows the vertices $\left(y_{(1)}, \frac{1}{n}\right), (y_{(2)}, \frac{2}{n}), \ldots, (y_{(n)}, \frac{n}{n})$ shown as star on the graph in Figure 2.5.

[16] Recall that we denote the ordered data values as $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ where $y_{(1)}$ is the smallest and $y_{(n)}$ is the largest.

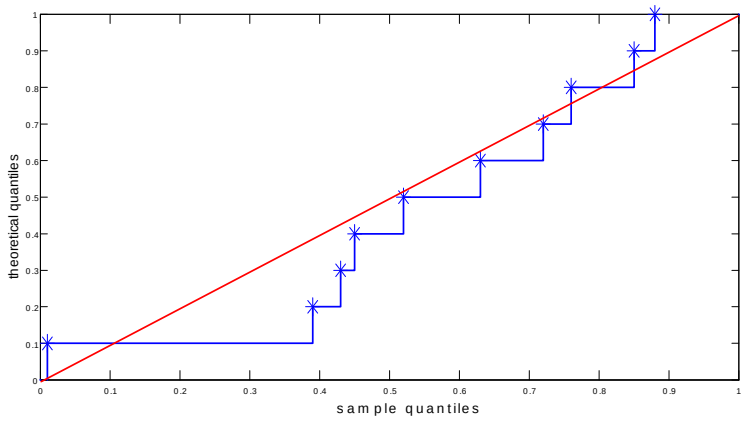


Figure 2.5: **The empirical cumulative distribution function for $n = 10$ data values and a superimposed Uniform$(0, 1)$ cumulative distribution function.**

By superimposing on this graph the theoretical Uniform$(0, 1)$ cumulative distribution function, which in this case is a straight line, we can see how well the theoretical distribution and empirical distribution agree. Since the sample is quite small we cannot expect a perfect straight line, but for larger samples we would expect much better agreement with the straight line.

Because the Uniform$(0, 1)$ cumulative distribution function is a straight line, it is easy to assess graphically how close the two curves fit, but what if the hypothesized distribution is Normal, whose cumulative distribution function is distinctly non-linear?

As an example we 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 2.6. One might hypothesize that the distribution of times between consecutive eruptions follows a Gaussian distribution. To see how well a Gaussian model fit the data we could superimpose a Gaussian cumulative distribution function on the plot of the empirical cumulative distribution function. The 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 2.6 the cumulative distribution function of a $G(72.3, 13.9)$ random variable is superimposed on the empirical cumulative distribution function for the data. Are the differences between the two curves sufficient that we would have to conclude a distribution other than the Gaussian? There are two other ways of examining the magnitude of these differences between the fitted model and the data. The first way is to plot the relative frequency histogram of the data and then superimpose the $G(72.3, 13.9)$ probability density

function. The second way is to use a qqplot which will be discussed in the next section.

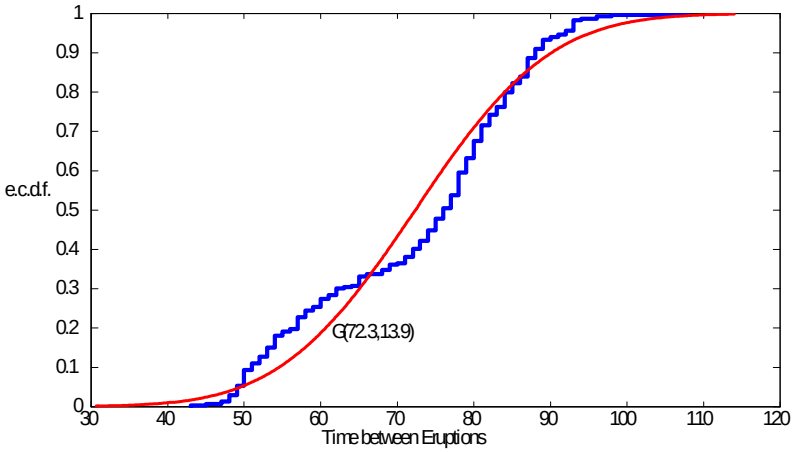


Figure 2.6: **Empirical c.d.f. of times between eruptions of Old Faithful and superimposed** $G\left(72.3, 13.9\right)$ **c.d.f.**

Figure 2.7 seems to indicate that the distribution of the times between eruptions is not very Normal because it appears to have two modes. The plot of the empirical cumulative distribution function did not show the shape of the distribution as clearly as the histogram. The empirical cumulative distribution function does allow us to determine the pth quantile or $100p$th percentile (the left-most value on the horizontal axis y_p where $\hat{F}(y_p) = p$). For example, from the empirical cumulative distribution function of the Old Faithful data, we see that the median time ($\hat{F}(\hat{m}) = 0.5$) between eruptions is around $\hat{m} = 78$.

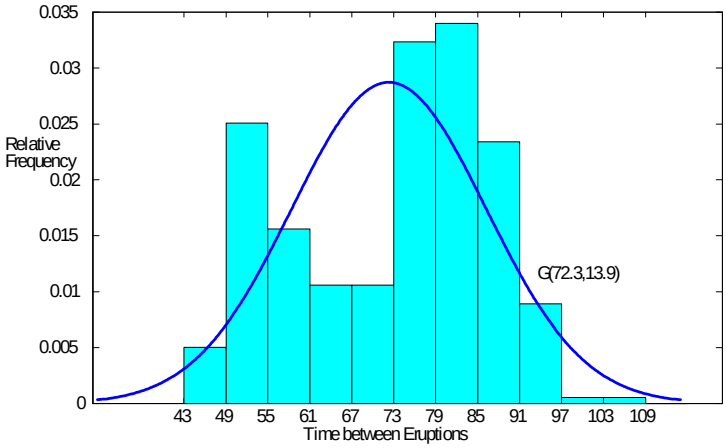


Figure 2.7: **Relative frequency histogram for times between eruptions of Old Faithful and superimposed** $G\left(72.3, 13.9\right)$ **p.d.f.**

Example 2.6.4 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 2.8 shows a plot of the empirical cumulative distribution function with the $G(1.62, 0.064)$ cumulative distribution function superimposed.

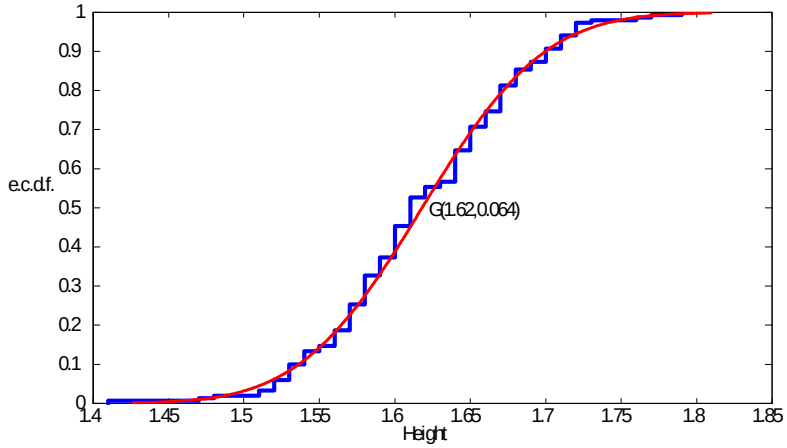


Figure 2.8: **Empirical c.d.f. of female heights and $G\,(1.62, 0.064)$ c.d.f.**

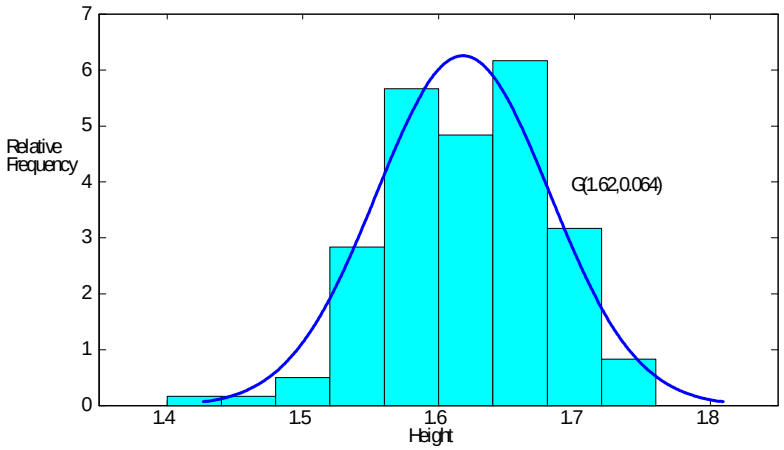


Figure 2.9: **Relative frequency histogram of female heights and $G\,(1.62, 0.064)$ p.d.f.**

Figure 2.9 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 Normal model seems reasonable for these data.

Qqplots

An alternative view, which is really just another method of graphing the empirical cumulative distribution function, tailored to the Normal distribution, is a graph called a *qqplot*. Suppose the data Y_i, $i = 1, 2, \ldots, n$ were in fact drawn from the $G(\mu, \sigma)$ distribution so that the standardized variables, after we order them from smallest $Y_{(1)}$ to largest $Y_{(n)}$, are

$$Z_{(i)} = \frac{Y_{(i)} - \mu}{\sigma}.$$

These behave like the ordered values from a sample of the same size taken from the $G(0, 1)$ distribution. Approximately what value do we expect $Z_{(i)}$ to take? If Φ denotes the standard Normal cumulative distribution function then for $0 < u < 1$

$$P(\Phi(Z) \leq u) = P(Z \leq \Phi^{-1}(u)) = \Phi(\Phi^{-1}(u)) = u$$

so that $\Phi(Z)$ has a Uniform distribution. It is easy to check that the expected value of the i'th largest value in a random sample of size n from a Uniform$(0, 1)$ distribution is equal to $\frac{i}{n+1}$[17] so we expect that the i/n'th quantile $\Phi(Z_{(i)})$ to be close to $\frac{i}{n+1}$. In other words we expect $Z_{(i)} = \left(Y_{(i)} - \mu\right)/\sigma$ to be approximately $\Phi^{-1}\left(\frac{i}{n+1}\right)$ or $Y_{(i)}$ to be roughly a linear function of $\Phi^{-1}\left(\frac{i}{n+1}\right)$. This is the basic argument underlying the qqplot. *If the distribution is actually Normal, then a plot* $\left(Y_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$, $i = 1, 2, \ldots, n$ *should be approximately linear* (subject to the usual randomness).

Similarly if the data obtain from an Exponential distribution we expect a plot of $\left(Y_{(i)}, F^{-1}\left(\frac{i}{n+1}\right)\right)$ to be approximately linear where $F^{-1}(u)$ is the inverse of the Exponential(1) cumulative distribution function given by $F^{-1}(u) = -\ln(1 - u)$.

Since reading qqplots is an art acquired from experience, it is a good idea to generate similar plots where we know the answer. This can be done by generating data from a known distribution and then plotting a qqplot. See the *R* code below and Chapter 2, Problem 18. A qqplot of 100 observations randomly generated from a $G(-2, 3)$ distribution is given in Figure 2.10. The theoretical quantiles are plotted on the horizontal axis and the empirical quantiles are plotted on the vertical axis. **Since the quantiles of the Normal distribution change more rapidly in the tails of the distribution, we expect the points at both ends of the line to lie further from the line.**

[17] This is intuitively obvious since n values $Y_{(i)}$ breaks the interval into $n + 1$ spacings, and it makes sense each should have the same expected length. For empirical evidence see http://www.math.uah.edu/stat/applets/OrderStatisticExperiment.html. More formally we must first show the p.d.f. of $Y_{(i)}$ is $\frac{n!}{(i-1)!(n-i)!}u^{i-1}(1-u)^{n-i}$ for $0 < u < 1$. Then find the integral $E(Y_{(i)}) = \int_0^1 \frac{n!}{(i-1)!(n-i)!}u^i(1-u)^{n-i}du = \frac{i}{n+1}$.

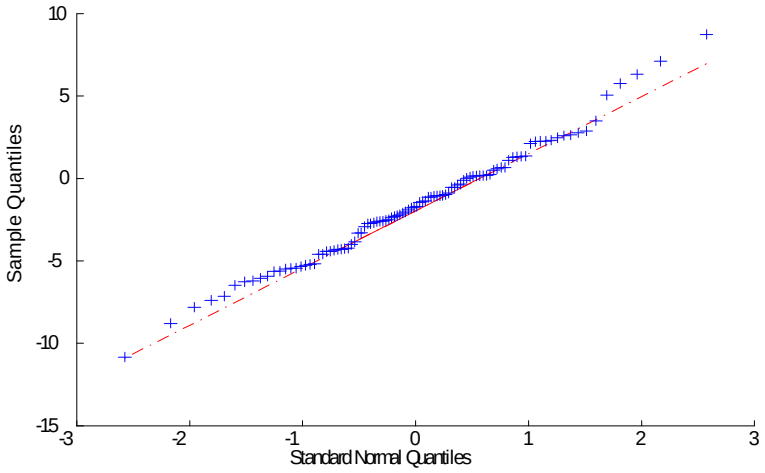


Figure 2.10: **Qqplot of a random sample of** 100 **observations from a** $G(-2, 3)$ **distribution**

A qqplot of the female heights is given in Figure 2.11. Overall the points lie reasonably along a straight line. The qqplot has a staircase look because the heights are rounded to the closest centimeter. As was the case for the relative frequency histogram and the empirical cumulative distribution function, the qqplot indicates that the Normal model is reasonable for these data.

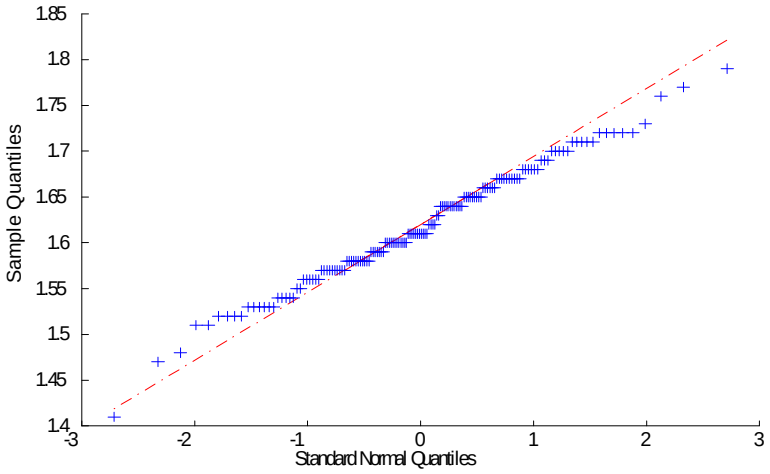


Figure 2.11: **Qqplot of heights of females**

A qqplot of the times between eruptions of Old Faithful is given in Figure 2.12. The points do not lie along a straight line which indicates as we saw before that the Normal 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.

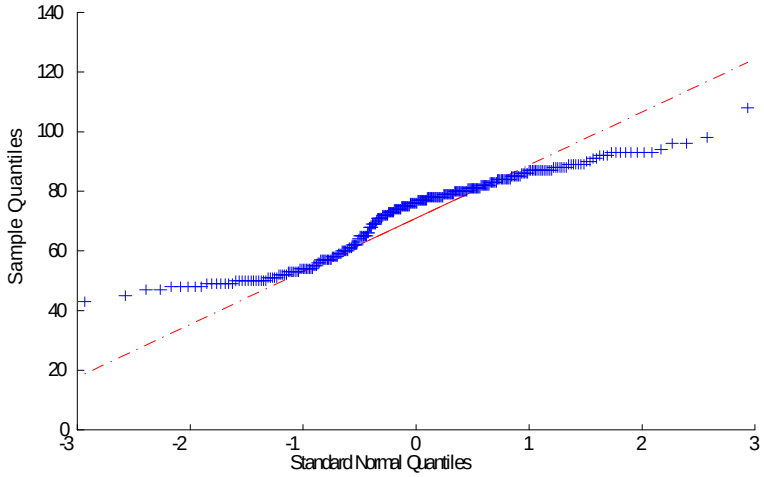


Figure 2.12: **Qqplot of times between eruptions of Old Faithful**

A qqplot of the lifetimes of brake pads (Example 1.3.4) is given in Figure 2.13. The points form a U-shaped curve. This pattern is consistent with the long right tail and positive skewness that we observed before. The Normal is not a reasonable model for these data.

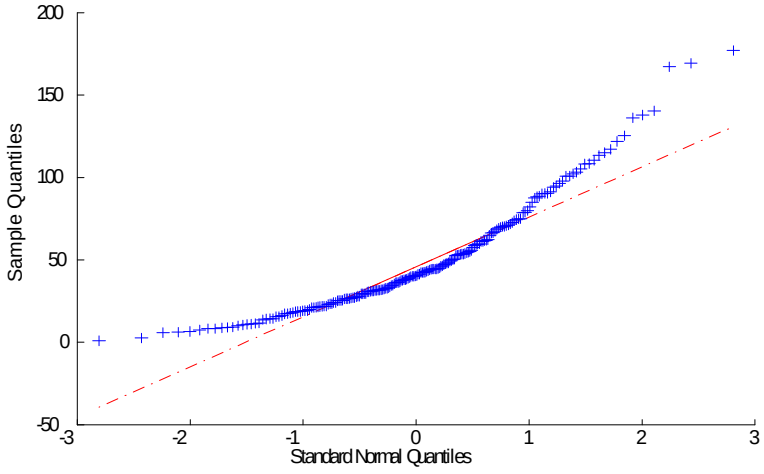


Figure 2.13: **Qqplot of lifetimes of brake pads**

A qqplot of the data in Figure 1.4 is given in Figure 2.14. These points form an S-shaped curve which is consistent with the fact that the data are reasonably symmetric but the data do not have tails like the Normal distribution.

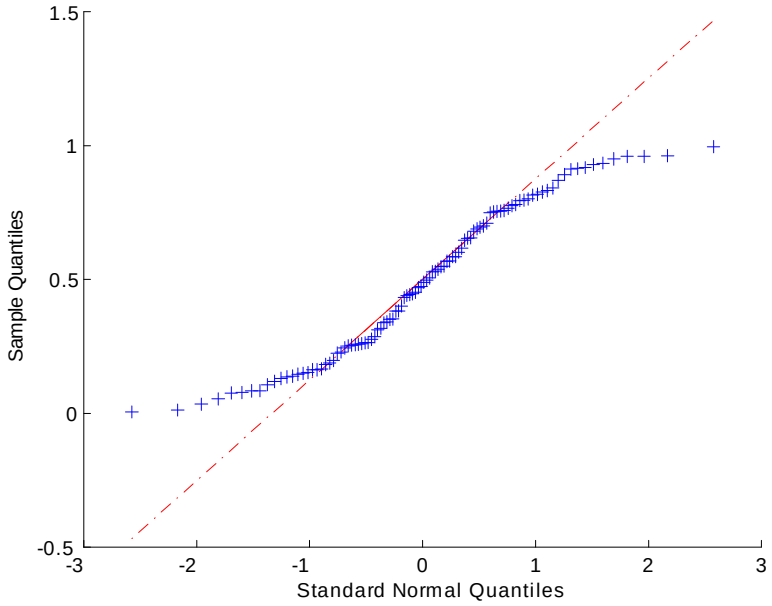


Figure 2.14: Qqplot of 100 observations

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

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

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

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

2. 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 functions *gbinom()* or *pbinom()* to determine the exact value of n.

3. 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

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

   (a) For observed data are $y_1, y_2, \ldots, y_n$, 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.

4. 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

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

   (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 during a two-minute interval, no transactions arrive.

   (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 variance; are they approximately the same? (Note that $E(Y) = Var(Y) = \theta$ for the Poisson model.)

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

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

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

   (b) Find the relative likelihood function $R(\theta)$. If $n = 20$ and $\sum_{i=1}^{20} y_i^2 = 72$ then plot $R(\theta)$.

6. 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 $\hat{\sigma}$.

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

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

   (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 $r(\theta)$.

8. 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 $\hat{\alpha}$ of α. What is the value of $\hat{\alpha}$ if $n = 50$ and $n_1 = 16$, $n_2 = 16$, $n_3 = 18$?

9. When Wayne Gretzky played for the Edmonton Oilers 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

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

   (a) Show that the likelihood function for θ based on the Poisson model and the data in the frequency table simplifies to

   $$L(\theta) = \theta^{1669} e^{-696\theta}, \quad \theta > 0.$$

   (b) Find the maximum likelihood estimate of θ.

   (c) Determine the expected frequencies based on the Poisson model and 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.)

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

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

(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. Show that

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

and determine c.

(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.

11. 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.

(a) Determine an estimate of θ from these data. What assumptions have you made to do this?

(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.

12. Run the following **R** code for checking the Gaussian model using histograms, empirical c.d.f.'s and qqplots

```
#  Gaussian Data Example
set.seed(456458)
yn<-rnorm(200,5,2) # 200 observations from G(5,2) distribution
c(mean(yn),sd(yn)) # display sample mean and standard deviation
fivenum(yn) # five number summary
skewness(yn) # sample skewness
kurtosis(yn) # sample kurtosis
#plot relative frequency histogram and superimpose Gaussian pdf
truehist(yn,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="y",ylab="ecdf",main="")
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 Normal Quantiles",main="Qqplot of Data")
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
c(mean(ye),sd(ye)) # display sample mean and standard deviation
fivenum(ye) # five number summary
skewness(ye) # sample skewness
kurtosis(ye) # sample kurtosis
#plot relative frequency histogram and superimpose Gaussian pdf
truehist(ye,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="y",ylab="ecdf",main="")
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 Normal Quantiles",main="Qqplot of Data")
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 above to assess whether it is reasonable to assume that the data have approximately a Gaussian distribution. Support your conclusion with clear reasons.

13. The marks for 100 students on a tutorial test in STAT 231 were:

3	5	11.5	13	13	13	13.5	13.5	13.5	13.5
14	14	14.5	14.5	14.5	15	15	15	15.5	15.5
15.5	16	16	16	16	16.6	16.5	17	17	17
17	17	17	17	17	17	17.5	17.5	18	18
18.5	18.5	18.5	18.5	19	19	19	19	19	19.5
19.5	19.5	20	20	20	20	20	20	20	20
20	20	20.5	20.5	20.5	20.5	21	21	21	21.5
21.5	21.5	22	22	22	22	22	22.5	22.5	22.5
23	23	23	23	23	23.5	24.5	25	25	25
25	25	25.5	26	26	26	26.5	27	27	30

The data are available in the file *tutorialtestdata.txt* posted on the course website. For these data

$$\sum_{i=1}^{100} y_i = 1914 \quad \text{and} \quad \sum_{i=1}^{100} y_i^2 = 38609.$$

The sample skewness is -0.50 and the sample kurtosis is 4.30.

A boxplot and qqplot of the data are given in Figures 2.15 and 2.16.

(a) Determine the five-number summary for these data.

(b) Determine the sample mean $\bar{y}$ and the sample standard deviation s for these data.

(c) 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 $Y \backsim G(\mu, \sigma)$.

(d) Find the interquartile range (IQR) for these data. Show that for Normally distributed data IQR $= 1.349\sigma$. How well do these data satisfy this relationship?

(e) Using both the numerical and graphical summaries for these data, assess whether it is reasonable to assume that the data are approximately Normally distributed. Be sure to support your conclusion with clear reasons.

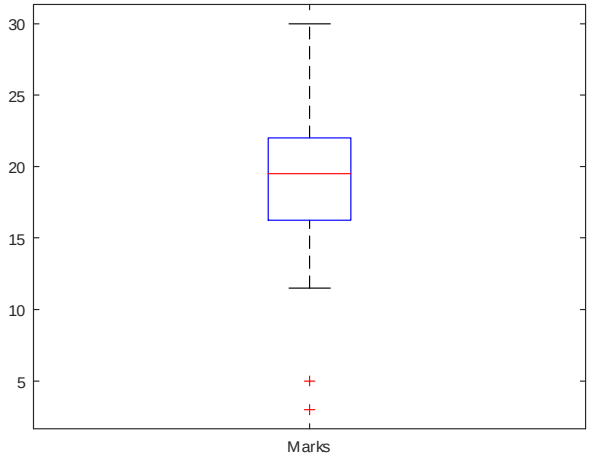


Figure 2.15: Boxplot of tutorial test marks

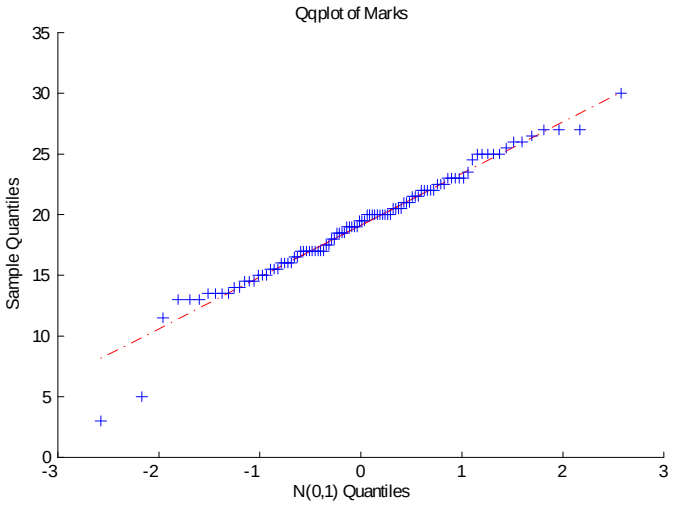


Figure 2.16: Qqplot of tutorial test marks

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

Heights of Elderly Women

142	145	145	145	146	147	147	147	147	148	148	149	150	150	150
150	150	150	151	151	151	151	151	151	152	152	152	152	152	152
152	152	152	152	152	152	153	153	153	153	153	153	153	153	153
153	153	153	153	153	153	153	153	154	154	154	154	154	154	154
154	154	154	154	155	155	155	155	155	155	155	155	155	155	155
155	155	155	155	155	155	155	155	155	155	156	156	156	156	156
156	156	156	156	156	156	156	156	156	156	156	156	156	156	156
157	157	157	157	157	157	157	157	157	157	157	157	157	157	157
157	157	157	157	157	158	158	158	158	158	158	158	158	158	158
158	158	158	158	158	158	158	158	158	158	158	158	158	158	158
158	158	158	158	158	158	159	159	159	159	159	159	159	159	159
159	159	159	159	159	159	159	159	160	160	160	160	160	160	160
160	160	160	160	160	160	160	160	160	160	160	160	160	160	161
161	161	161	161	161	161	161	161	161	161	161	161	161	161	161
161	161	161	161	162	162	162	162	162	162	162	162	162	162	162
162	162	162	162	162	162	162	163	163	163	163	163	163	163	163
163	163	163	163	163	163	163	163	163	163	163	163	163	163	163
163	163	163	163	163	163	163	164	164	164	164	164	164	164	164
164	164	164	164	164	164	164	164	164	165	165	165	165	165	165
165	165	165	165	165	165	165	165	165	165	165	165	166	166	166
166	166	166	166	166	166	166	166	167	167	167	167	167	167	167
168	168	168	168	168	168	169	169	169	169	169	169	169	169	170
170	170	170	170	170	170	170	170	170	170	171	171	171	173	174
173	174	176	177	178	178									

For these data

$$\sum_{i=1}^{351} y_i = 56081 \quad \sum_{i=1}^{351} y_i^2 = 8973063$$

(a) Determine the sample mean $\bar{y}$ and the sample standard deviation s for these data.

(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 $Y \backsim G(\mu, \sigma)$.

(c) Find the sample skewness and sample kurtosis for these data. Are these values close to what you would expect for Normally distributed data?

(d) Find the five-number summary for these data.

(e) Find the IQR for these data. Does the IQR agree with what you expect for Normally distributed data?

(f) Construct a relative frequency histogram and superimpose a Gaussian probability density function with $\mu = \bar{y}$ and $\sigma = s$.

(g) Construct an empirical distribution function for these data and superimpose a Gaussian cumulative distribution function with $\mu = \bar{y}$ and $\sigma = s$.

(h) Draw a boxplot for these data.

(i) Plot a qqplot for these data. Do you observe anything unusual about the qqplot? Why might cause this?

(j) Based on the above information indicate whether it is reasonable to assume a Gaussian distribution for these data.

15. 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) Assuming that for **each** sex the heights Y in the population from which the samples were drawn is adequately represented by $Y \sim G(\mu, \sigma)$, obtain the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ 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, or 6 ft).

(d) A simpler estimate of $P(Y > 1.83)$ that doesn't use the Gaussian model is

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

where here $n = 150$. 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).

16. The qqplot of the brake pad data in Figure 2.13 indicates that the Normal distribution is not a reasonable model for these data. Sometimes transforming the data gives a data set for which the Normal model is more reasonable. A log transformation is often used. Plot a qqplot of the log lifetimes and indicate whether the Normal distribution is a reasonable model for these data. The data are posted on the course website.

17. In a large population of males ages $40-50$, the proportion who are regular smokers is α where $0<\alpha<1$ and the proportion who have hypertension (high blood pressure) is β where $0<\beta<1$. 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?

18. **Interpreting qqplots:** Consider the following data sets defined by R commands. For each generate the Normal qqplot using `qqnorm(y)` and on the basis of the qqplot determine whether the underlying distribution is symmetric, light-tailed, heavy tailed, whether the skewness is positive, negative or approximately zero, and whether the kurtosis is larger or smaller than that of the Gaussian, i.e. 3. Repeat changing the sample size $n=100$ to $n=25$. How much more difficult is it in this case to draw a clear conclusion?

    (a) `y<-rnorm(100)`

    (b) `y<-runif(100)`

    (c) `y<-rexp(100)`

    (d) `y<-rgamma(100,4,1)`

    (e) `y<-rt(100,3)`

    (f) `y<-rcauchy(100)`

19. A qqplot was generated for 100 values of a variate. See Figure 2.17. Based on this qqplot, answer the following questions:

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

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

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

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

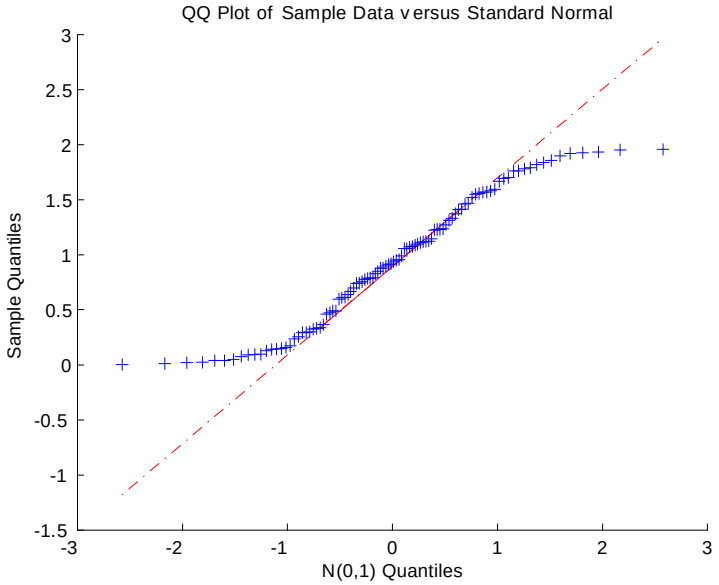


Figure 2.17: Qqplot for 100 observations

20. **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?

21. **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. **Warning:** The maximum likelihood estimate is not found by solving $l'(\theta) = 0$.

22. **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 $Y \backsim$ 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 does part (b) give 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.

23. **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 a covariate (explanatory variable) 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 of planning and conducting a study as a set of steps. We describe below the set of steps to which we assign the acronym PPDAC

- **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 we will collect the data.
- **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.

PPDAC has been designed to emphasize the statistical aspects of empirical studies. We develop each of the five steps in more detail below. Several examples of the use of PPDAC in an empirical study will be given. We identify the steps in the following example.

Example 3.1
The following news item was published by the University of Sussex, UK on February 16, 2015. It 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."

Here are the five steps:

- **Problem:** To study the effect of four different mental exercises related to non-drinking on the drinking behaviour of young people.
- **Plan:** Recruit university 211 students aged 18-25 in the United Kingdom and assign the students to one of the four mental exercises. (The article in the British Journal of Health Psychology indicated that academic departments across English universities were asked to forward a pre-prepared recruitment message to their students containing a URL to an online survey.) Collect information from the students via online surveys at various points including how much alcohol they had drunk the previous week.
- **Data:** The data collected included which mental exercise group the student was in and information about their alcoholic consumption in the week before they completed the various online surveys.
- **Analysis:** Look at differences in alcoholic consumption between the four groups.
- **Conclusion:** The study found that completing mental exercises relating to non-drinking was more effective in encouraging safer drinking behaviour than completing a drinks diary alone.

Note that in the Problem step, we describe **what** we are trying to learn or **what** questions we want to answer. The Plan step describes **how** the data are to be measured and collected. In the Data step, the Plan is executed. The Analysis step corresponds to what many people think Statistics is all about. We carry out both simple and complex calculations to process the data into information. Finally, in the Conclusion step, we answer the questions formulated at the Problem step.

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 (as in the above example), 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.

To discuss the steps of PPDAC in more detail we need to introduce a number of technical terms. Every subject has its own jargon, i.e. 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

1. Problem

The elements of the Problem address questions starting with "What"

- What conclusions are we trying to draw?
- What group of things or people do we want the conclusions to apply?
- What variates can we define?
- What is(are) the question(s) we 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. 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 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 effects the value of their investments.

In the second type of 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 experiment in which the experimenter manipulates the values of the explanatory variates is referred to as an *experimental study.* On the other hand in the study of whether second-hand smoke causes asthma, it is unlikely that the experimenter would

be able to manipulate the explanatory variate and so the experimenter needs to rely on a potentially less informative *observational study*, one that depends on data that is collected without the ability to control explanatory variates. We will see in Chapter 8 how an empirical study must be carefully designed in order to answer such causative questions. Important considerations in an observational study are the design of the survey and questionnaire, who to ask, what to ask, how many to ask, where to sample etc.

Defining the Problem

The first step in describing the Problem is to define the *units* and the *target population* or *target process*.

Definition 14 *The target population or process is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.*

In Chapter 1, we considered a survey of teenagers in Ontario in a specific week to learn about their smoking behaviour. In this example the units are teenagers in Ontario at the time of the survey and the target population is all such teenagers.

In another example, we considered the comparison of two machines with respect to the volume of liquid in cans being filled. The units are the individual cans. The target population (or perhaps it is better to call it a process) is all such cans filled now and into the future under current operating conditions. Sometimes we will be vague in specifying the target population, i.e. "cans filled under current conditions" is not very clear. What do we mean by current conditions, for example?

Definition 15 *A variate is a characteristic associated with each unit.*

For each teenager (unit) in the target population, the variate of primary interest is whether or not the teenager smokes. 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 is a variate. The machine that filled the can is another variate. A key point to notice is that the values of the variates change from unit to unit in the population. There are usually many variates associated with each unit. At this stage, we will be interested in only those that help specify the questions of interest.

Definition 16 *An attribute is a function of the variates over a population.*

We specify the questions of interest in the Problem in terms of attributes of the target population. In the smoking example, one important attribute is the proportion of teenagers in the target population. In the can-filling example, the attributes of interest were the average volume and the variability of the volumes for all cans filled by each machine under current conditions. Possible questions of interest (among others) are:

"What proportion of teenagers 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 or a scatterplot of one variate versus another over the whole population.

It is very important that the Problem step contain clear questions about one or more attributes of the target population.

2. Plan

In most cases, we cannot calculate the attributes of interest for the target population directly because we can only examine a subset of the units in the target population. This may be due to lack of resources and time, as in the smoking survey or a physical impossibility as in the can-filling study where we can only look at cans available now and not in the future. Or, in an even more difficult situation, we may be forced to carry out a clinical trial using mice because it is unethical to use humans and so we do not examine any units in the target population. Obviously there will be uncertainty in our answers. The purpose of the Plan step is to decide what units we will examine (the *sample*), what data we will collect and how we will do so. The Plan depends on the questions posed in the Problem step.

Definition 17 *The study population or study process is the collection of units available to be included in the study.*

Often the study population is a subset of the target population (as in the teenage smoking survey). However, in many medical applications, the study population consists of laboratory animals whereas the target population consists of humans. In this case the units in the study population are laboratory animals and the units in the target population are humans. In the development of new products, we may want to draw conclusions about a production process in the future but we can only look at units produced in a laboratory in a pilot process. In this case, the study units are not part 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. Therefore the study population excludes those people without telephones or with unlisted numbers.

The study population is often not identical to the target population.

Definition 18 *If the attributes in the study population differ from the attributes in the target population then the difference is called study error.*

We cannot quantify study error but must rely on context experts to know, for example, that conclusions from an investigation using mice will be relevant to the human target population. We can however warn the context experts of the possibility of such error, especially when the study population is very different from the target population.

Definition 19 *The sampling protocol is the procedure used to select a sample of units from the study population. 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 we "take what we can get". Sample size is usually driven by economics or availability. We will show in later chapters how we can use the model to help with sample size determination.

Definition 20 *If the attributes in the sample differ from the attributes in the study population the difference is called sample error.*

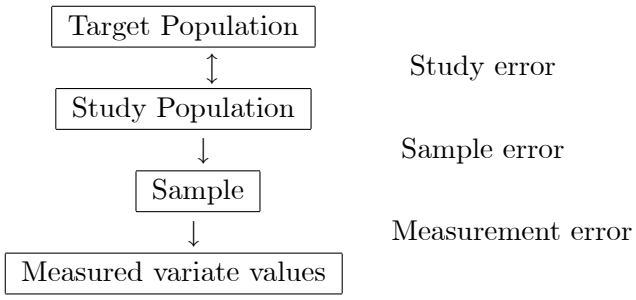
Even with random sampling, we are looking at only a subset of the units in the study population. Differing sampling protocols are likely to produce different sample errors. Also, since we do not know the values of the study population attributes, we cannot know the sample error. However, we can use the model to get an idea of how large this error might be. These ideas are discussed in Chapter 4.

We must decide which variates we are going to measure or determine for the units in the sample. For any attributes of interest, as defined in the Problem step, we will certainly measure the corresponding variates for the units in the sample. As we shall see, we may also decide to measure other variates that can aid the analysis. In the smoking survey, we will try to determine whether each teenager in the sample smokes or not (this requires a careful definition) and also many demographic variates such as age and sex so that we can compare the smoking rate across age groups, sex etc. In experimental studies, the experimenters assign the value of a variate 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 21 *If the measured value and the true value of a variate are not identical the difference is called measurement error.*

Measurement errors are usually unknown. In practice, we need to ensure that the measurement systems used do not contribute substantial error to the conclusions. We may have to study the measurement systems which are used in separate studies to ensure that this is so.

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, 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 the Conclusions from 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.

3. Data

The object 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.

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.

Missing data and response bias

Suppose we 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 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 respond at all or change some of their responses to favour the police. 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[18] of the resulting response bias. See for example the CBC story "Census Mourned on World Statistics Day"[19].

4. Analysis

In Chapter 1 we discussed different methods of summarizing the data using numerical and graphical summaries. A key step in formal analyses is the selection of an appropriate model that can describe the data and how it was collected. In Chapter 2 we discussed methods for checking the fit of the model. We also need to describe the Problem in terms of the model parameters and properties. You will see many more formal analyses in subsequent chapters.

5. Conclusions

The purpose of the Conclusion step is to answer the questions posed in the Problem. In other words, the Conclusion is directed by the Problem. An attempt should be made

[18] http://www.youtube.com/watch?v=0A7ojjsmSsY

[19] http://www.cbc.ca/news/technology/story/2010/10/20/long-form-census-world-statistics-day.html

to quantify (or at least discuss) potential errors as described in the Plan step and any limitations to the conclusions.

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.

Figure 3.1: **Crankshaft with arrow pointing to "journal"**

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 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 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. 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.

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

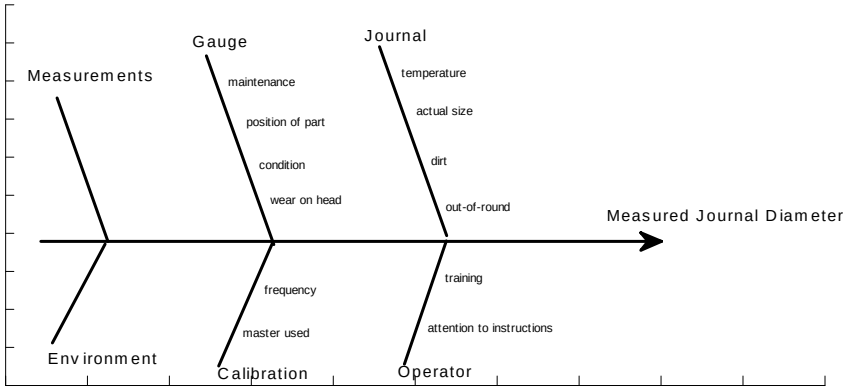


Figure 3.2: **Fishbone diagram for variation in measured journal diameter**

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 seen. 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. 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 $Y_{ij} \backsim G(\mu_i, \sigma_m)$ then

$$P(\mu_i - 2\sigma_m \leq Y_{ij} \leq \mu_i + 2\sigma_m) = 0.95$$

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 $\pm 2\,(1.37) = \pm 2.74$.

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 population). 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.

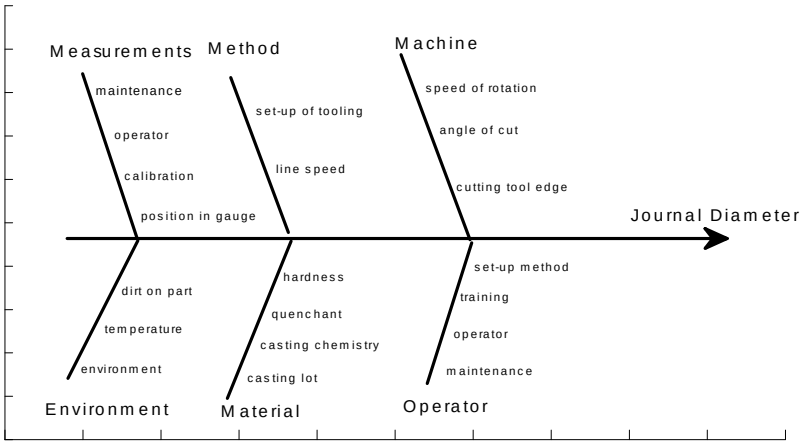


Figure 3.3: **Fishbbone diagram for cause-and-effect**

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 (ten 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 $+16$.

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 represents 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,

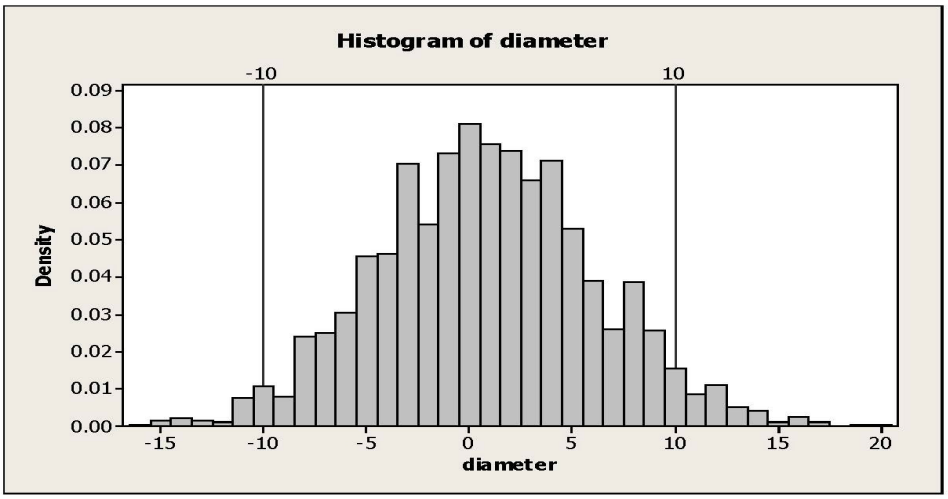


Figure 3.4: **Histogram of 2000 measured values from the gauge memory**

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 populations 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.

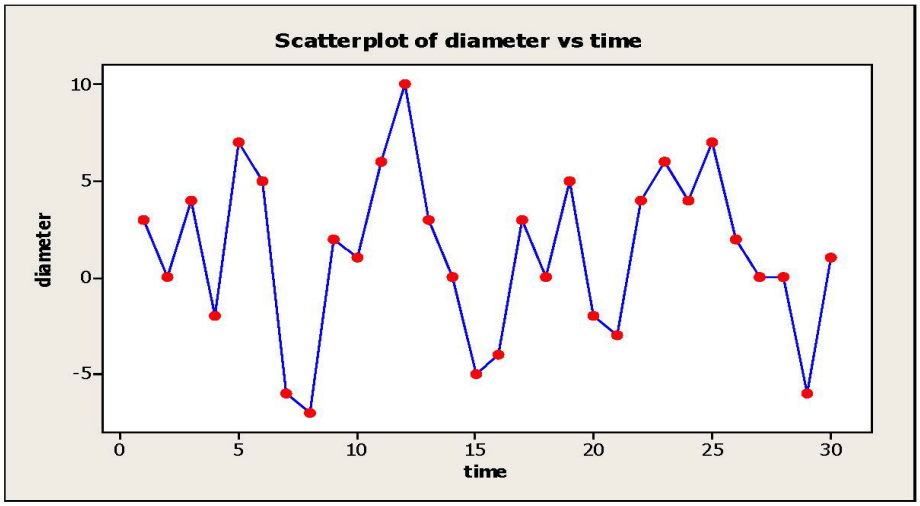


Figure 3.5: **Scatter plot of diameter versus time**

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.

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. 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) Describe one possible source of study error is?

   (h) Describe one possible source of sample error?

   (i) Estimate the attributes of interest for the study population based on the given data.

2. **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 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) Give two variates of interest in this problem and specify the type of variate for each.

(f) Define a suitable study population for this study.

(g) What is the sampling protocol?

(h) What is a possible source of study error is?

(i) What is a possible source of sample error?

(j) What is a possible source of measurement error?

(k) 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?

3. **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 are the two most important variates in this study and what is their type?

(f) What is the attribute of interest in the target population?

(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.

4. 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.

5. 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.

6. 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?

7. 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 for variation in the population or process being studied which includes the attributes which are to be estimated.

(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 for estimating the unknown attributes based on the observed data and determining the uncertainty in the estimates. The unknown attributes are usually represented by unknown parameters θ 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, we may not be able to use the estimates based on it. For the lifetimes of brake pads data introduced in Example 1.3.4, a Gaussian model does not appear to be suitable (see Chapter 2, Problem 16).

(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 $\bar{y}$ is an estimate of θ, but not usually equal to it. How far away from θ is $\bar{y}$ likely to be? If we take a sample of only $n = 50$ units, would we expect the estimate $\bar{y}$ 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

$$\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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 $\hat{\theta} = \bar{y}$ 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*[20], 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 $\hat{\theta} = \bar{y}$ 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.

[20] See the video at www.watstat.ca called "What is a sampling distribution?".

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

$$\tilde{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is a random variable and $\hat{\theta} = \bar{y}$ 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 22 *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 $Y_1, Y_2, \ldots, Y_n$ we can find its distribution, at least in principle. Once we know the sampling distribution of an estimator $\tilde{\theta}$ then we are in a position to express the uncertainty in an estimate. In Examples 4.2.1-4.2.3 we examine ways of finding the sampling distribution, at least approximately. We also look at *the probability that the estimator $\tilde{\theta}$ is "close" to θ.*

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 $Y_i \backsim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$. The maximum likelihood estimator of μ is

$$\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

From properties of Gaussian random variables we know that the sampling distribution of $\tilde{\mu} = \bar{Y}$ is $G(\mu, \sigma/\sqrt{n})$.

If we knew σ we could determine how often the estimator $\tilde{\mu} = \bar{Y}$ is within a specified amount of the mean. For example, if the variate is height and heights are measured in meters than we could determine how often the estimator $\tilde{\mu} = \bar{Y}$ is within 0.01 meters of the true mean μ as follows:

$$\begin{aligned} P\left(|\tilde{\mu} - \mu| \leq 0.01\right) &= P\left(\mu - 0.01 \leq \bar{Y} \leq \mu + 0.01\right) \\ &= P\left(\frac{-0.01}{\sigma/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.01}{\sigma/\sqrt{n}}\right) \\ &= P\left(-0.01\sqrt{n}/\sigma \leq Z \leq 0.01\sqrt{n}/\sigma\right) \quad \text{where } Z \sim G(0,1). \end{aligned}$$

Suppose $\sigma = 0.07$ meters. If $n = 50$ 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 the estimator $\tilde{\mu} = \bar{Y}$ is within 0.01 meters of the true but unknown mean height μ in the 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 $\tilde{\mu} = \bar{Y}$ 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

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \backsim N(0,1) \quad \text{approximately.}$$

This result could be used, for example, 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 $\theta \in [0,1]$. See Problem 1.

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 $\tilde{\mu} = \bar{Y}$ is in the case of the population which has the distribution of variate values as indicated in Table 4.1.

Table 4.1: Distribution of Variate Values in Finite Population

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

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 standard

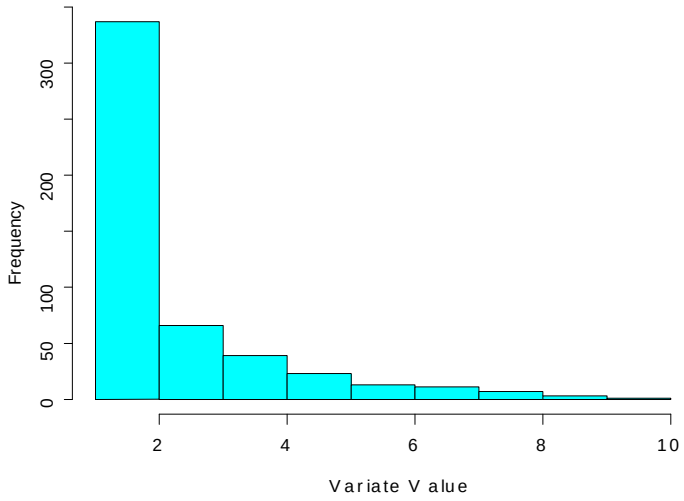


Figure 4.1: Histogram of the variate values for the finite population of Table 4.1

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 $\tilde{\mu} = \bar{Y}$ is we need the sampling distribution of $\bar{Y}$. 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 same mean $\bar{y}$ for this sample is calculated and saved in a vector. These two steps are repeated k times. The k values $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ can then be considered

as a random sample from the distribution of $\tilde{\mu} = \bar{Y}$, 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 Problem 2.

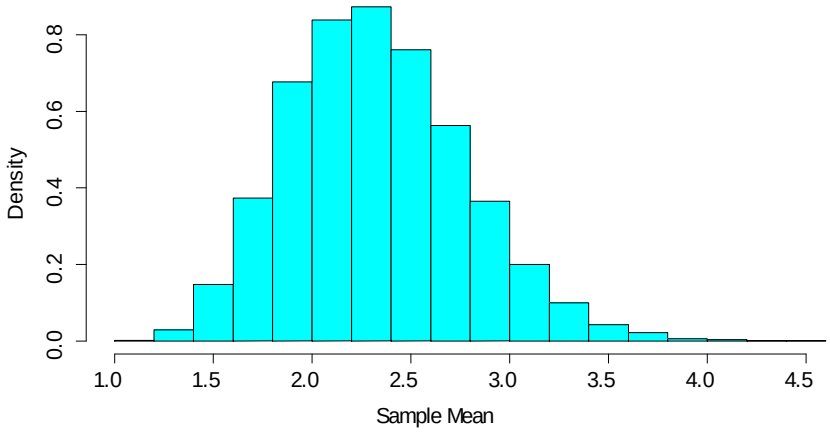


Figure 4.2: **Relative frequency histogram of means from 10000 samples of size 15 drawn from the population defined by Table 4.1**

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 values $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{10000}$ and then plotting the relative frequency histogram. It can be shown[21] that

$$E\left(\bar{Y}\right) = \mu = 2.362 \quad \text{and} \quad sd\left(\bar{Y}\right) \approx \frac{\sigma}{\sqrt{n}} = \frac{1.7433}{\sqrt{15}} = 0.4501$$

Do the results of the simulation agree with these statements? Does the distribution look like a Gaussian distribution? What do you notice about the symmetry of the distribution?

Based on this simulation we can approximate $P\left(\left|\bar{Y} - 2.362\right| \leq 0.5\right)$, the probability that the sample mean $\bar{Y}$ is within 0.5 of the population mean $\mu = 2.362$, by determining 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 $P\left(\left|\bar{Y} - 2.362\right| \leq 0.5\right)$ be affected? See Problem 2.

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 $\hat{\theta}$.

[21] 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 estimates and estimators we have discussed so far are often referred to as *point estimates* and *point estimators* respectively. 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, $L(\mathbf{y})$ and $U(\mathbf{y})$, are both functions of the observed data $\mathbf{y}$. If we let $L(\mathbf{Y})$ and $U(\mathbf{Y})$ represent the random variables associated with $L(\mathbf{y})$ and $U(\mathbf{y})$, then $[L(\mathbf{Y}), U(\mathbf{Y})]$ is called a random interval since the endpoints are random variables. The probability that the parameter θ falls in the random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$ is $P(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})]) = P[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})]$. This probability tells us how good the rule is by which the interval estimate was obtained. It tells us, for example, how often we would expect the statement $\theta \in [L(\mathbf{y}), U(\mathbf{y})]$ to be true if we were to draw many random samples from the same population and each time we constructed the interval $[L(\mathbf{y}), U(\mathbf{y})]$ based on the observed data $\mathbf{y}$. For example, suppose $P[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})] = 0.95$. If we drew a large number of random samples and each time we constructed the interval $[L(\mathbf{y}), U(\mathbf{y})]$ from the data $\mathbf{y}$, then we would expect the true value of the parameter to lie in 95% of these constructed intervals. This means we can be **reasonably confident** that if we construct one interval based on one observed data set $\mathbf{y}$, then the interval $[L(\mathbf{y}), U(\mathbf{y})]$ will contain the true value of the unknown parameter θ. In general, uncertainty in a point estimate is explicitly stated by giving the interval estimate along with the probability $P(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})])$.

We will discuss this idea of confidence related to interval estimates in more detail in Section 4.4. First we show how the likelihood function can be used to construct interval 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 requiring that the parameter falls in 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 23 *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 that

$$0 \leq R(\theta) \leq 1 \quad \textit{for all } \theta \in \Omega.$$

Definition 24 *A $100p\%$ likelihood interval for θ is the set $\{\theta : \ R(\theta) \geq p\}$.*

Actually, $\{\theta : R(\theta) \geq p\}$ is not necessarily an interval unless $R(\theta)$ is unimodal, but this is the case for all models that we consider here. The motivation for this approach is that the values of θ that give large values of $L(\theta)$ and hence $R(\theta)$, are the most plausible in light of the data. The main challenge is to decide what p to choose; we show later that choosing $p \in [0.10, 0.15]$ is often useful. If you return to the likelihood function for the Harris/Decima poll (Example 2.2.1) in Figure 2.2, the interval that the pollsters provided, which was 26 ± 2.2 percent, looks like it was constructed such that the values of the likelihood at the endpoints is around $1/10$ of its maximum value so p is between 0.10 and 0.15.

Example 4.3.1 Polls

Let θ be the proportion of people in a large population who have a specific characteristic. Suppose n persons are randomly selected for a poll and y people are observed to have the characteristic of interest. If we let Y be the number who have the characteristic, then $Y \sim \text{Binomial}(n, \theta)$ is a reasonable model. As we have seen previously the likelihood function is

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

and the maximum likelihood estimate of θ is the sample proportion $\hat{\theta} = y/n$. The relative likelihood function is

$$R(\theta) = \frac{\theta^y (1-\theta)^{n-y}}{\hat{\theta}^y (1-\hat{\theta})^{n-y}} \quad \text{for } 0 < \theta < 1.$$

Figure 4.3 shows the relative likelihood functions $R(\theta)$ for two polls:

$$\begin{aligned} \text{Poll 1} &: \ n = 200, \ y = 80 \\ \text{Poll 2} &: \ n = 1000, \ y = 400. \end{aligned}$$

In each case $\hat{\theta} = 0.40$, but the relative likelihood function is more "concentrated" around $\hat{\theta}$ for the larger poll (Poll 2). The 10% likelihood intervals also reflect this:

$$\begin{aligned} \text{Poll 1} &: \ R(\theta) \geq 0.1 \ \text{ for } \ 0.33 \leq \theta \leq 0.47 \\ \text{Poll 2} &: \ R(\theta) \geq 0.1 \ \text{ for } \ 0.37 \leq \theta \leq 0.43. \end{aligned}$$

The graph also shows the log relative likelihood function.

Definition 25 *The log relative likelihood function is*

$$r(\theta) = \log R(\theta) = \log \left[\frac{L(\theta)}{L(\hat{\theta})} \right] = l(\theta) - l(\hat{\theta}) \quad \textit{for } \theta \in \Omega$$

where $l(\theta) = \log L(\theta)$ is the log likelihood function.

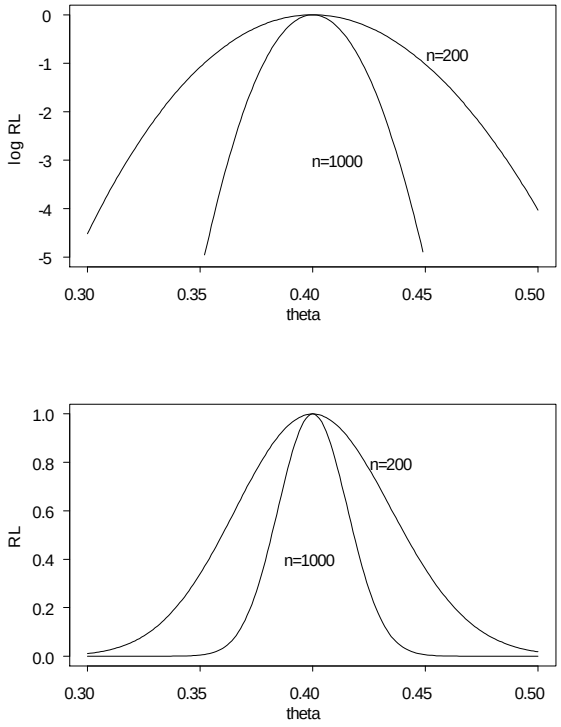


Figure 4.3: **Relative likelihood function and log relative likelihood function for a Binomial model**

It is often more convenient to compute $r(\theta)$ instead of $R(\theta)$ and to compute a $100p\%$ likelihood interval using the fact that $R(\theta) \geq p$ if and only if $r(\theta) \geq \log p$. While both plots are unimodal and have identical locations of the maximum, they differ in terms of the shape. The plot of the relative likelihood function resembles a Normal probability density function in shape while that of the log relative likelihood resembles a quadratic function of θ. Likelihood intervals become narrower as the sample size increases (see, for example, Figure 4.3), which reflects the fact that larger samples contain more information about θ. Likelihood intervals intervals cannot usually be found explicitly. They must be found numerically by using a function like `uniroot` in R or they can be read from graph of $R(\theta)$ or $r(\theta) = \log R(\theta)$.

Table 4.2 gives rough guidelines for interpreting likelihood intervals. *These are only guidelines for this course. The interpretation of a likelihood interval must always be made in the context of a given study.*

Table 4.2: Interpretation of 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 one apparent shortcoming of likelihood intervals so far is that we do not know how probable it is that a given interval will contain the true parameter value. As a result we also do not have a basis for the choice of p. Sometimes it is argued that values like $p = 0.10$ or $p = 0.05$ make sense because they rule out parameter values for which the probability of the observed data is less than 1/10 or 1/20 of the probability when $\theta = \hat{\theta}$. However, a more satisfying approach is to apply the sampling distribution ideas in Section 4.2 to the interval estimates. This leads to the concept of confidence intervals, which we describe next. In Section 4.6 we revisit likelihood intervals and show that they are also confidence intervals.

The idea of a likelihood interval for a parameter θ can also be extended to the case of a vector of parameters $\boldsymbol{\theta}$. In this case $R(\boldsymbol{\theta}) \geq P$ gives likelihood "regions" for $\boldsymbol{\theta}$[22].

4.4 Confidence Intervals and Pivotal Quantities

Suppose we assume that the model chosen for the data $\mathbf{y}$ is correct and that the interval estimate for the parameter θ is given by $[L(\mathbf{y}), U(\mathbf{y})]$. To quantify the uncertainty in the interval estimate we look at an important property of the corresponding interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$ called the *coverage probability* which is defined as follows.

Definition 26 *The value*

$$C(\theta) = P\left[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})\right] \tag{4.1}$$

is called the coverage probability *for the interval estimator* $[L(\mathbf{Y}), U(\mathbf{Y})]$.

A few words are in order about the meaning of the probability statement in (4.1). The parameter θ is an unknown fixed constant associated with the population. It is **not** a random variable and therefore does not have a distribution. The statement (4.1) can be interpreted in the following way. Suppose we were about to draw a random sample of the same size from the same population and the true value of the parameter was θ. Suppose also that we knew that we would construct an interval of the form $[L(\mathbf{y}), U(\mathbf{y})]$ once we

[22] For a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ of unknown parameters, we may want to obtain interval estimates for individual parameters θ_j, $j = 1, 2, \ldots, k$ or for a function $\psi = h(\theta_1, \theta_2, \ldots, \theta_k)$. For example, suppose a model has two parameters θ_1, θ_2 and a likelihood function $L(\theta_1, \theta_2)$ based on observed data. We define the relative likelihood function as $R(\theta_1, \theta_2) = L(\theta_1, \theta_2)/L(\hat{\theta}_1, \hat{\theta}_2)$. The set of pairs (θ_1, θ_2) which satisfy $R(\theta_1, \theta_2) \geq p$ is called a **100p% likelihood region** for (θ_1, θ_2).

had collected the data. Then the probability that θ will be contained in this new interval is $C(\theta)$[23].

How then does $C(\theta)$ assist in the evaluation of interval estimates? In practice, we try to find intervals for which $C(\theta)$ is fairly close to 1 (values 0.90, 0.95 and 0.99 are often used) while keeping the interval fairly narrow. Such interval estimates are called *confidence intervals*[24].

Definition 27 *A* $100p\%$ confidence interval *for a parameter is an interval estimate* $[L(\mathbf{y}), U(\mathbf{y})]$ *for which*

$$P\left[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})\right] = p \tag{4.2}$$

where p *is called the* confidence coefficient.

Suppose $p = 0.95$ and we drew many random samples from the model. Suppose also that each time we observed a random sample, we constructed a 95% confidence interval based on the observed data. Then (4.2) indicates that 95% of these constructed intervals would contain the true value of the parameter θ (and of course 5% do not). This gives us some confidence that for a particular sample, the true value of the parameter is contained in the confidence interval constructed from the sample.

The following example illustrates that the confidence coefficient sometimes does not depend on the unknown parameter θ.

Example 4.4.1 Gaussian distribution with known standard deviation

Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, 1)$ distribution, that is, $\mu = E(Y_i)$ is unknown but $sd(Y_i) = 1$ is known. Consider the interval

$$\left[\overline{Y} - 1.96n^{-1/2}, \overline{Y} + 1.96n^{-1/2}\right]$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the sample mean. Since $\overline{Y} \sim G(\mu, 1/\sqrt{n})$, then

$$\begin{aligned}
&P\left(\overline{Y} - 1.96/\sqrt{n} \leq \mu \leq \overline{Y} + 1.96/\sqrt{n}\right) \\
&= P\left[-1.96 \leq \sqrt{n}\left(\overline{Y} - \mu\right) \leq 1.96\right] \\
&= P\left(-1.96 \leq Z \leq 1.96\right) \\
&= 0.95
\end{aligned}$$

where $Z \sim G(0,1)$. Thus the interval $[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}]$ is a 95% confidence interval for the unknown mean μ. The confidence coefficient was determined without knowing

[23] When we use the observed data y, $L(y)$ and $U(y)$ are numerical values not random variables. We do not know whether $\theta \in [L(y), U(y)]$ or not. $P\left[L(y) \leq \theta \leq U(y)\right]$ makes no more sense than $P\left(1 \leq \theta \leq 3\right)$ since $L(y), \theta, U(y)$ are all numerical values: there is no random variable to which the probability statement can refer.

[24] See the video at www.watstat.com called "What is a confidence interval".

the value of the unknown parameter μ. We will see below that this is a useful property for an interval estimator to have.

We repeat the very important interpretation of a 95% confidence interval (since so many people get the interpretation incorrect!). Suppose the experiment which was used to estimate μ was conducted a large number of times and each time a 95% confidence interval for μ was constructed using the observed data and the interval $[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}]$. Then, approximately 95% of these constructed intervals would contain the true, but unknown value of μ. Since we only have one interval $[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}]]$ we do not know whether it contains the true value of μ or not. We can only say that we are 95% confident that the interval $[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}]$ 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. **Warning:** $P\left(\mu \in [\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}]\right) = 0.95$ is an incorrect statement!!!

If in Example 4.4.1 a particular sample of size $n = 16$ had observed mean $\bar{y} = 10.4$, then the observed 95% confidence interval would be $[\bar{y} - 1.96/4,\ \bar{y} + 1.96/4]$, or $[9.91,\ 10.89]$. We **cannot** say that $P\left(\mu \in [9.91,\ 10.89]\right) = 0.95$. We can **only** say that we are 95% confident that the interval $[9.91,\ 10.89]$ contains μ.

Confidence intervals become narrower as the size of the sample on which they are based increases. For example, note the effect of n in Example 4.4.1. The width of the confidence interval is $2(1.96)/\sqrt{n}$ which decreases as n increases. We noted this earlier for likelihood intervals. We will see in Section 4.6 that likelihood intervals are a type of confidence interval.

Recall that the coverage probability for the interval in Example 4.4.1 did not depend on the unknown parameter μ. This is a highly desirable property because we would like to know the coverage probability without knowing the value of the unknown parameter. We next consider a general method for finding confidence intervals which have this property.

Pivotal Quantities

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* $P\left(Q \geq a\right)$ *and* $P\left(Q \leq b\right)$ *depend on* a *and* b *but not on* θ *or any other unknown information.*

We now describe how a pivotal quantity can be used to construct a confidence interval. We begin with the statement $P\left[a \leq Q(\mathbf{Y}; \theta) \leq b\right] = p$ 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

$$\begin{aligned} p &= P\left[a \leq Q(\mathbf{Y};\theta) \leq b\right] = P\left[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})\right] \\ &= P\left(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})]\right), \end{aligned}$$

the interval $[L(\mathbf{y}), U(\mathbf{y})]$ is a $100p\%$ confidence interval for θ. The confidence coefficient for the interval $[L(\mathbf{y}), U(\mathbf{y})]$ is equal to p which does not depend on θ. The confidence coefficient does depend on a and b, but these are determined by the known distribution of $Q(\mathbf{Y};\theta)$.

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. Since

$$Q = Q(\mathbf{Y};\mu) = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1)$$

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 $0.95 = P(-1.96 \leq Z \leq 1.96)$ where $Z \backsim G(0,1)$. Since $Q(\mathbf{Y};\mu) = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$ we have

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\ &= P\left(\overline{Y} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{Y} + 1.96\sigma/\sqrt{n}\right), \end{aligned}$$

so that

$$\left[\bar{y} - 1.96\sigma/\sqrt{n},\ \bar{y} - 1.96\sigma/\sqrt{n}\right]$$

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 $Z \backsim G(0,1)$ 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 value from the $N(0,1)$ tables. Many two-sided confidence intervals we will encounter will take a similar form.

Exercise: Show that
(*a*) $\bar{y} \pm 1.645\sigma/\sqrt{n}$ is a 90% confidence interval for μ
(*b*) $\bar{y} \pm 2.576\sigma/\sqrt{n}$ is a 99% confidence interval for μ.

Since $P\left(\mu \in [\bar{Y} - 1.645\sigma/\sqrt{n}, \infty)\right) = 0.95$, the interval $[\bar{y} - 1.645\sigma/\sqrt{n}, \infty)$ is also a 95% confidence interval for μ. The interval $[\bar{y} - 1.645\sigma/\sqrt{n}, \infty)$ 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.

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; more accurately, we call Q_n an *approximate pivotal quantity.*

Example 4.4.3 Approximate confidence interval for Binomial model

Suppose $Y \sim \text{Binomial}(n, \theta)$. From the Central Limit Theorem we know that for large n, $Q_1 = (Y - n\theta)/[n\theta(1-\theta)]^{1/2}$ has approximately a $G(0,1)$ distribution. It can also be shown that the distribution of

$$Q_n = Q_n(Y; \theta) = \frac{Y - n\theta}{[n\tilde{\theta}(1-\tilde{\theta})]^{1/2}}$$

where $\tilde{\theta} = Y/n$, is also close to $G(0,1)$ for large n. Q_n is an approximate pivotal quantity which can be used to construct confidence intervals for θ. For example,

$$\begin{aligned} 0.95 &\approx P(-1.96 \le Q_n \le 1.96) \\ &= P\left(\tilde{\theta} - 1.96\left[\tilde{\theta}(1-\tilde{\theta})/n\right]^{1/2} \le \theta \le \tilde{\theta} + 1.96\left[\tilde{\theta}(1-\tilde{\theta})/n\right]^{1/2}\right). \end{aligned}$$

Thus

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \tag{4.3}$$

gives an approximate 95% confidence interval for θ where $\hat{\theta} = y/n$ and y is the observed data.

As a numerical example, suppose we observed $n = 100$, $y = 18$. Then (4.3) gives $0.18 \pm 1.96\,[0.18(0.82)/100]^{1/2}$ or $[0.115,\ 0.255]$.

Remark: It is important to understand that confidence intervals may vary a great deal when we take repeated samples. For example, in Example 4.4.3, ten samples of size $n = 100$

which were simulated for a population with $\theta = 0.25$ gave the following approximate 95% confidence intervals for θ:

$$\begin{array}{lllll} [0.20, 0.38] & [0.14, 0.31] & [0.23, 0.42] & [0.22, 0.41] & [0.18, 0.36] \\ [0.14, 0.31] & [0.10, 0.26] & [0.21, 0.40] & [0.15, 0.33] & [0.19, 0.37] \end{array}$$

For larger samples (larger n), the confidence intervals are narrower and will have better agreement. See Problem 5.

Choosing a Sample Size

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 4.4.5 Sample size and estimation of a Binomial probability

Suppose we want to estimate the probability θ from a Binomial experiment in which $Y \backsim$ Binomial(n, θ) distribution. We use the approximate pivotal quantity

$$Q = \frac{Y - n\theta}{[n\tilde{\theta}(1-\tilde{\theta})]^{1/2}}$$

which was introduced in Example 4.4.3 and which has approximately a $G(0, 1)$ distribution 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

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

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 that $2\,(0.03)$ (or even narrower, say $2\,(0.025)$), we need n large enough so that

$$1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq 0.03$$

or

$$n \geq \left(\frac{1.96}{0.03}\right)^2 \hat{\theta}(1-\hat{\theta}).$$

Of course we don't know what $\hat{\theta}$ is because we have not taken a sample, but we note that the interval is the widest when $\hat{\theta} = 0.5$. So to be conservative, we find n such that

$$n \geq \left(\frac{1.96}{0.03}\right)^2 (0.5)^2 \approx 1067.1$$

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'll 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 $P(|\tilde{\theta} - \theta| \leq 0.03) = 0.95$, 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.

Sample sizes can be similarly determined so as to give confidence intervals of some desired length in other settings. We consider this topic again in Section 4.7 for the $G\,(\mu, \sigma)$ distribution.

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

$$\frac{50000}{6} \times \$20 = \$166,667$$

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

$$\$20 \times \frac{2401}{3} \approx \$16,000$$

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

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad \text{for } x > 0 \tag{4.4}$$

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 4.4 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 $k > 30$, the probability density function resembles that of a $N\,(k, 2k)$ probability density function.*

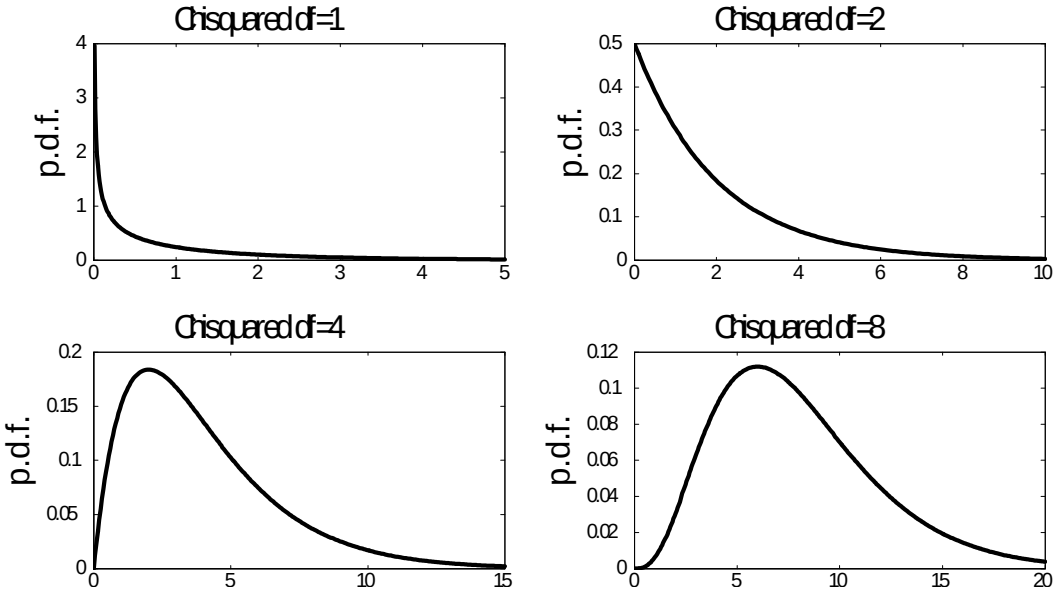


Figure 4.4: **Chi-squared probability density functions for k = 1,2,4,8**

The cumulative distribution function, $F(x;k)$, can be given in closed algebraic form for even values of k. In R the function $dchisq(x,k)$ gives the probability density function $f(x;k)$ and $pchisq(x,k)$ gives the cumulative distribution function $F(x;k)$ for the $\chi^2(k)$ distribution. A table with selected values is given at the end of these course notes.

If $X \backsim \chi^2\,(k)$ then

$$E\,(X) = k \;\text{ and }\; Var(X) = 2k.$$

This result follows by first showing that

$$E\left(X^j\right) = \frac{2^j\Gamma(k/2+j)}{\Gamma(k/2)} \quad \text{for} \quad j = 1, 2, \ldots.$$

This is true since

$$\begin{aligned} E\left(X^j\right) &= \int_0^\infty x^j \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} dx \\ &= \int_0^\infty \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)+j-1} e^{-x/2} dx \quad \text{let } y = x/2 \text{ or } x = 2y \\ &= \int_0^\infty \frac{1}{2^{k/2}\Gamma(k/2)} (2y)^{(k/2)+j-1} e^{-y} 2dy = \frac{2^j}{\Gamma(k/2)} \int_0^\infty y^{(k/2)+j-1} e^{-y} dy \\ &= \frac{2^j\Gamma(k/2+j)}{\Gamma(k/2)}. \end{aligned}$$

Letting $j = 1$ we obtain

$$E(X) = 2\Gamma(k/2+1)/\Gamma(k/2) = 2(k/2) = k.$$

Letting $j = 2$ we obtain

$$E\left(X^2\right) = 2^2\Gamma(k/2+2)/\Gamma(k/2) = 4(k/2+1)(k/2) = k(k+2)$$

and therefore

$$Var(X) = E\left(X^2\right) - [E(X)]^2 = 2k.$$

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 $S = \sum_{i=1}^n W_i \sim \chi^2(\sum_{i=1}^n k_i)$.*

For a proof of this result see Problem 18.

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. Then*

$$P(W \le w) = P(-\sqrt{w} \le Z \le \sqrt{w}) = \Phi(\sqrt{w}) - \Phi(-\sqrt{w}) \quad \text{for } w > 0$$

and the probability density function of W is

$$\begin{aligned}&\frac{d}{dw}\left[\Phi(\sqrt{w})-\Phi(-\sqrt{w})\right]\\&=\left[\phi(\sqrt{w})+\phi(-\sqrt{w})\right]\left(\frac{1}{2}w^{-1/2}\right)\\&=\frac{1}{\sqrt{2\pi}}w^{-1/2}e^{-w/2}\quad \textit{for } w>0\end{aligned}$$

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 $S \backsim \chi^2(n)$.*

Proof. *Since $Z_i \backsim G(0,1)$ then by Theorem 30, $Z_i^2 \backsim \chi^2(1)$ and the result follows by Theorem 29.*

■

The following will be helpful in Chapter 5.

Useful Results:

1. If $W \backsim \chi^2(1)$ then $P(W \geq w) = 2\left[1 - P(Z \leq \sqrt{w})\right]$ where $Z \backsim G(0,1)$.
2. If $W \backsim \chi^2(2)$ then $W \backsim Exponential(2)$ 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 $T \backsim t(k)$ to indicate that the random variable T has a Student t distribution with k degrees of freedom. In Figure 4.5 the probability density function $f(t;k)$ for $k=2$ is plotted together with the $G(0,1)$ probability density function.

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,

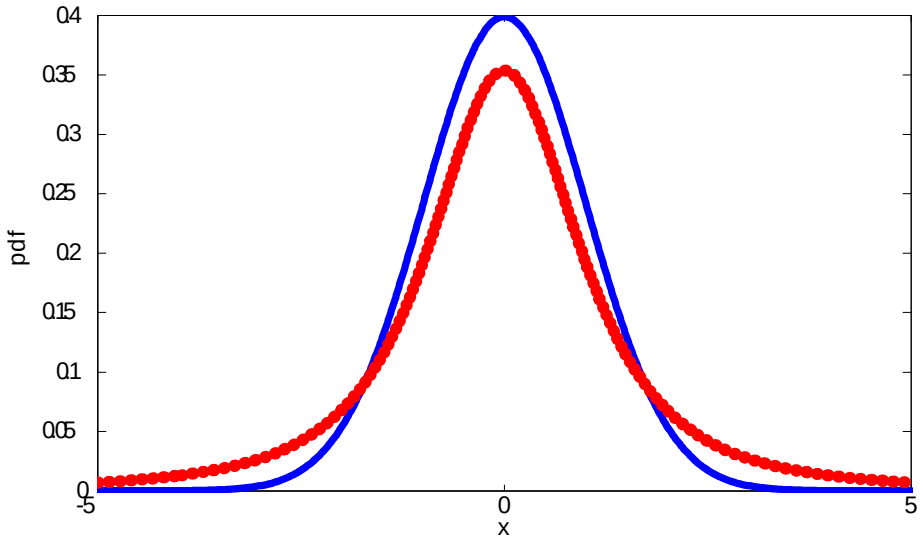


Figure 4.5: **Probability density functions for $t\,(2)$ distribution (dashed red) and $G\,(0,1)$ distribution (solid blue)**

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 20 at the end of this chapter considers some properties of $f(x;k)$.

Probabilities for the t distribution are available from tables at the end of these notes[25] or computer software. In R, the cumulative distribution function $F(t;k) = P(T \leq t;k)$ where $T \sim t\,(k)$ is obtained using *pt(t,k)*. For example, *pt(1.5,10)* gives $P(T \leq 1.5;10) = 0.918$.

The t distribution arises as a result of the following theorem involving the ratio of a $N(0,1)$ random variable and an independent Chi-squared random variable. We will not attempt to prove this theorem here.

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.***

[25] See the video at www.watstat.ca called "Using the t table".

4.6 Likelihood-Based Confidence Intervals

We will now show that likelihood intervals are also confidence intervals. Recall the relative likelihood

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$$

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 $Z \backsim N(0,1)$.*

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

$$\begin{aligned} P(\Lambda(\theta) \leq -2\log p) &= P\left(-2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \leq -2\log p\right) \\ &\approx P(W \leq -2\log p) \quad \text{where } W \backsim \chi^2(1) \\ &= P\left(|Z| \leq \sqrt{-2\log p}\right) \quad \text{where } Z \backsim N(0,1) \\ &= 2P\left(Z \leq \sqrt{-2\log p}\right) - 1. \end{aligned}$$

as required.

■

Example: If $p = 0.1$ then

$$\begin{aligned} q &= 2P\left(Z \leq \sqrt{-2\log(0.1)}\right) - 1 \quad \text{where} \quad Z \backsim G(0,1) \\ &= 2P(Z \leq 2.15) - 1 = 0.96844 \end{aligned}$$

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*

$$p = 2P(Z \leq a) - 1 \quad \textit{where} \quad Z \backsim N(0,1)$$

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 likelihood interval $\left\{\theta : R(\theta) \geq e^{-a^2/2}\right\}$ is

$$\begin{aligned} P\left(\frac{L(\theta)}{L(\tilde{\theta})} \geq e^{-a^2/2}\right) &= P\left(-2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \leq a^2\right) \\ &\approx P\left(W \leq a^2\right) \quad \text{where} \quad W \backsim \chi^2(1) \quad \text{by Theorem 33} \\ &= 2P(Z \leq a) - 1 \quad \text{where} \quad Z \backsim N(0,1) \\ &= p \end{aligned}$$

as required.

■

Example:

Since

$$0.95 = 2P(Z \leq 1.96) - 1 \quad \text{where} \quad Z \backsim N(0,1)$$

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 interval for Binomial model

For Binomial data with n trials and y successes the relative likelihood function is (see Example 4.3.1)

$$R(\theta) = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} \quad \text{for } 0 < \theta < 1.$$

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 4.6 we can read off the 15% likelihood interval which is $[0.31, 0.495]$ which is also an approximate 95% confidence interval.

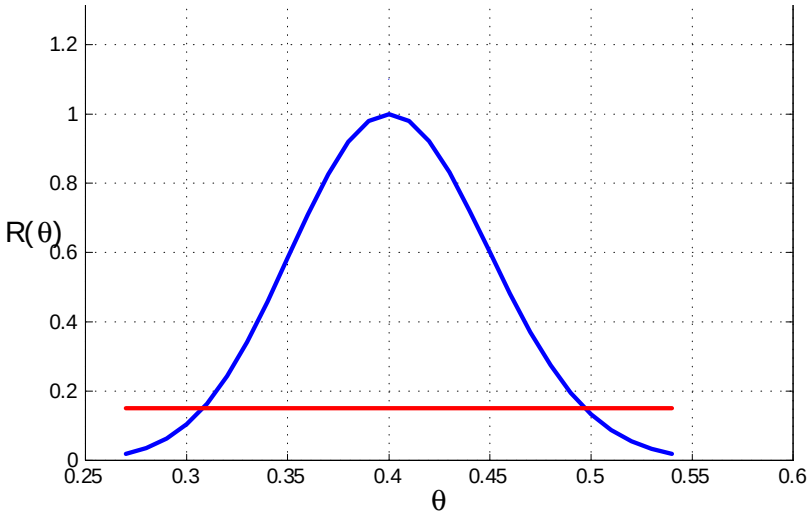


Figure 4.6: **Relative likelihood function for Binomial with** $n = 100$ and $y = 40$

We can compare this interval to the approximate 95% confidence interval based on

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

which gives the interval $[0.304, 0.496]$. The two intervals differ slightly (they are both based on approximations) but are very close.

Suppose $n = 30$ and $\hat{\theta} = 0.1$. From the graph of the relative likelihood function given in Figure 4.7 we can read off the 15% likelihood interval which is $[0.03, 0.24]$ which is also an approximate 95% confidence interval.

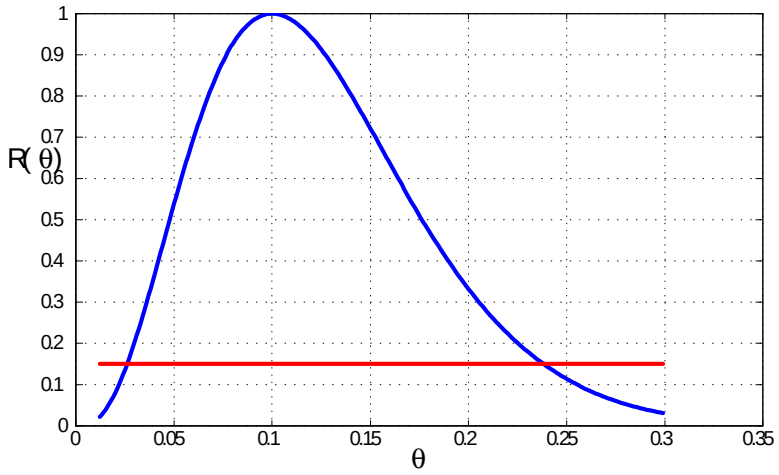


Figure 4.7: **Relative likelihood function for Binomial with** $n = 30$ and $y = 3$

We can compare this to the approximate 95% confidence interval based on

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \tag{4.5}$$

which gives the interval $[-0.0074, 0.2074]$ 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 (4.5) 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 4.7 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. If $\hat{\theta}$ is close to 0.5 or n is large then the likelihood interval will be fairly symmetric and there will be little difference in the two approximate confidence intervals as we saw in the previous example in which n was equal to 100 and $\hat{\theta}$ was equal to 0.4.

4.7 Confidence Intervals for Parameters in the $G(\mu, \sigma)$ Model

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. We have already seen in Section 2.2 that the maximum likelihood estimators of μ and σ^2 are

$$\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and } \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

A closely related point estimator of σ^2 is the sample variance,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

which differs from $\tilde{\sigma}^2$ only by the choice of denominator. Indeed if n is large there is very little difference between S^2 and $\tilde{\sigma}^2$. Note that the sample variance has the advantage that it is an "unbiased" estimator, that is, $E(S^2) = \sigma^2$. This follows since

$$E\left[(Y_i - \mu)^2\right] = Var\left(Y_i\right) = \sigma^2, \;\; E\left[(\bar{Y} - \mu)^2\right] = Var\left(\bar{Y}\right) = \frac{\sigma^2}{n}$$

and

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} \left[(Y_i - \mu) - (\bar{Y} - \mu)\right]^2\right] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (Y_i - \mu)^2 - n(\bar{Y} - \mu)^2\right] \\ &= \frac{1}{n-1} \left\{\sum_{i=1}^{n} E\left[(Y_i - \mu)^2\right] - nE\left[(\bar{Y} - \mu)^2\right]\right\} \\ &= \frac{1}{n-1} \left[n\sigma^2 - n\left(\frac{\sigma^2}{n}\right)\right] = \frac{1}{n-1}\left[(n-1)\,\sigma^2\right] \\ &= \sigma^2. \end{aligned}$$

We now consider interval estimation for μ and σ.

Confidence Intervals for μ

If σ were known then

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1) \tag{4.6}$$

would be a pivotal quantity that could 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 variance S in Z, then we still have a pivotal quantity. We will write the pivotal quantity in terms of S. The pivotal quantity is

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \tag{4.7}$$

Since S, unlike σ, is a random variable in (4.7) 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 $\bar{Y}$ and sample variance S^2. Then*

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \backsim t\,(n-1)\,. \tag{4.8}$$

To see how this result follows from Theorem 32 let

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1)$$

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[26]. By Theorem 32 with $k = n-1$, we have

$$\frac{Z}{\sqrt{U/k}} = \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t\,(n-1)\,.$$

In other words if we replace σ in the pivotal quantity (4.6) 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 are inherited from the degrees of freedom of the Chi-squared random variable U or from S^2.

We now show how to use the t distribution to obtain a confidence interval for μ when σ is unknown. Since (4.8) has a t distribution with $n-1$ degrees of freedom which is a completely known distribution, we can use this pivotal quantity to construct a $100p\%$ confidence interval for μ. Since the t distribution is symmetric we determine the constant a such that $P(-a \leq T \leq a) = p$ using the t tables provided in these course notes or R. Note that, due to symmetry, $P(-a \leq T \leq a) = p$ is equivalent to $P\,(T \leq a) = (1+p)\,/2$ (you should verify this) and since the t tables tabulate the cumulative distribution function $P\,(T \leq t)$, it is easier to find a such that $P\,(T \leq a) = (1+p)\,/2$. Then since

$$\begin{aligned} p &= P(-a \leq T \leq a) \\ &= P\left(-a \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq a\right) \\ &= P\left(\bar{Y} - aS/\sqrt{n} \leq \mu \leq \bar{Y} + aS/\sqrt{n}\right) \end{aligned}$$

a $100p\%$ confidence interval for μ is given by

$$\left[\bar{y} - as/\sqrt{n},\ \bar{y} + as/\sqrt{n}\right]. \tag{4.9}$$

(Note that if we attempted to use (4.6) 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.

[26] The proof of the remarkable result that, for a random sample from a Normal distribution, the sample mean and the sample variance are independent random variables, is beyond the scope of this course.

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.

Example 4.7.1 IQ test

Scores Y for an IQ test administered to ten year old children in a very large population have close to a $G(\mu, \sigma)$ distribution. A random sample of 10 children in a particular large inner city school obtained test scores as follows:

$$103,\ 115,\ 97,\ 101,\ 100,\ 108,\ 111,\ 91,\ 119,\ 101$$

$$\sum_{i=1}^{10} y_i = 1046 \ \text{ and } \ \sum_{i=1}^{10} y_i^2 = 110072.$$

We wish to use these data to estimate the parameter μ which represents the mean test score for ten year old children at this school. Since

$$P(T \leq 2.262) = 0.975 \quad \text{for } T \sim t\,(9)\,,$$

a 95% confidence interval for μ based on (4.9) is $\bar{y} \pm 2.262s/\sqrt{10}$ or

$$\left[\bar{y} - 2.262s/\sqrt{10},\ \bar{y} + 2.262s/\sqrt{10}\right].$$

For the given data $\bar{y} = 104.6$ and $s = 8.57$, so the confidence interval is 104.6 ± 6.13 or $[98.47, 110.73]$.

Behaviour as $n \to \infty$

As 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. Second as the degrees of freedom increase, the t distribution approaches the Gaussian so that the quantiles of the t distribution approach that of the $G(0,1)$ distribution. For example, if in Example 4.7.1 we knew that $\sigma = 8.57$ then we would use the 95% confidence interval $\bar{y} \pm 1.96\,(8.57)\,/\sqrt{n}$ instead of $\bar{y} \pm 2.262\,(8.57)\,/\sqrt{n}$ with $n = 10$. In general for large n, the width of the confidence interval gets narrower as n increases (but at the rate $1/\sqrt{n}$) so the confidence intervals shrink to include only the point $\bar{y}$.

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 Normal 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 Normal 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 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, which we will not prove.

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. Then the random variable*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{4.10}$$

has a Chi-squared distribution with $n-1$ degrees of freedom.

While we will not prove this result, we should at least try to explain the puzzling number of degrees of freedom $n-1$, which, at first glance, seems wrong since $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is the sum of n squared Normal random variables. Does this contradict Corollary 31? It is true that each $W_i = (Y_i - \bar{Y})$ 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 19.) It is easy to see that $W_1, W_2, \ldots, W_n$ are not independent random variables since $\sum_{i=1}^{n} W_i = 0$ 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, $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} W_i^2$ 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 both of the Chi-squared and of the t distribution. In both cases, the degrees of freedom are inherited from S^2 and are related to the dimension of the subspace inhabited by the terms in the sum for S^2, that is, $W_i = Y_i - \bar{Y}$, $i = 1, 2, \ldots, n$.

We will now show how we can use Theorem 37 to construct a $100p\%$ confidence interval for the parameter σ^2 or σ. First note that (4.10) is a pivotal quantity since its distribution is completely known. Using Chi-squared tables 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

$$\begin{aligned} p &= P(a \leq U \leq b) \\ &= P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) \\ &= P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) \\ &= P\left(\sqrt{\frac{(n-1)S^2}{b}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{b}}\right) \end{aligned}$$

a $100p\%$ confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{b},\ \frac{(n-1)s^2}{a}\right] \tag{4.11}$$

and a $100p\%$ confidence interval for σ is

$$\left[\sqrt{\frac{(n-1)s^2}{b}},\ \sqrt{\frac{(n-1)s^2}{a}}\right]. \tag{4.12}$$

The choice for a, b is not unique. For convenience, a and b are usually chosen such that

$$P(U \leq a) = P(U > b) = \frac{1-p}{2} \tag{4.13}$$

where $U \sim \chi^2(n-1)$. Note that since the Chi-squared tables provided in these course notes tabulate the cumulative distribution function, $P(U \leq u)$, this means using the tables 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 (4.11) and (4.12) are called *equal-tailed* confidence intervals. The choice (4.13) 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

$$\left[0,\ \sqrt{\frac{(n-1)s^2}{a}}\right].$$

Example 4.7.2 Optical glass lenses

A manufacturing process produces wafer-shaped pieces of optical glass for lenses. Pieces must be very close to 25 mm thick, and only a small amount of variability around this can be tolerated. If Y represents the thickness of a randomly selected piece of glass then, to a close approximation, $Y \sim G(\mu, \sigma)$. The parameter σ represents the standard deviation of the population of lens thicknesses produced by this manufacturing process. Periodically, random samples of $n = 15$ pieces of glass are selected and the values of μ and σ are estimated to see if they are consistent with $\mu = 25$ and with σ being under 0.02 mm. On one such occasion the observed data were

$$\bar{y} = 25.009,\ s = 0.013.$$

From the sample standard deviation we can determine

$$\sum_{i=1}^{15}(y_i - \bar{y})^2 = (14)\, s^2 = 0.002347.$$

To obtain a 95% confidence interval for σ we determine a and b such that

$$P\,(U \leq a) = \frac{1 - 0.95}{2} = 0.025 \ \text{ and } \ P\,(U \leq b) = \frac{1 + 0.95}{2} = 0.975$$

where $U \sim \chi^2\,(14)$. From Chi-squared tables or R we obtain

$$P\,(U \leq 5.63) = 0.025 \ \text{ and } \ P\,(U \leq 26.12) = 0.975$$

so $a = 5.63$ and $b = 26.12$. Substituting these values along with $(14)\, s^2 = 0.002347$ into (4.12) we obtain

$$\left[\sqrt{\frac{0.002347}{26.12}},\ \sqrt{\frac{0.002347}{5.63}}\right] = [0.0095,\ 0.0204]\,.$$

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 $P(6.57 \leq U < \infty) = 0.95$ we can obtain a one-sided 95% confidence interval for σ which is given by

$$\left[0, \sqrt{\frac{(n-1)\, s^2}{a_1}}\right] = \left[0, \sqrt{\frac{0.002347}{6.57}}\right] = [0, 0.0189]$$

and the value 0.02 is not in the interval. Why are the intervals different? Both cover the true value of the parameter σ for 95% of all samples so they have the same confidence coefficient. However the one-sided interval, since it allows smaller (as small as zero) values on the left end of the interval, can achieve the same coverage with a smaller right end-point. If our primary concern was for values of σ being too large, that is, for an upper bound for the interval, then the one-sided interval is the one that should be used for this purpose.

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 effects 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 an 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 $\hat{\mu} = \bar{y}$ and s respectively. Our best point estimate of Y based on the data we have already observed is $\hat{\mu}$ with corresponding estimator $\tilde{\mu} = \bar{Y} \backsim N\left(\mu, \sigma^2/n\right)$. To obtain an interval of values for Y we note that $Y \backsim G(\mu, \sigma)$ independently of $\tilde{\mu} = \bar{Y} \backsim n\left(\mu, \sigma^2/n\right)$. Since $E(Y - \tilde{\mu}) = \mu - \mu = 0$ and $Var(Y - \tilde{\mu}) = \sigma^2 + \sigma^2/n$ therefore

$$Y - \tilde{\mu} = Y - \bar{Y} \backsim N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right).$$

Also

$$\frac{Y - \bar{Y}}{S\sqrt{1 + \frac{1}{n}}} \backsim t(n-1)$$

is a pivotal quantity which can be used to obtain an interval of values for Y. Let a be the value such that $P(-a \leq T \leq a) = p$ or $P(T \leq a) = (1+p)/2$ which is obtained from t tables or by using R. Since

$$\begin{aligned} p &= P(-a \leq T \leq a) \\ &= P\left(-a \leq \frac{Y - \bar{Y}}{S\sqrt{1 + \frac{1}{n}}} \leq a\right) \\ &= P\left(\bar{Y} - aS\sqrt{1 + \frac{1}{n}} \leq Y \leq \bar{Y} + aS\sqrt{1 + \frac{1}{n}}\right) \end{aligned}$$

therefore

$$\left[\bar{y} - as\sqrt{1 + \frac{1}{n}},\ \bar{y} + as\sqrt{1 + \frac{1}{n}}\right] \tag{4.14}$$

is an interval of values for the future observation Y with confidence coefficient p. The interval (4.14) 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 (4.14) 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 $Var(Y) = \sigma^2$.

Example 4.7.2 Revisited Optical glass lenses

Suppose in Example 4.7.2 a 95% prediction interval is required for a glass lens drawn at random from the population of glass lenses. Now $\bar{y} = 25.009$, $s = 0.013$ and for $T \sim t(14)$ we have $P(T \leq 2.1448) = (1 + 0.95)/2 = 0.975$. Therefore a 95% prediction interval for this new lens is given by

$$\begin{aligned} & \left[25.009 - 2.1448\,(0.013)\sqrt{1 + \frac{1}{15}},\ 25.009 + 2.1448\,(0.013)\sqrt{1 + \frac{1}{15}}\right] \\ = \ & [25.009 - 0.0288, 25.009 + 0.0288] \\ = \ & [24.980, 25.038]\,. \end{aligned}$$

Note that this interval is much wider than a 95% confidence interval for $\mu =$ the mean of the population of lens thicknesses produced by this manufacturing process which is given by

$$\begin{aligned} & \left[25.009 - 2.1448\,(0.013)/\sqrt{15},\ 25.009 + 2.1448\,(0.013)/\sqrt{15}\right] \\ = \ & [25.009 - 0.007, 25.009 + 0.007] \\ = \ & [25.002, 25.016]\,. \end{aligned}$$

4.8 A Case Study: Testing Reliability of Computer Power Supplies[27]

Components of electronic products often must be very reliable, that is, they must perform over long periods of time without failing. Consequently, manufacturers who supply components to a company that produces, e.g. personal computers, must satisfy the company that their components are reliable.

Demonstrating that a component is highly reliable is difficult because if the component is used under "normal" conditions it will usually take a very long time to fail. It is generally not feasible for a manufacturer to carry out tests on components that last for years (or even months, in most cases) and therefore they use what are called *accelerated life tests*. These involve placing high levels of stress on the components so that they fail in much less than the normal time. If a model relating the level of stress to the lifetime of the component is

[27] Optional

known then such experiments can be used to estimate lifetime at normal stress levels for the population from which the experimental units are taken.

Table 4.3: Lifetimes (in hours) from an accelerated life test experiment

$70°C$	$60°C$	$50°C$	$40°C$
2	1	55	78
5	20	139	211
9	40	206	297
10	47	263	556
10	56	347	600*
11	58	402	600*
64	63	410	600*
66	88	563	600*
69	92	600*	600*
70	103	600*	600*
71	108	600*	600*
73	125	600*	600*
75	155	600*	600*
77	177	600*	600*
97	209	600*	600*
103	224	600*	600*
115	295	600*	600*
130	298	600*	600*
131	352	600*	600*
134	392	600*	600*
145	441	600*	600*
181	489	600*	600*
242	600*	600*	600*
263	600*	600*	600*
283	600*	600*	600*

Notes: Lifetimes are given in ascending order; asterisks(*) denote censored observations.

We consider below some life test experiments on power supplies for personal computers, with ambient temperature being the stress factor. As the temperature increases, the lifetimes of components tend to decrease and at a temperature of around 70° Celsius the average lifetimes tend to be of the order of 100 hours. The normal usage temperature is around 20° C. The data in Table 4.3 show the lifetimes (i.e. times to failure) y_i of components tests at each of 40°, 50°, 60° and 70° C. The experiment was terminated after 600 hours and for temperatures 40°, 50° and 60° some of the 25 components being tested had still not failed. Such observations are called *censored observations*: we only know in each

case that the lifetime in question was over 600 hours. In Table 4.3 the asterisks denote the censored observations. Note the data have been organized so that the lifetimes are listed first followed by the censored times.

It is known from past experience that, at each temperature level, lifetimes are approximately Exponentially distributed; let us therefore suppose that at temperature t, $(t = 40, 50, 60, 70)$, component lifetimes Y have an Exponential distribution with probability density function

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

where $E(Y) = \theta_t$ is the mean lifetime of components subjected to temperature t.

We begin by determining the likelihood function for the experiment at $t = 60°$. The data are $y_1, y_2, \ldots, y_{25}$ where we note that $y_{23} = 600$, $y_{24} = 600$, $y_{25} = 600$ are censored observations. We assume these data arise from an Exponential(μ) distribution where we let $\mu = \theta_{60}$ for the moment for convenience.

The contribution to the likelihood function for an observed lifetime y_i is

$$f(y_i; \mu) = \frac{1}{\mu} e^{-y_i/\mu}.$$

For the censored observations we only know that the lifetime is greater than 600. Since

$$\begin{aligned} P(Y; \mu) &= P(Y > 600; \mu) \\ &= \int_{600}^{\infty} \frac{1}{\mu} e^{-y/\mu} dy \\ &= e^{-600/\mu} \end{aligned}$$

the contribution to the likelihood function of each observation censored at 600 is

$$e^{-600/\mu}.$$

Therefore the likelihood function for μ based on the data $y_1, y_2, \ldots, y_{25}$ is

$$\begin{aligned} L(\mu) &= \left[\prod_{i=1}^{22} \frac{1}{\mu} e^{-y_i/\mu}\right] \left[\prod_{i=23}^{25} e^{-600/\mu}\right] \\ &= \mu^{-k} e^{-s/\mu} \quad \text{for } \mu > 0 \end{aligned}$$

where $k = 22 =$ the number of uncensored observations and $s = \sum_{i=1}^{22} y_i + 3(600) = 5633 =$ sum of all lifetimes and censored times.

Question 1 Show that the maximum likelihood estimate of μ is given by $\hat{\mu} = s/k$ and thus $\hat{\theta}_{60} = s/k$.

Question 2 Assuming that the Exponential model is correct, the likelihood function for θ_t, $t = 40, 50, 60, 70$ can be obtained using the method above and is given by

$$L(\theta_t) = \theta_t^{-k_t} e^{-s_t/\theta_t}$$

where k_t = number of uncensored observations at temperature t and s_t = sum of all lifetimes and censored times at temperature t.

Find the maximum likelihood estimates of $\hat{\theta}_t$, $t = 40, 50, 60, 70$. Graph the relative likelihood functions for θ_{40} and θ_{70} on the same graph and comment on any qualitative differences.

Question 3 Graph the empirical cumulative distribution function for $t = 40$. Note that, due to the censoring, the empirical cumulative distribution function $\hat{F}(y)$ is constant and equal to one for $y \geq 600$. On the same plot graph the cumulative distribution function for an Exponential($\hat{\theta}_{40}$). What would you conclude about the fit of the Exponential model for $t = 40$? Repeat this exercise for $t = 50$. What happens if you use this technique to check the Exponential model for $t = 60$ and 70?

Questions 4 Engineers use a model (called the Arrhenius model) that relates the mean lifetime of a component to the ambient temperature. The model states that

$$\theta_t = \exp\left(\alpha + \frac{\beta}{t + 273.2}\right) \tag{4.15}$$

where t is the temperature in degrees Celsius and α and β are parameters. Plot the points $\left(\log \hat{\theta}_t,\ (t + 273.2) - 1\right)$ for $t = 40, 50, 60, 70$. If the model is correct why should these points lie roughly along a straight line? Do they?

Using the graph give rough point estimates of α and β. Extrapolate the line or use your estimates of α and β to estimate θ_{20}, the mean lifetime at $t = 20°$ C which is the normal operating temperature.

Question 5 Question 4 indicates how to obtain a rough point estimate of

$$\theta_{20} = \exp\left(\alpha + \frac{\beta}{20 + 273.2}\right).$$

Suppose we wanted to find the maximum likelihood estimate of θ_{20}. This would require the maximum likelihood estimates of α and β which requires the joint likelihood function of α and β. Explain why this likelihood is given by

$$L(\alpha, \beta) = \prod_{t=40}^{70} \theta_t^{-k_t} e^{-s_t/\theta_t}$$

where θ_t is given by (4.15). (Note that the product is only over $t = 40, 50, 60, 70$.) Outline how you might attempt to get an interval estimate for θ_{20} based on the likelihood function for α and β. If you obtained an interval estimate for θ_{20}, would you have any concerns about indicating to the engineers what mean lifetime could be expected at 20°C? (Explain.)

Question 6 Engineers and statisticians have to *design* reliability tests like the one just discussed, and considerations such as the following are often used:

Suppose that the mean lifetime at 20°C is supposed to be about 90,000 hours and that at 70°C you know from past experience that it is about 100 hours. If the model (4.15) holds, determine what α and β should be approximately and thus what θ is roughly equal to at 40°, 50° and 60°C. How might you use this information in deciding how long a period of time to run the life test? In particular, give the approximate expected number of uncensored lifetimes from an experiment that was terminated after 600 hours.

4.9 Chapter 4 Problems

1. 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 $\theta \in [0, 1]$.

2. The following code 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 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.

3. 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 $R(\theta) = 0.1$ which can be used to determine a 10% 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(theta,Rtheta,type="l") # plots Relative Likelihood Function
# draw a horizontal line at 0.10
abline(a=0.1,b=0,col="red",lwd=2)
title(main="Binomial Likelihood for y=15 and n=40")
```

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 10% likelihood interval.

4. 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 $R(\theta) = 0.1$ which can be used to determine a 10% 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(theta,Rtheta,type="l")
# draw a horizontal line at 0.10
abline(a=0.1,b=0,col="red",lwd=2)
title(main="Poisson Likelihood for ybar=5 and n=25")
```

Modify this code for larger sample sizes $n = 100$, $n = 400$ and observe what happens to the width of the 10% likelihood interval.

5. The following R Code generates k approximate Binomial(n, θ) confidence intervals 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-thetahat)<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.

6. 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 an approximate 95% confidence interval is less than 2 (0.02)?

7. In the U.S.A. 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.

8. 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).

9. For Chapter 2, Problem 3 (b) determine an approximate 95% confidence interval for θ by using 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.

10. For Chapter 2, Problem 5(b) determine an approximate 95% confidence interval for θ by using 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.

11. For Chapter 2, Problem 7(b) determine an approximate 95% confidence interval for θ by using 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.

12. Recall Chapter 2, Problem 8.

    (a) 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. Is α very accurately determined?

    (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 the accuracy of estimation in the two cases.

13. For Chapter 2, Problem 10(c) determine an approximate 95% confidence interval for θ by using 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.

14. 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 $\hat{\theta}$ 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.

15. A manufacturing process produces fibers of varying lengths. The length of a fiber Y is a continuous random variable with probability density function

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

    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 $E\left(\bar{Y}\right)$ and $Var\left(\bar{Y}\right)$ using the result in (a).

    (d) Justify the statement

$$P\left(-1.96 \leq \frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \leq 1.96\right) \approx 0.95.$$

    (e) Explain 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).

16. 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?

17. **The Chi-squared distribution:**

(a) Determine the following using the χ^2 tables provided in the Course Notes:

(i) If $X \backsim \chi^2(10)$ find $P(X \leq 2.6)$ and $P(X > 16)$.

(ii) If $X \backsim \chi^2(4)$ find $P(X > 15)$.

(iii) If $X \backsim \chi^2(40)$ 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 $Y \backsim N(40, 80)$.

(iv) If $X \backsim \chi^2(25)$ find a and b such that $P(X \leq a) = 0.025$ and $P(X > b) = 0.025$.

(v) If $X \backsim \chi^2(12)$ 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 χ^2 tables:

(i) If $X \backsim \chi^2(1)$ find $P(X \leq 2)$ and $P(X > 1.4)$.

(ii) If $X \backsim \chi^2(2)$ find $P(X \leq 2)$ and $P(X > 3)$.

(d) If $X \backsim G(3,2)$ and $Y_i \backsim Exponential(2)$, $i = 1, 2, \ldots, 5$ all independently then what is the distribution of $W = \sum_{i=1}^{5} Y_i + \left(\frac{X-3}{2}\right)^2$?

(e) If $X_i \backsim \chi^2(i)$, $i = 1, 2, \ldots, 10$ independently then what is the distribution of $\sum_{i=1}^{10} X_i$?

18. **Properties of the Chi-squared distribution:** Suppose $X \sim \chi^2(k)$ with probability density function given by

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad \text{for } y > 0.$$

(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

$$M(t) = E\left(e^{tX}\right) = (1-2t)^{-k/2} \quad \text{for } t < \frac{1}{2}$$

and use this to show that $E(X) = k$ and $Var(X) = 2k$.

(d) Prove Theorem 29 using moment generating functions.

19. Suppose $Y_i \backsim N\left(\mu, \sigma^2\right)$, $i = 1, 2, \ldots, n$ independently and let $W_i = Y_i - \bar{Y}$, $i = 1, 2, \ldots, n$.

(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 $Cov\left(Y_i, \bar{Y},\right) = \frac{1}{n}$, $i = 1, 2, \ldots, n$. Note that this result along with the result in (a) implies that

$$W_i = Y_i - \bar{Y} \backsim N\left(0, \sigma^2\left(1 - \frac{1}{n}\right)\right), \quad i = 1, 2, \ldots, n.$$

(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.

20. **Student's t distribution:** Suppose $T \backsim t(k)$.

    (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 tables provided in the Course Notes to answer the following:

    (i) If $T \backsim t(10)$ find $P(T \leq 0.88)$, $P(T \leq -0.88)$ and $P(|T| \leq 0.88)$.

    (ii) If $T \backsim t(17)$ find $P(|T| > 2.90)$.

    (iii) If $T \backsim t(30)$ 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 $Z \backsim N(0,1)$.

    (iv) If $T \backsim t(18)$ find a and b such that $P(T \leq a) = 0.025$ and $P(T > b) = 0.025$.

    (v) If $T \backsim t(13)$ 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).

21. **Limiting t distribution:** Suppose $T \backsim t(k)$ 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.

22. 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.

    (a) Assuming that survival times are Exponentially distributed with mean θ days, graph the relative likelihood function for these data and obtain an approximate 90% confidence interval for θ. This interval may be obtained from the graph of the relative likelihood function or by using the function `uniroot` in R.

    (b) Show that $m = \theta \ln 2$ is the median survival time. Using the interval obtained in (a), give an approximate 90% confidence interval for m.

23. **Exact confidence intervals for θ for Exponential data:**

    (a) If $Y \backsim$ Exponential(θ) then show that $W = 2Y/\theta$ has a $\chi^2(2)$ distribution. (Hint: compare the probability density function of W with (4.4)).

    (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

    $$U = 2\sum_{i=1}^{n} Y_i/\theta \sim \chi^2(2n).$$

    This result implies that U is a pivotal quantity which can be used to obtain confidence intervals for θ.

    (c) Refer to the data in the previous problem. Using the fact that

    $$P(43.19 \leq W \leq 79.08) = 0.90$$

    where $W \sim \chi^2(60)$ obtain a 90% confidence interval for θ based on U. Compare this with the approximate confidence interval for θ obtained in the previous problem.

24. 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 which is also an approximate 95% confidence 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?

(d) What conditions would need to be satisfied to make the assumptions and analysis in (a) to (c) valid?

(e) If $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Poisson(θ) distribution then the random variable

$$\frac{\bar{Y} - \theta}{\sqrt{\bar{Y}/n}}$$

has approximately a $N(0,1)$ distribution. Show how this result leads to an approximate 95% confidence interval for θ given by

$$\bar{y} \pm 1.96\sqrt{\frac{\bar{y}}{n}}.$$

Using this result determine the approximate 95% confidence intervals for each company based on the result. Compare these intervals with the intervals obtained in (c).

25. 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 $Y_i \backsim G(\mu, \sigma)$, $i = 1, 2, \ldots, 19$ 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.

26. 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 Normal model for the female heights and a different Normal 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?

27. 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 $Y_i \backsim G(\mu, \sigma)$, $i = 1, 2, \ldots, 12$ 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.

28. 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 placed 12 detectors 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

$$\sum_{i=1}^{12} y_i = 1249.6 \quad \text{and} \quad \sum_{i=1}^{12} (y_i - \bar{y})^2 = 971.43.$$

To analyze these data assume the model $Y_i \backsim G(\mu, \sigma)$, $i = 1, 2, \ldots, 12$ independently where μ and σ are unknown parameters. University researchers obtained a $13th$ 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.

29. A manufacturer wishes to determine the mean breaking strength (force) μ of a type of string to "within 0.5 kilograms", which we interpret as requiring that the 95% confidence interval for a μ should have length at most 1 kilogram. If breaking strength Y of strings tested are $G(\mu, \sigma)$ and if 10 preliminary tests gave $\sum_{i=1}^{10}(y_i - \bar{y})^2 = 45$, how many additional measurements would you advise the manufacturer to take?

30. 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 Normal distribution with mean μ and variance σ_1^2, whereas for method II, measurements $Y_1, Y_2, \ldots, Y_n$ have a Normal distribution with mean μ and variance σ_2^2.

    (a) Assuming that σ_1^2 and σ_2^2 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

    $$\tilde{\mu} = \frac{w_1\bar{X} + w_2\bar{Y}}{w_1 + w_2}$$

    and the estimator $(\bar{X} + \bar{Y})/2$. Why is $\tilde{\mu}$ a better estimator?

31. **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.

32. **Challenge Problem:** A sequence of random variables $\{X_n\}$ is said to *converge in probability* to the constant c if for all $\epsilon > 0$,

    $$\lim_{n\to\infty} P(|X_n - c| \geq \epsilon) = 0$$

    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 $\tilde{\theta}_n = \max(X_1, X_2, \ldots, X_n)$ is consistent for θ.

(ii) Let $X \sim$ Binomial(n, θ). Show that $\tilde{\theta}_n = X/n$ is consistent for θ.

33. **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?

34. **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

$$Z_n = \sum_{i=1}^{n} \frac{(Y_i - \mu)}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma}$$

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 $G(0, 1)$. (Hint: For any real number a, $(1 + a/n)^n \to e^a$ as $n \to \infty$.)

5. TESTS OF HYPOTHESES

5.1 Introduction

[31]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 53 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?

When studying tests of hypotheses 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.

Statistical tests of hypotheses are 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, which may not always be specified. In many cases the alternative hypothesis is

[31]See the video at www.watstat.ca called "A Test of Significance".

simply that H_0 *is not true.*

We will outline the logic of tests of hypotheses 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 easily in the our given 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 hypothesis tests. 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 and there is no evidence to support (2).

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: Interpretation of $p-values$

$p-value$	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

$$\begin{aligned} p - value &= P\left(|Y - 100| \geq 10\right) \quad \text{where } Y \sim \text{Binomial}(200, 0.5) \\ &= P\left(\frac{|Y - 100|}{\sqrt{200\,(0.5)\,(0.5)}} \geq \frac{10}{\sqrt{200\,(0.5)\,(0.5)}}\right) \\ &\approx P\left(|Z| \geq 1.41\right) \quad \text{where } \; Z \backsim N\left(0, 1\right) \\ &= 2\left[1 - P\left(Z \leq 1.41\right)\right] \\ &= 2\left(1 - 0.92073\right) \\ &= 0.15854 \end{aligned}$$

so there is no evidence against the hypothesis that I was guessing.

Example 5.1.2 Test of hypothesis for Binomial

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, $Y \backsim \text{Binomial}(n, 1/6)$ distribution. A reasonable test statistic would then be either $D_1 = |Y - n/6|$ or (if we wanted to focus on the possibility that θ was bigger than $1/6$), $D = \max\left[(Y - n/6), 0\right]$.

Suppose that $n = 180$ tosses gave $y = 44$. Using $D = \max\left[(Y - n/6), 0\right]$, we get $d = \max\left[(44 - 180/6), 0\right] = 14$. The $p - value$ (calculated using R) is

$$\begin{aligned} p - value &= P(D \geq 14; H_0) \\ &= P(Y \geq 44) \quad \text{where } Y \sim \text{Binomial}(180, 1/6) \\ &= \sum_{y=44}^{180} \binom{180}{y} \left(\frac{1}{6}\right)^{y} \left(\frac{5}{6}\right)^{180-y} \\ &= 0.005 \end{aligned}$$

which provides strong evidence against H_0, and suggests that θ is bigger than $1/6$. This is an example of a one-sided test which is described in more detail below.

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 in light of the observed data.

Note that we do **not** claim that H_0 is true, only that there is no evidence in light of the 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.

A final point is that once we have specified a test statistic D, we need to be able to compute the $p - value$ for the observed data. Calculating probabilities involving D brings us back to distribution theory. In most cases the exact $p - value$ is difficult to determine mathematically, and we must use either an approximation or computer simulation. Fortu-

nately, for the tests considered in Section 5.3 we can use an approximation based on the Chi-squared distribution.

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 Tests of Hypotheses 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

$$\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

As usual we prefer to use the sample variance estimator

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

to estimate σ^2.

Recall from Chapter 4 that

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \backsim t\,(n-1)\,.$$

We use this pivotal quantity to construct a test of hypothesis for the parameter μ when the standard deviation σ is unknown.

Hypothesis Tests for μ

For a Normally distributed population, we may wish to test a hypothesis $H_0 : \mu = \mu_0$, where μ_0 is some specified value. The values of the parameter to be considered when H_0 is not true are called the alternative hypothesis and denoted by H_A. If the alternative hypothesis is $H_A : \mu \neq \mu_0$ then the test statistic

$$D = \frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \tag{5.1}$$

makes intuitive sense. We obtain the $p - value$ using the fact that

$$\frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \backsim t\,(n-1)$$

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

$$\begin{aligned} p-value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(|T| \geq d) \quad \text{where } T \sim t\,(n-1) \\ &= 1 - P(-d \leq T \leq d) \\ &= 2\left[1 - P\left(T \leq d\right)\right]. \end{aligned}$$

One-sided hypothesis tests

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

$$D = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

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

$$\begin{aligned} p-value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(T \geq d) \\ &= 1 - P\left(T \leq d\right) \quad \text{where } T \sim t\,(n-1). \end{aligned}$$

This type of example is often called a one-sided test of hypothesis.

In Example 5.1.2, the hypothesis of interest was $H_0 : \theta = 1/6$ where θ was the probability that the upturned face was a one. If the alternative of interest is that θ is not equal to $1/6$ then the alternative hypothesis is $H_A : \theta \neq 1/6$ and the test statistic $D = |Y - n/6|$ is a good choice. If the alternative of interest is that θ is bigger than $1/6$ then the alternative hypothesis is $H_A : \theta > 1/6$ and the test statistic $D = \max\left[(Y - n/6), 0\right]$ is a better choice.

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

$$\begin{array}{lll} A: & \bar{y} = 1.0061, & s = 0.0230, \quad d = 0.839 \\ B: & \bar{y} = 0.9810, & s = 0.0170, \quad d = 3.534. \end{array}$$

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 tables 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 tables 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

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{10}} \sim t\left(9\right).$$

For t tables we have $P(T \leq 2.2622) = 0.975$ and a 95% confidence interval for μ is given by

$$\bar{y} \pm 2.2622 s/\sqrt{10} = 0.981 \pm 0.012 \quad \text{or} \quad [0.969, 0.993]\,.$$

Evidently scale B consistently understates the weight but the bias in measuring the 1 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.

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

$$p - value \ \geq 0.05$$

$$\text{if and only if } P\left(\frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}};\ H_0 : \mu = \mu_0 \text{ is true}\right) \geq 0.05$$

$$\text{if and only if } \ P\left(|T| \ \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05 \quad \text{where } T \sim t\,(n-1)$$

$$\text{if and only if } \ P\left(|T| \ \leq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \leq 0.95$$

$$\text{if and only if } \ \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \ \leq \ a \quad \text{where } P\,(|T| \ \leq a) = 0.95$$

$$\text{if and only if } \ \mu_0 \ \in \ \left[\bar{y} - as/\sqrt{n}, \bar{y} + as/\sqrt{n}\right]$$

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 inside a 95% confidence interval for μ (assuming we use the same pivotal quantity).

More generally, suppose we have data $\mathbf{y}$, a model $f\,(\mathbf{y}; \theta)$ and we use the same pivotal quantity to construct a confidence interval for θ and a test of the hypothesis $H_0 : \theta = \theta_0$. Then the parameter value $\theta = \theta_0$ is inside a $100q\%$ confidence interval for θ if and only if the $p - value$ for testing $H_0 : \theta = \theta_0$ is greater than $1 - q$.

Example 5.2.1 Revisited

For the weigh scale example a 95% confidence interval for the mean μ for the second scale 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.)

Hypothesis tests 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

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2(n-1)$$

to construct confidence intervals for the parameter σ. We may also wish to test a hypothesis such as $H_0 : \sigma = \sigma_0$. 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 $p - value = 2P(U \geq u)$.

Example 5.2.2

For the manufacturing process in Example 4.7.2, test the hypothesis $H_0 : \sigma = 0.008$ (0.008 is the desired or target value of σ the manufacturer would like to achieve). Note that since the value $\sigma = 0.008$ is outside the two-sided 95% confidence interval for σ in Example 4.5.2, 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 follow the procedure above:

1. $u = (n-1)s^2/\sigma_0^2 = (14)\, s^2/\left(0.008\right)^2 = 0.002347/\left(0.008\right)^2 = 36.67$

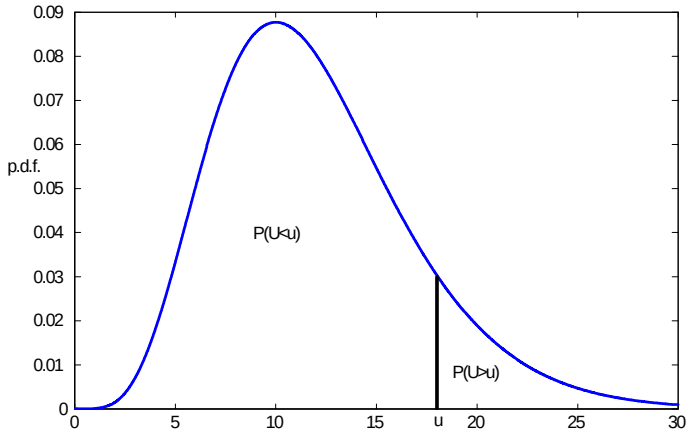


Figure 5.1: **Picture of large observed** u

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 is obtain using R. Alternatively if we use the Chi-squared tables provided in these notes we obtain $P(U \leq 31.319) = 0.995$ so

$$p-value < 2(1-0.995) = 0.01$$

In either case we have that the $p-value < 0.01$ and thus there is strong evidence based on the observed data against $H_0 : \sigma = 0.008$. Since the observed value of $s = \sqrt{0.002347/14} = 0.0129$ is greater than 0.008, the data suggest that σ is bigger than 0.008.

5.3 Likelihood Ratio Tests of Hypotheses - 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, $L(\theta_0)/L(\tilde{\theta})$[32], appears to be a natural statistic for testing $H_0 : \theta = \theta_0$. This only leaves determining the distribution of $L(\theta_0)/L(\tilde{\theta})$ under H_0 so we can determine $p - values$. Equivalently, we usually work instead with a simple function of $L(\theta_0)/L(\tilde{\theta})$. We use the likelihood ratio statistic which was introduced in Chapter 4:

$$\Lambda(\theta_0) = -2 \log \left[\frac{L(\theta_0)}{L(\tilde{\theta})} \right]. \tag{5.4}$$

We choose this particular function because, if $H_0 : \theta = \theta_0$ is true, then $\Lambda(\theta_0) \backsim \chi^2(1)$. 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 $p - value$.

To determine the $p - value$ we first calculate the observed value of $\Lambda(\theta_0)$, denoted by $\lambda(\theta_0)$ and given by

$$\lambda(\theta_0) = -2 \log \left[\frac{L(\theta_0)}{L(\hat{\theta})} \right] = -2 \log R(\theta_0)$$

[32] Recall that $L(\theta) = L(\theta; \mathbf{y})$ is a function of the observed data $\mathbf{y}$ and therefore replacing $\mathbf{y}$ by the corresponding random variable $\mathbf{Y}$ means that $L(\theta; \mathbf{Y})$ is a random variable. Therefore 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 $\tilde{\theta} = g(\mathbf{Y})$.

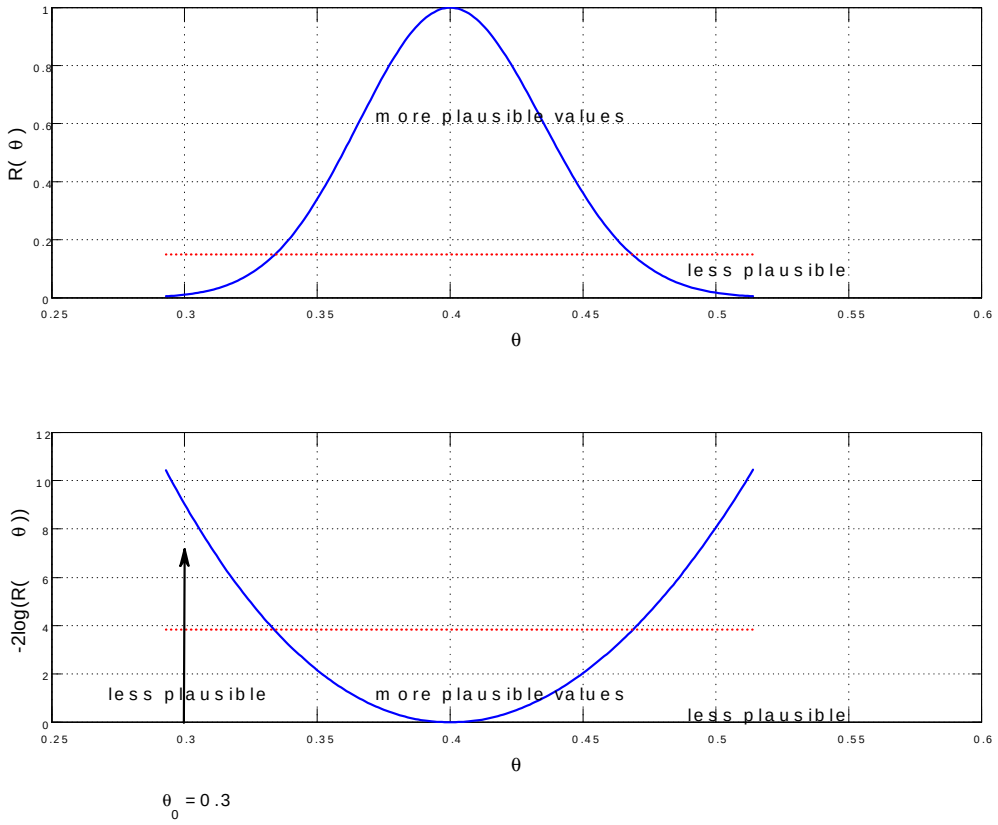


Figure 5.2: **Top panel:** Graph of the relative likelihood function.
Bottom Panel: $\Lambda(\theta) = -2\log R(\theta)$. Note that $\Lambda(\theta_0)$ is relatively large when $R(\theta_0)$ is small.

where $R(\theta_0)$ is the relative likelihood function evaluated at $\theta = \theta_0$. The approximate $p - value$ is then

$$\begin{aligned} p - value &\approx P[W \geq \lambda(\theta_0)] \quad \text{where} \quad W \sim \chi^2(1) \\ &= P\left(|Z| \geq \sqrt{\lambda(\theta_0)}\right) \quad \text{where} \quad Z \backsim G(0,1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{\lambda(\theta_0)}\right)\right] \end{aligned} \tag{5.5}$$

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 $\Lambda(\theta_0)$ 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, $\Lambda(\theta_0)$ 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

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} \quad \text{for} \quad 0 < \theta < 1$$

the likelihood ratio test statistic for testing the hypothesis $H_0 : \theta = \theta_0$ is

$$\Lambda(\theta_0) = -2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] = -2\log\left[\frac{\theta^y(1-\theta)^{n-y}}{\tilde{\theta}^y(1-\tilde{\theta})^{n-y}}\right]$$

where $\tilde{\theta} = Y/n$ is the maximum likelihood estimator of θ. The observed value of $\Lambda(\theta_0)$ is

$$\begin{aligned}\lambda(\theta_0) &= -2\log R(\theta_0) = -2\log\left[\frac{\theta_0^y(1-\theta_0)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}}\right] \\ &= -2\log\left[\left(\frac{\theta_0}{\hat{\theta}}\right)^y\left(\frac{1-\theta_0}{1-\hat{\theta}}\right)^{n-y}\right]\end{aligned}$$

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

$$\begin{aligned}&p - value \approx P(W \geq 2.003) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{2.003}\right)\right] \quad \text{where } Z \backsim G(0,1) \\ &= 2[1 - P(Z \leq 1.42)] = 2(1 - 0.9222) \\ &= 0.1556\end{aligned}$$

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

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{-n} e^{-n\bar{y}/\theta}}{\hat{\theta}^{-n} e^{-n\bar{y}/\hat{\theta}}} = \left(\frac{\hat{\theta}}{\theta}\right)^n e^{n\left(1-\hat{\theta}/\theta\right)} \quad \text{for } \theta > 0.$$

The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0$ is

$$\Lambda(\theta_0) = -2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] = -2\log\left[\left(\frac{\tilde{\theta}}{\theta}\right)^n e^{n\left(1-\tilde{\theta}/\theta\right)}\right]$$

where $\tilde{\theta} = \bar{Y}$ and the observed value of $\Lambda(\theta_0)$ is

$$\lambda(\theta_0) = -2\log R(\theta_0) = -2\log\left[\left(\frac{\hat{\theta}}{\theta_0}\right)^n e^{n\left(1-\hat{\theta}/\theta_0\right)}\right].$$

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

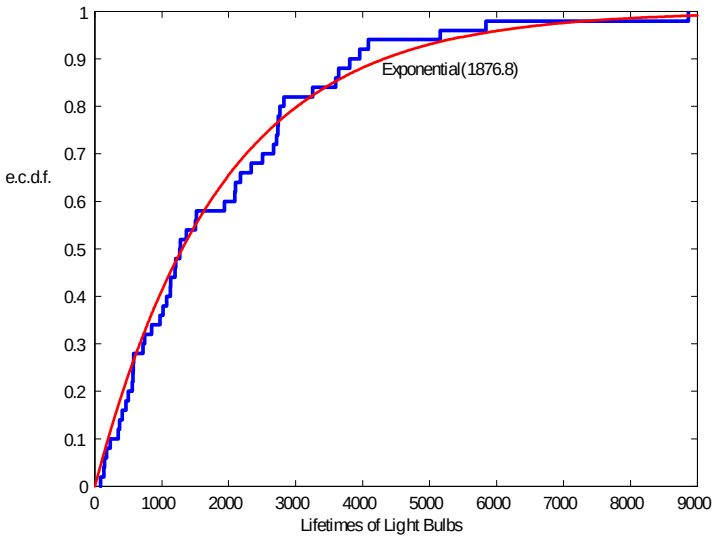


Figure 5.3: Empirical c.d.f. and Exponential(1876.8) c.d.f.

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

$$\begin{aligned} \lambda(2000) &= -2\log R(2000) = -2\log\left[\left(\frac{1876.8}{2000}\right)^{50} e^{50(1-1876.8/2000)}\right] \\ &= -2\log(0.9058) = 0.1979. \end{aligned}$$

(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

$$\begin{aligned} p - value &\approx P(W \geq 0.1979) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.1979}\right)\right] \quad \text{where } Z \backsim G(0,1) \\ &= 2[1 - P(Z \leq 0.44)] = 2(1 - 0.67003) = 0.65994 \end{aligned}$$

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

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

Let us begin with the (rather unrealistic) assumption that the standard deviation σ has a known value and so the only unknown parameter is μ. In this case the likelihood function for an observed sample $y_1, y_2, \ldots, y_n$ from this distribution is

$$L(\mu) = \prod_{i=1}^{n} f(y_i; \mu, \sigma) = (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \quad \text{for } \mu \in \Re.$$

Using the algebraic identity[33]

$$\sum_{i=1}^{n}(y_i-\mu)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2.$$

we can write the likelihood function as

$$\begin{aligned} L(\mu) &= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right]\right\} \\ &= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\} \exp\left[-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right] \end{aligned}$$

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 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

$$\tilde{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

The relative likelihood function can be written as

$$R(\mu) = \frac{L(\mu)}{L(\hat{\mu})} = \exp\left[-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right] \quad \text{for } \mu \in \Re$$

[33] You should be able to verify the identity $\sum_{i=1}^{n}(y_i-c)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-c)^2$ for any value of c

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

$$\begin{aligned}\Lambda(\mu_0) &= -2\log\left[\frac{R(\mu_0)}{R(\tilde{\mu})}\right] \\ &= -2\log\left\{\exp\left[-\frac{n(\bar{Y}-\mu_0)^2}{2\sigma^2}\right]\right\} \quad \text{since } \tilde{\mu} = \bar{Y} \\ &= \frac{n(\bar{Y}-\mu_0)^2}{\sigma^2} \\ &= \left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}\right)^2. \end{aligned} \tag{5.6}$$

The purpose for writing the likelihood ratio statistic in the form (5.6) is to draw attention to the fact that $\Lambda(\mu_0)$ is the square of the standard Normal random variable $\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}$ and therefore has exactly a $\chi^2(1)$ distribution. Of course it is not clear in general that the likelihood ratio test statistic has an approximate $\chi^2(1)$ distribution, but in this special case, the distribution of $\Lambda(\mu_0)$ is clearly $\chi^2(1)$ for all values of n.

5.4 Likelihood Ratio Tests of Hypotheses - Multiparameter[34]

Let the data $\mathbf{Y}$ 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 alone. 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)

$$\Lambda = 2l(\tilde{\boldsymbol{\theta}}) - 2l(\tilde{\boldsymbol{\theta}}_0) = -2\log\left[\frac{L(\tilde{\boldsymbol{\theta}}_0)}{L(\tilde{\boldsymbol{\theta}})}\right] \tag{5.7}$$

[34] Optional

and let

$$\lambda = 2l(\hat{\boldsymbol{\theta}}) - 2l(\hat{\boldsymbol{\theta}}_0) = -2\log\left[\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right]$$

denote an observed value of Λ. If the observed value λ is very large, then there is evidence against H_0 (confirm that this means $L(\hat{\boldsymbol{\theta}})$ is much larger than $L(\hat{\boldsymbol{\theta}}_0)$). In this case 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. Again, 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 here.

Example 5.4.3 Comparison of two Poisson means

In Problem 15 of Chapter 4 some data were given on the numbers of failures per month for each of two companies' photocopiers. To a good approximation we can 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

$$P(X = x;\mu_A) = \frac{\mu_A^x \exp(-\mu_A)}{x!} \quad \text{for } x = 0, 1, \ldots \text{ and } \mu_A > 0.$$

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

$$P(Y = y;\mu_B) = \frac{\mu_B^y \exp(-\mu_B)}{y!} \quad \text{for } y = 0, 1, \ldots \text{ and } \mu_B > 0$$

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 $\Omega = \{(\mu_A, \mu_B) : \mu_A > 0, \mu_B > 0\}$.

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 $\Omega_0 = \{(\mu, \mu) : \mu > 0\}$ 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 $(x_1, \ldots, x_n)$ is

$$L_1(\mu_A) = \prod_{i=1}^{n} f(x_i; \mu_A) = \prod_{i=1}^{n} \frac{\mu_A^{x_i} \exp(-\mu_A)}{x_i!} \quad \text{for } \mu_A > 0$$

or more simply

$$L_1(\mu_A) = \prod_{i=1}^{n} \mu_A^{x_i} \exp(-\mu_A) \quad \text{for } \mu_A > 0.$$

Similarly the likelihood function for μ_B only based on $(y_1, y_2, \ldots, y_m)$ is given by

$$L_2(\mu_B) = \prod_{j=1}^{m} \mu_B^{y_j} \exp(-\mu_B) \quad \text{for } \mu_B > 0.$$

Since the data from A and B are independent, the likelihood function for $\boldsymbol{\theta} = (\mu_A, \mu_B)$ is obtained as a product of the individual likelihoods

$$\begin{aligned} L(\boldsymbol{\theta}) = L(\mu_A, \mu_B) &= L_1(\mu_A) \times L_2(\mu_B) \\ &= \prod_{i=1}^{n} \mu_A^{x_i} \exp(-\mu_A) \prod_{j=1}^{m} \mu_B^{y_j} \exp(-\mu_B) \quad \text{for } (\mu_A, \mu_B) \in \Omega \end{aligned}$$

and the log likelihood function for $\boldsymbol{\theta} = (\mu_A, \mu_B)$ is

$$l(\boldsymbol{\theta}) = -n\mu_A - m\mu_B + \left(\sum_{i=1}^{n} x_i\right) \log \mu_A + \left(\sum_{j=1}^{m} y_j\right) \log \mu_B. \tag{5.9}$$

The number of failures in twelve consecutive months for company A and company B's copiers are given below; there were the same number of copiers from each company in use so $n = m = 12$

Company A:	16	14	25	19	23	12	22	28	19	15	18	29
Company B:	13	7	12	9	15	17	10	13	8	10	12	14

We note that $\sum_{i=1}^{12} x_i = 240$ and $\sum_{j=1}^{12} y_j = 140$.

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:

$$-12 + \frac{240}{\mu_A} = 0$$
$$-12 + \frac{140}{\mu_B} = 0$$

The maximum likelihood estimates of μ_A and μ_B (unconstrained) are $\hat{\mu}_A = 240/12 = 20.0$ and $\hat{\mu}_B = 140/12 = 11.667$. That is, $\hat{\boldsymbol{\theta}} = (20.0, 11.667)$.

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

$$\begin{aligned} l(\mu, \mu) &= -12\mu + 240 \log \mu - 12\mu + 140 \log \mu \\ &= -24\mu + 380 \log \mu \quad \text{for } \mu > 0. \end{aligned}$$

Solving $\partial l(\mu, \mu)/\partial \mu = 0$, we find $\hat{\mu} = 380/24 = 15.833 (= \hat{\mu}_A = \hat{\mu}_B)$; that is,
$\hat{\boldsymbol{\theta}}_0 = (15.833, 15.833)$.

The next step is to compute the observed value of the likelihood ratio statistic, which from (5.7) is

$$\begin{aligned} \lambda &= 2l(\hat{\boldsymbol{\theta}}) - 2l(\hat{\boldsymbol{\theta}}_0) \\ &= 2l(20.0, 11.667) - 2l(15.833, 15.833) \\ &= 2\,(682.92 - 669.60) \\ &= 26.64 \end{aligned}$$

Finally, we compute the approximate $p - value$ for the test, which by (5.8) is

$$\begin{aligned} p - value &= P(\Lambda \geq 26.64; H_0) \\ &\approx P(W \geq 26.64) \quad \text{where } W \sim \chi^2\,(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{26.64}\right)\right] \quad \text{where } \; Z \backsim G\,(0, 1) \\ &\approx 0. \end{aligned}$$

Our conclusion is that there is very strong evidence against the hypothesis $H_0 : \mu_A = \mu_B$. The data indicate that Company B's copiers have a lower rate of failure than Company A's copiers.

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 a lot, but we could also give a confidence interval in order to express the uncertainty in such estimates.

Example 5.4.4 Likelihood ratio tests of hypotheses for σ for $G(\mu, \sigma)$ model for 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

$$L(\boldsymbol{\theta}) = L(\mu, \sigma) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]$$

and the log likelihood function is

$$l(\mu, \sigma) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + c$$

where

$$c = \log\left[(2\pi)^{-n/2}\right]$$

does not depend on μ or σ. The maximum likelihood estimators of (μ, σ) in the unconstrained case are

$$\tilde{\mu} = \bar{Y}$$
$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Under the constraint imposed by $H_0 : \sigma = \sigma_0$ the maximum likelihood estimator of the parameter μ is also $\bar{Y}$ so the likelihood ratio statistic is

$$\begin{aligned}
\Lambda(\sigma_0) &= 2l\left(\bar{Y}, \tilde{\sigma}\right) - 2l\left(\bar{Y}, \sigma_0\right) \\
&= -2n\log(\tilde{\sigma}) - \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2n\log(\sigma_0) + \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
&= 2n\log\left(\frac{\sigma_0}{\tilde{\sigma}}\right) + \left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}^2}\right)n\tilde{\sigma}^2 \\
&= n\left[\left(\frac{\tilde{\sigma}^2}{\sigma_0^2} - 1\right) - \log\left(\frac{\tilde{\sigma}^2}{\sigma_0^2}\right)\right].
\end{aligned}$$

This is not as obviously a Chi-squared random variable. It 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 obtaining 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 $p-value$

$$\begin{aligned} p-value &\approx P(W > \lambda(\sigma_0)) \text{ where } W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{\lambda(\sigma_0)}\right)\right] \text{ where } Z \backsim G(0,1) \end{aligned}$$

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 16(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

$$f(y_1, y_2, \ldots, y_k; \theta_1, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_k^{y_k} \quad \text{for } 0 \leq y_j \leq n \text{ where } \sum_{j=1}^{k} y_j = n.$$

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 $\Omega = \{(\theta_1, \theta_2, \ldots, \theta_k) : 0 \leq \theta_j \leq 1, \text{ where } \sum_{j=1}^{k} \theta_j = 1\}$ has dimension $k-1$. The parameter in the model assuming H_0 is $\boldsymbol{\theta}_0 = (\theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), .., \theta_k(\boldsymbol{\alpha}))$ and the parameter space $\Omega_0 = \{(\theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), .., \theta_k(\boldsymbol{\alpha})) : \text{for all } \boldsymbol{\alpha}\}$ 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

$$\lambda = 2l(\hat{\boldsymbol{\theta}}) - 2l(\hat{\boldsymbol{\theta}}_0)$$

is the observed value of Λ. We will give specific examples of the Multinomial model in Chapter 7.

5.5 Chapter 5 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

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \backsim G(0, 1)$$

with $\sigma = 2$.

5. 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 placed 12 detectors 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

   $$\sum_{i=1}^{12} y_i = 1249.6 \quad \text{and} \quad \sum_{i=1}^{12} {y_i}^2 = 131096.44.$$

   To analyze these data assume the model

   $$Y_i \backsim G(\mu, \sigma), \quad i = 1, 2, \ldots, 12 \text{ independently}$$

   where μ and σ are unknown parameters.

   (a) Test the hypothesis $H_0 : \mu = 105$. Determine a 95% confidence interval for μ.

   (b) Determine a 95% confidence interval for σ.

   (c) As a statistician what would you say to the university researchers about the accuracy and precision of the detectors?

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

   $$\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \backsim \chi^2(n)$$

   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 $Y \backsim \text{Binomial}(6000, \theta)$ 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:

$$P\left(\frac{\bar{Y} - \theta}{\sqrt{\theta/n}} \leq c\right) \approx P\left(Z \leq c\right) \quad \text{where} \quad Z \sim N\left(0, 1\right).$$

(c) Why is the discrepancy measure $D = \left|\bar{Y} - 3\right|$ 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 $5\,(b)$ 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 8?

11. For Chapter 2, Problem 7 (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 9?

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 10 (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 10?

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°. The 20 failure times $y_1, y_2, \ldots, y_{20}$ were (in hours):

374.2	544.0	1113.9	509.4	1244.3
551.9	853.2	3391.2	297.0	63.1
250.2	678.1	379.6	1818.9	1191.1
162.8	1060.1	1501.4	332.2	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 $y \leftarrow runif(n)$ 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. **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

$$\sum_{i=1}^{n}(Y_i - \mu_0)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2.$$

    (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.

17. **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*[38], and the β_j's are called the *regression coefficients*.

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

$$Y \sim G(\mu_O, \sigma_O) \quad \text{for observations from the old machine}$$
$$Y \sim G(\mu_N, \sigma_N) \quad \text{for observations from the new machine.}$$

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 buildings[39]

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.

Table 6.1: Size and Price of 30 Buildings

Size	Price	Size	Price	Size	Price
3.26	226.2	0.86	532.8	0.38	636.4
3.08	233.7	0.80	563.4	0.38	657.9
3.03	248.5	0.77	578.0	0.38	597.3
2.29	360.4	0.73	597.3	0.38	611.5
1.83	415.2	0.60	617.3	0.38	670.4
1.65	458.8	0.48	624.4	0.34	660.6
1.14	509.9	0.46	616.4	0.26	623.8
1.11	525.8	0.45	620.9	0.24	672.5
1.11	523.7	0.41	624.3	0.23	673.5
1.00	534.7	0.40	641.7	0.20	611.8

[38] The word *regression* is an historical term introduced in the 19th century in connection with these models.
[39] This reference can be found in earlier course notes for Oldford and MacKay, STAT 231 Ch. 16

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 are given in Table 6.1 and plotted in Figure 6.1. They include the size of the building x (in $m^2/10^5$) and the selling price y (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 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 is roughly inversely proportional to the 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

$$Y_i \sim G\left(\beta_0 + \beta_1 x_i, \sigma\right) \quad \text{for } i = 1, 2, \ldots, n \ \text{ independently}$$

where β_0 and β_1 are unknown parameters. We assume a common standard deviation σ for the observations.

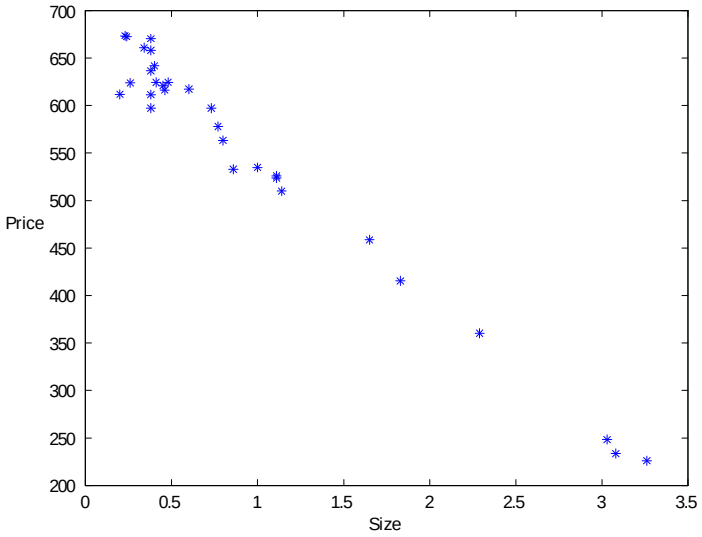


Figure 6.1: **Scatterplot of price versus building size**

Example 6.1.3 Strength of steel bolts

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.

The data below show the breaking strengths y of six steel bolts at each of five different bolt diameters x. The data are plotted in Figure 6.2.

Diameter x	0.10	0.20	0.30	0.40	0.50
	1.62	1.71	1.86	2.14	2.45
Breaking	1.73	1.78	1.86	2.07	2.42
Strength	1.70	1.79	1.90	2.11	2.33
	1.66	1.86	1.95	2.18	2.36
	1.74	1.70	1.96	2.17	2.38
	1.72	1.84	2.00	2.07	2.31

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.

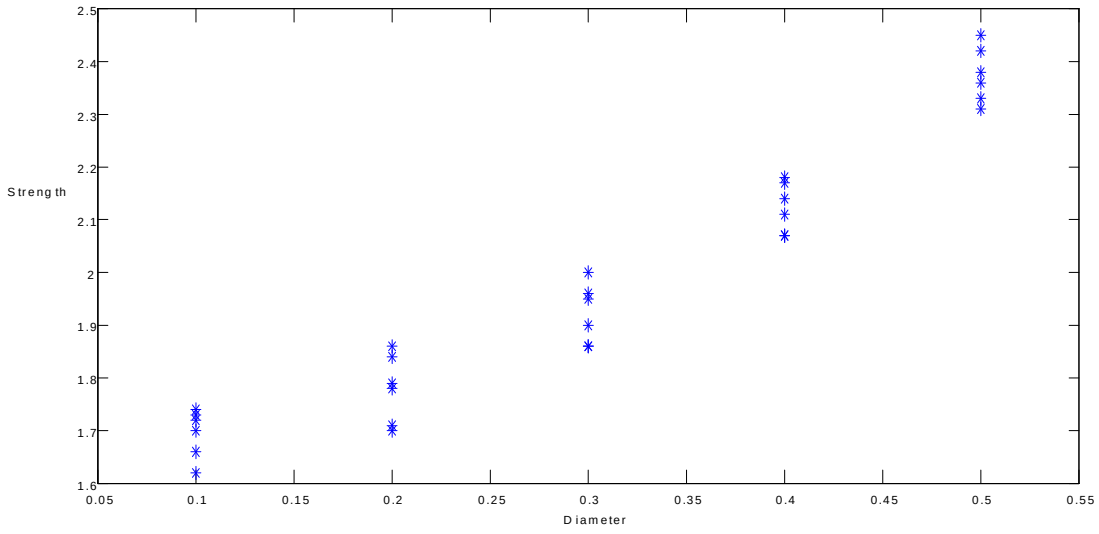


Figure 6.2: **Scatterplot of strength versus bolt diameter**

Remark: Sometimes the model (6.1) is written a little differently as

$$Y_i = \mu(\mathbf{x}_i) + R_i \text{ where } R_i \sim G(0, \sigma).$$

This splits Y_i into a deterministic component, $\mu(\mathbf{x}_i)$, and a random component, R_i.

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

$$Y_i = \mu + R_i \text{ where } R_i \sim G(0, \sigma). \tag{6.2}$$

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 $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. This estimator is also a "least squares estimator". $\bar{Y}$ has the property that it is closer to the data than any other constant, or

$$\min_{\mu} \sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

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

[40]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 likelihood function for (α, β, σ) is

$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

[40]See the video at www.watstat.ca called "Regression and Crickets3".

or more simply

$$L(\alpha,\beta,\sigma)=\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta x_i\right)^2\right].$$

The log likelihood function is

$$l(\alpha,\beta,\sigma)=-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta x_i\right)^2.$$

To obtain the maximum likelihood estimates we solve the three equations

$$\begin{aligned}
\frac{\partial l}{\partial\alpha} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta x_i\right)=\frac{n}{\sigma^2}\left(\bar{y}-\alpha-\beta\bar{x}\right)=0 &(6.4)\\
\frac{\partial l}{\partial\beta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\alpha-\beta x_i\right)x_i=\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_iy_i-\alpha\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2\right)=0 &(6.5)\\
\frac{\partial l}{\partial\sigma} &= -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}\left(y_i-\alpha-\beta x_i\right)^2=0
\end{aligned}$$

simultaneously. We obtain the maximum likelihood estimators

$$\tilde{\beta}=\frac{S_{xy}}{S_{xx}}, \tag{6.6}$$

$$\tilde{\alpha}=\bar{Y}-\tilde{\beta}\bar{x}, \tag{6.7}$$

$$\tilde{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\tilde{\alpha}-\tilde{\beta}x_i)^2 \tag{6.8}$$

where

$$\begin{aligned}
S_{xx}&=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})x_i\\
S_{xy}&=\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})=\sum_{i=1}^{n}(x_i-\bar{x})Y_i\\
S_{yy}&=\sum_{i=1}^{n}(Y_i-\bar{Y})^2
\end{aligned}$$

The alternative expressions for S_{xy} and S_{yy} [41] are easy to obtain.

We will use

$$S_e^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\tilde{\alpha}-\tilde{\beta}x_i)^2=\frac{1}{n-2}(S_{yy}-\tilde{\beta}S_{xy})$$

[41] Since $\sum_{i=1}^{n}(x_i-\bar{x})=0$,
$\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})x_i-\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})x_i$
and
$\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})=\sum_{i=1}^{n}(x_i-\bar{x})Y_i-\bar{Y}\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})Y_i$

as the estimator of σ^2 rather than the maximum likelihood estimator $\tilde{\sigma}^2$ given by (6.8) since it can be shown that $E\left(S_e^2\right) = \sigma^2$. Note that S_e^2 can be more easily calculated using

$$S_e^2 = \frac{1}{n-2}(S_{yy} - \tilde{\beta} S_{xy})$$

which follows since

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2 &= \sum_{i=1}^{n}(Y_i - \bar{Y} + \tilde{\beta}\bar{x} - \tilde{\beta}x_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\tilde{\beta}\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) + \tilde{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= S_{yy} - 2\tilde{\beta}S_{xy} + \tilde{\beta}\left(\frac{S_{xy}}{S_{xx}}\right)S_{xx} \\
&= S_{yy} - \tilde{\beta}S_{xy}.
\end{aligned}$$

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

$$g(\alpha, \beta) = \sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2.$$

Such estimates are called *least squares estimates.* To find the least squares estimates we need to solve the two equations

$$\begin{aligned}
\frac{\partial g}{\partial \alpha} &= \sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = n(\bar{y} - \alpha - \beta\bar{x}) = 0 \\
\frac{\partial g}{\partial \beta} &= \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = \sum_{i=1}^{n}x_i y_i - \alpha\sum_{i=1}^{n}x_i - \beta\sum_{i=1}^{n}x_i^2 = 0.
\end{aligned}$$

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. We now show how to obtain confidence intervals based on the model (6.3).

Distribution of the estimator $\tilde{\beta}$

Notice that we can rewrite the expression for $\tilde{\beta}$ as

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n}a_i Y_i \quad \text{where } a_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

to make it clear that $\tilde{\beta}$ is a linear combination of the Normal random variables Y_i and is therefore Normally distributed with easily obtained expected value and variance. In fact it is easy to show that these non-random coefficients satisfy $\sum_{i=1}^{n} a_i = 0$ and $\sum_{i=1}^{n} a_i x_i = 1$ and $\sum_{i=1}^{n} a_i^2 = 1/S_{xx}$. Therefore

$$\begin{aligned} E(\tilde{\beta}) &= \sum_{i=1}^{n} a_i E(Y_i) = \sum_{i=1}^{n} a_i (\alpha + \beta x_i) \\ &= \beta \sum_{i=1}^{n} a_i x_i \quad \text{since } \sum_{i=1}^{n} a_i = 0 \\ &= \beta \quad \text{since } \sum_{i=1}^{n} a_i x_i = 1. \end{aligned}$$

Similarly

$$\begin{aligned} Var(\tilde{\beta}) &= \sum_{i=1}^{n} a_i^2 Var(Y_i) \quad \text{since the } Y_i \text{ are independent random variables} \\ &= \sigma^2 \sum_{i=1}^{n} a_i^2 \\ &= \frac{\sigma^2}{S_{xx}} \quad \text{since } \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{xx}}. \end{aligned}$$

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

Confidence intervals for β are important because the parameter β represents the increase in the mean value of Y, resulting from an increase of one unit in the value of x. As well, if $\beta = 0$ then x has no effect on Y (within this model).

Since

$$\frac{\tilde{\beta} - \beta}{\sigma/\sqrt{S_{xx}}} \sim G(0, 1)$$

holds independently of

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2) \tag{6.9}$$

then by Theorem 32 it follows that

$$\frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \backsim t(n-2). \tag{6.10}$$

This pivotal quantity can be used to obtain confidence intervals for β and to construct tests of hypotheses about β.

Using t-tables or R find the constant a such that $P(-a \leq T \leq a) = p$ where $T \sim t(n-2)$. Since

$$\begin{aligned} p &= P(-a \leq T \leq a) = P\left(-a \leq \frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \leq a\right) \\ &= P\left(\tilde{\beta} - aS_e/\sqrt{S_{xx}} \leq \beta \leq \tilde{\beta} + aS_e/\sqrt{S_{xx}}\right), \end{aligned}$$

therefore a $100p\%$ confidence interval for β is given by

$$\hat{\beta} \pm as_e/\sqrt{S_{xx}} = \left[\hat{\beta} - as_e/\sqrt{S_{xx}}, \hat{\beta} + as_e/\sqrt{S_{xx}}\right]$$

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

$$\begin{aligned} p - value &= P\left(|T| \geq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right) \\ &= 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right)\right] \quad \text{where} \quad T \backsim t(n-2). \end{aligned}$$

Note also that (6.9) can be used to obtain confidence intervals or tests for σ, but these are usually of less interest than inference about β or the other quantities below.

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,

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x}),$$

since $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$. Since

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i$$

we can rewrite $\tilde{\mu}(x)$ as

$$\tilde{\mu}(x) = \bar{Y} + \tilde{\beta}(x - \bar{x}) = \sum_{i=1}^{n} a_i Y_i \quad \text{where} \quad a_i = \frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}}. \tag{6.11}$$

Since $\tilde{\mu}(x)$ is a linear combination of Gaussian random variables it has a Gaussian distribution. We can use (6.11) to determine the mean and variance of the random variable $\tilde{\mu}(x)$. You should verify the following properties of the coefficients a_i:

$$\sum_{i=1}^{n} a_i = 1, \quad \sum_{i=1}^{n} a_i x_i = x \quad \text{and} \quad \sum_{i=1}^{n} a_i^2 = \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}.$$

Therefore

$$\begin{aligned} E[\tilde{\mu}(x)] &= \sum_{i=1}^{n} a_i E(Y_i) = \sum_{i=1}^{n} a_i (\alpha + \beta x_i) \\ &= \alpha \left(\sum_{i=1}^{n} a_i \right) + \beta \left(\sum_{i=1}^{n} a_i x_i \right) \\ &= \alpha + \beta x \quad \text{since } \sum_{i=1}^{n} a_i = 1 \text{ and } \sum_{i=1}^{n} a_i x_i = x \\ &= \mu(x). \end{aligned}$$

In other words $\tilde{\mu}(x)$ is an unbiased estimator of $\mu(x)$. Also

$$\begin{aligned} Var\left[\tilde{\mu}(x)\right] &= \sum_{i=1}^{n} a_i^2 Var(Y_i) \quad \text{since the } Y_i \text{ are independent random variables} \\ &= \sigma^2 \sum_{i=1}^{n} a_i^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]. \end{aligned}$$

Note that the variance of $\tilde{\mu}(x)$ is smallest in the middle of the data, or when x is close to $\bar{x}$ 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 (6.9) then by Theorem (32) we obtain the pivotal quantity

$$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2) \tag{6.12}$$

which can be used to obtain confidence intervals for $\mu(x)$ in the usual manner. Using t-tables or R find the constant a such that $P(-a \le T \le a) = p$ where $T \sim t(n-2)$. Since

$$\begin{aligned} p &= P(-a \le T \le a) = P\left(-a \le \frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \le a\right) \\ &= P\left(\tilde{\mu}(x) - aS_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}} \le \mu(x) \le \tilde{\mu}(x) + aS_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right), \end{aligned}$$

a $100p\%$ confidence interval for $\mu(x)$ is given by

$$\left[\hat{\mu}(x) - as_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}},\ \hat{\mu}(x) + as_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right] \tag{6.13}$$

where $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$,

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$$

and S_{yy} and S_{xy} are replaced by their observed values.

Remark: Note that since $\alpha = \mu(0)$, a 95% confidence interval for α, is given by (6.13) with $x = 0$ which gives

$$\hat{\alpha} \pm as_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}} \tag{6.14}$$

In fact one can see from (6.14) 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. In the following example it refers to a building of area $x = 0$, which is nonsensical!

Remark: The results of the analyses below can be obtained using the R function lm, with the command $lm(y \sim x)$. We give the detailed results below to illustrate how the calculations are made. In R, *summary(lm(y∼x))* gives a lot of useful output.

Example 6.1.2 Revisited Price versus size of commercial buildings

Example 6.1.2 gave data on the selling price per square meter y and area x of commercial buildings. Figure 6.1 suggested that a linear regression model of the form $E(Y|x) = \alpha + \beta x$ would be reasonable. For the given data

$$n = 30,\ \bar{x} = 0.9543,\ \bar{y} = 548.9700,\ S_{xx} = 22.9453,\ S_{xy} = -3316.6771,\ S_{yy} = 489,624.723$$

so we find

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-3316.6771}{22.9453} = -144.5469,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 548.9700 - (-144.5469)\,(0.9543) = 686.9159,$$

$$s_e^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy}) = \frac{1}{28}\left[489624.723 - (-144.5469)\,(-3316.6771)\right] = 364.6199,$$

and $s_e = 19.0950$.

(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.) Since $\hat{\beta}$ is negative this implies that the larger sized buildings tend to sell for less per square meter. (The estimate $\hat{\beta} = -144.55$ indicates a drop in average price of \$144.55 per square meter for each increase of one unit in x; remember x's units are $m^2(10^5)$).

The line $y = \hat{\alpha} + \hat{\beta}x$ is often called the *fitted regression line for y on x*. If we plot the fitted line on the same graph as the scatterplot of points (x_i, y_i), $i = 1, 2, \ldots, n$ as in Figure 6.3, we see the fitted line passes close to the points.

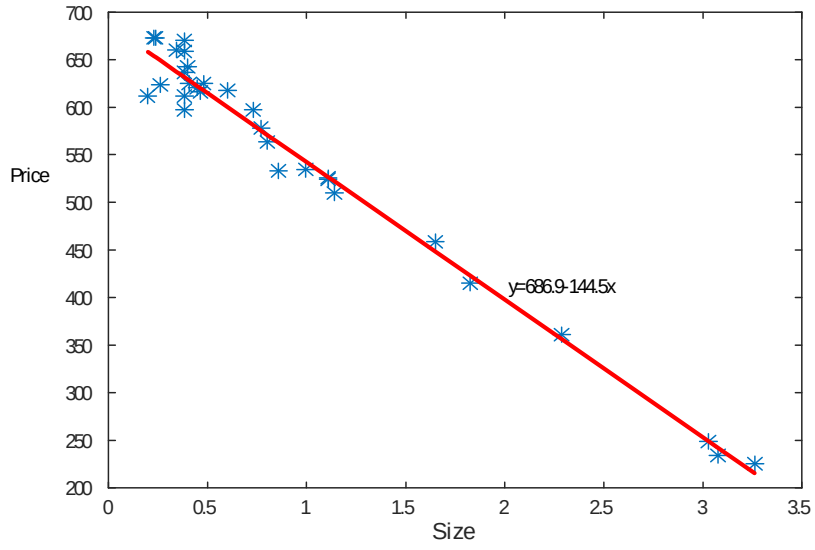


Figure 6.3: **Scatterplot and fitted line for building price versus size**

A confidence interval for β is not of major interest in the setting here, where the data were called on to indicate a fair assessment value for a large building with $x = 4.47$. One

way to address this is to estimate $\mu(x)$ when $x = 4.47$. We get the maximum likelihood estimate for $\mu(4.47)$ as

$$\hat{\mu}(4.47) = \hat{\alpha} + \hat{\beta}(4.47) = \$40.79$$

which we note is much below the assessed value of \$75 per square meter. However, one can object that there is uncertainty in this estimate, and that it would be better to give a confidence interval for $\mu(4.47)$. Using (6.13) and $P(T \leq 2.0484) = 0.975$ for $T \sim t\,(28)$ we get a 95% confidence interval for $\mu(4.47)$ as

$$\begin{aligned} & \hat{\mu}(4.47) \pm 2.0484 s_e \sqrt{\frac{1}{30} + \frac{(4.47 - \bar{x})^2}{S_{xx}}} \\ = \; & \$40.79 \pm \$29.58 \\ = \; & [\$11.21,\ \$70.37]\,. \end{aligned}$$

Thus the assessed value of \$75 is outside this interval.

However (playing lawyer for the assessor), we could raise another objection: we are considering a **single** building but we have constructed a confidence interval for the average of all buildings of size $x = 4.47(\times 10^5)m^2$. The constructed confidence interval is for a point **on** the line, not a point Y generated by adding to $\alpha + \beta(4.47)$ the random error $R \sim G\,(0, \sigma)$ which has a non-negligible variance. This suggests that what we should do is **predict** the y value for a building with $x = 4.47$, instead of estimating $\mu(4.47)$. We will temporarily leave the example in order to develop a method for this.

Prediction Interval for Future Response

Suppose we want 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. We can obtain a pivotal quantity that can be used to give a prediction interval (or interval "estimate") for the future response Y, as follows.

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 $\mu(x)$. We have derived its distribution as

$$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right).$$

Moreover the error in the point estimator of Y is given by

$$Y - \tilde{\mu}(x) = Y - \mu(x) + \mu(x) - \tilde{\mu}(x) = R + [\mu(x) - \tilde{\mu}(x)]\,. \tag{6.15}$$

Since R is independent of $\tilde{\mu}(x)$ (it is not connected to the existing sample), (6.15) is the sum of independent Normally distributed random variables and is consequently Normally distributed. Since

$$\begin{aligned} E\left[Y-\tilde{\mu}(x)\right] &= E\left\{R+\left[\mu(x)-\tilde{\mu}(x)\right]\right\} \\ &= E(R)+E\left[\mu(x)\right]-E\left[\tilde{\mu}(x)\right] \\ &= 0+\mu(x)-\mu(x)=0. \end{aligned}$$

and

$$\begin{aligned} Var\left[Y-\tilde{\mu}(x)\right] &= Var(Y)+Var\left[\tilde{\mu}(x)\right] \\ &= \sigma^2+\sigma^2\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right]. \end{aligned}$$

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

$$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}\sim G\left(0,1\right). \tag{6.16}$$

Since (6.16) holds independently of (6.9) then by Theorem (32) we obtain the pivotal quantity

$$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}\backsim t\left(n-2\right).$$

For an interval estimate with confidence coefficient p we choose a such that $p=P(-a\leq T\leq a)$ where $T\sim t\left(n-2\right)$. 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 interval

$$\left[\hat{\mu}(x)-as_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}},\ \hat{\mu}(x)+as_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}\right]. \tag{6.17}$$

This interval is usually called a $100p\%$ *prediction interval* instead of a confidence interval, since Y is not a parameter but a "future" observation.

Remark: Care must be taken in constructing prediction intervals for values of x which lie outside the interval of observed $x_i's$ since this assumes that the linear relationship holds beyond the observed data. This is dangerous since there are no data to support the assumption.

Example 6.1.2 Revisited Price versus size of commercial buildings

Let us obtain a 95% prediction interval for Y when $x = 4.47$. Using (6.17) and the fact that $P(T \leq 2.0484) = 0.975$ when $T \sim t\,(28)$ we obtain

$$\begin{aligned} & \tilde{\mu}(4.47) \pm 2.0484 s_e \sqrt{1 + \frac{1}{30} + \frac{(4.47 - \bar{x})^2}{22.945}} \\ = \quad & \$40.79 \pm \$49.04 = [-\$8.25,\ \$89.83] \end{aligned}$$

The lower limit is negative, which is nonsensical. This happened because we were using a Gaussian model (Gaussian random variables Y can be positive or negative) in a setting where the price Y must be positive. Nonetheless, the Gaussian model fits the data reasonably well. We might just truncate the prediction interval and take it to be [0, \$89.83].

Now we find that the assessed value of \$75 is inside this interval. On this basis it's difficult to say that the assessed value is unfair (though it is towards the high end of the prediction interval). Note also that the value $x = 4.47$ of interest is well outside the interval of observed x values which was $[0.20, 3.26]$) in the data set of 30 buildings. Thus any conclusions we reach are based on an assumption that the linear model $E\,(Y|x) = \alpha + \beta x$ applies beyond $x = 3.26$ at least as far as $x = 4.47$. This may or may not be true, but we have no way to check it with the data we have.

There is a slight suggestion in Figure 6.3 that $Var(Y)$ may be smaller for larger x values. There is not sufficient data to check this either. We mention these points because an important companion to every statistical analysis is a qualification of the conclusions based on a careful examination of the applicability of the assumptions underlying the analysis.

Remark: Note from (6.13) and (6.17) that the confidence interval for $\mu(x)$ and the prediction interval for Y are wider the further away x is from $\bar{x}$. Thus, as we move further away from the "middle" of the x's in the data, we get wider and wider intervals for $\mu(x)$ and Y.

Example 6.2.1 Strength of steel bolts

Recall the data given in Example 6.1.3, 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 6.4 shows a plot of points (x_i^2, y_i) which looks quite linear. Because of this let us assign a new variable 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

$$n = 30,\ \bar{x}_1 = 0.11,\ \bar{y} = 1.979,\ S_{x_1x_1} = 0.2244,\ S_{x_1y} = 0.6368,\ S_{yy} = 1.88147$$

so we find

$$\begin{aligned}
\hat{\beta} &= \frac{S_{x_1y}}{S_{x_1x_1}} = \frac{0.6368}{0.2244} = 2.8378, \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}_1 = 1.979 - (2.8378)\,(0.11) = 1.6668, \\
s_e^2 &= \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{x_1y}) = \frac{1}{28}\left[1.88147 - (2.8378)\,(0.6368)\right] = 0.002656, \\
\text{and } s_e &= 0.05154.
\end{aligned}$$

The fitted regression line $y = \hat{\alpha} + \hat{\beta}x_1$ is shown on the scatterplot in Figure 6.4. The model appears to fit the data well.

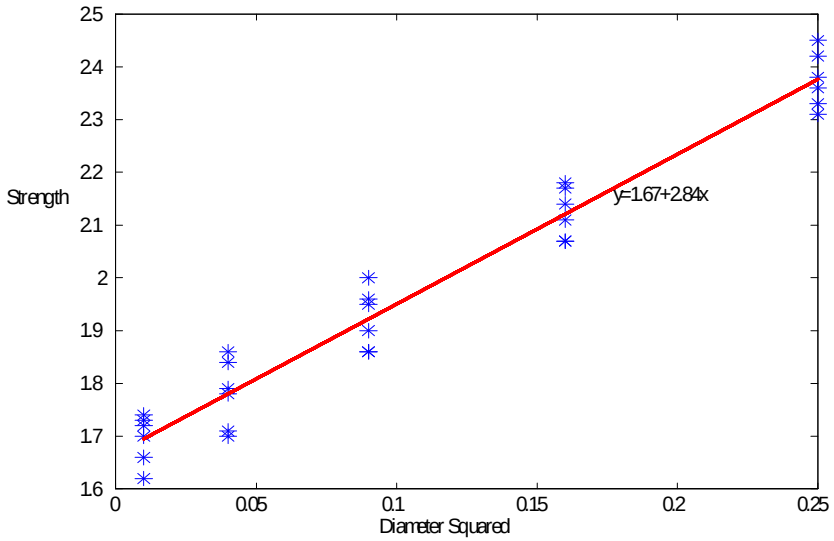


Figure 6.4: **Scatterplot plus fitted line for strength versus diameter squared**

The parameter β represents the increase in average strength $\mu(x_1)$ from increasing $x_1 = x^2$ by one unit. Using the pivotal quantity (6.10) and the fact that $P(T \leq 2.0484) = 0.975$ for $T \sim t\,(28)$, we obtain the 95% confidence interval for β as

$$\begin{aligned}
& \hat{\beta} \pm 2.0484 s_e / \sqrt{S_{xx}} \\
= \ & 2.8378 \pm 0.2228 \\
= \ & [2.6149,\ 3.0606]\,.
\end{aligned}$$

Table 6.2
Summary of Distributions for Simple Linear Regression

Random variable	**Distribution**	**Mean or degrees of freedom**	**Standard Deviation**
$\tilde{\beta} = \frac{S_{xy}}{S_{xx}}$	Gaussian	$E(\tilde{\beta}) = \beta$	$std(\tilde{\beta}) = \sigma \left[\frac{1}{S_{xx}}\right]^{1/2}$
$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{s_{xx}}}$ where $S_e^2 = \frac{1}{n-2}\left(S_{yy} - \tilde{\beta} S_{xy}\right)$	Student t	degrees of freedom $= n-2$	
$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$	Gaussian	$E(\tilde{\alpha}) = \alpha$	$std(\tilde{\alpha}) = \sigma \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]^{1/2}$
$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$	Gaussian	$E\left[\tilde{\mu}(x)\right]$ $= \mu(x)$ $= \alpha + \beta x$	$std\left[\tilde{\mu}(x)\right]$ $= \sigma \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}$
$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$	Student t	degrees of freedom $= n-2$	
$Y - \tilde{\mu}(x)$	Gaussian	$E\left[Y - \tilde{\mu}(x)\right]$ $= 0$	$std\left[Y - \tilde{\mu}(x)\right]$ $= \sigma \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}$
$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$	Student t	degrees of freedom $= n-2$	
$\frac{(n-2)S_e^2}{\sigma^2}$	Chi-squared	degrees of freedom $= n-2$	

Checking the Model Assumptions for Simple Linear Regression

There are two main components in Gaussian linear response models:

(1) The assumption that Y_i (given any covariates x_i) is Gaussian with constant standard deviation σ.

(2) The assumption that $E(Y_i) = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.

Models should always be checked. In problems with only one x covariate, a plot of the fitted line superimposed on the scatterplot of the data (as in Figures 6.3 and 6.4) shows pretty clearly how well the model fits. If there are two or more covariates in the model, residual plots, which are described below, are very useful for checking the model assumptions.

Residuals are defined as the difference between the observed response and the fitted values. 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. The residuals are given by

$$\begin{aligned} \hat{r}_i &= y_i - \hat{\mu}_i \\ &= y_i - \hat{\alpha} - \hat{\beta} x_i \quad \text{for } i = 1, 2, \ldots, n. \end{aligned}$$

The idea behind the $\hat{r}_i$'s is that they can be thought of as "observed" R_i's. This isn't exactly correct since we are using $\hat{\mu}_i$ instead of μ_i in $\hat{r}_i$, but if the model is correct, then the $\hat{r}_i$'s should behave roughly like a random sample from the $G(0, \sigma)$ distribution. The $\hat{r}_i$'s do have some features that can be used to check the model assumptions. Recall that the maximum likelihood estimate of α is $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ which implies that $\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$ or

$$0 = \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i$$

so that the *average of the residuals is always zero.*

Residual plots can be used to check the model assumptions. Here are three residual plots which can be used:

(1) Plot points $(x_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$. If the model is satisfactory the points should lie more or less horizontally within a constant band around the line $\hat{r}_i = 0$ (see Figure 6.5).

(2) Plot points $(\hat{\mu}_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$. If the model is satisfactory the points should lie more or less horizontally within a constant band around the line $\hat{r}_i = 0$.

(3) Plot a Normal qqplot of the residuals $\hat{r}_i$. If the model is satisfactory the points should lie more or less along a straight line.

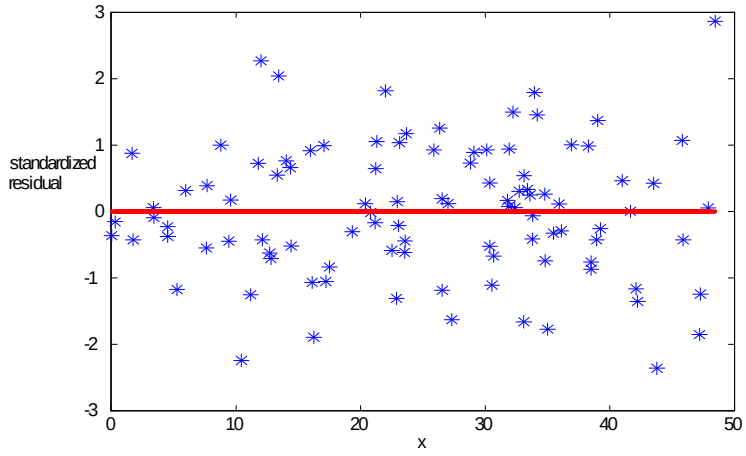


Figure 6.5: **Residual plot for example in which model assumptions hold**

Departures from the "expected" pattern may suggest problems with the model. For example, Figure 6.6 plot suggests the mean function $\mu_i = \mu(x_i)$ is not correctly specified. The pattern of points suggests that assuming a quadratic form for the mean such as $\mu(x_i) = \alpha + \beta x_i + \gamma x_i^2$ might give a better fit to the data than $\mu(x_i) = \alpha + \beta x_i$.

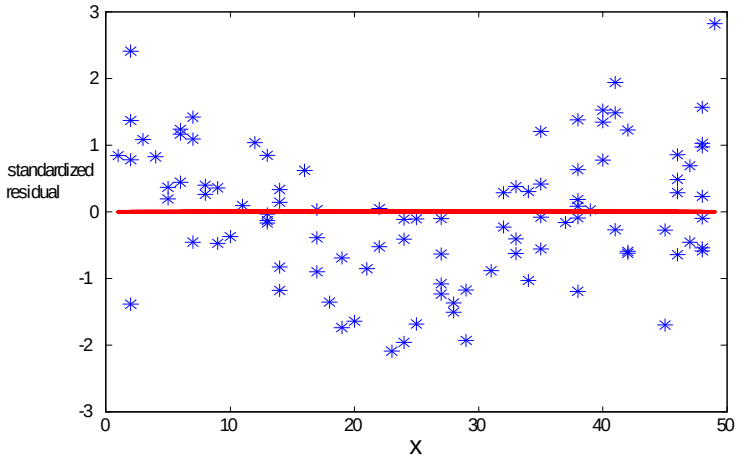


Figure 6.6: **Example of residual plot which indicates that assumption $E(Y_i) = \alpha + \beta x_i$ does not hold**

Figure 6.7 suggests that for these data the variance is non-constant. Sometimes transforming the response can solve this problem. Transformations such as $\log y$ and $\sqrt{y}$ are frequently used.

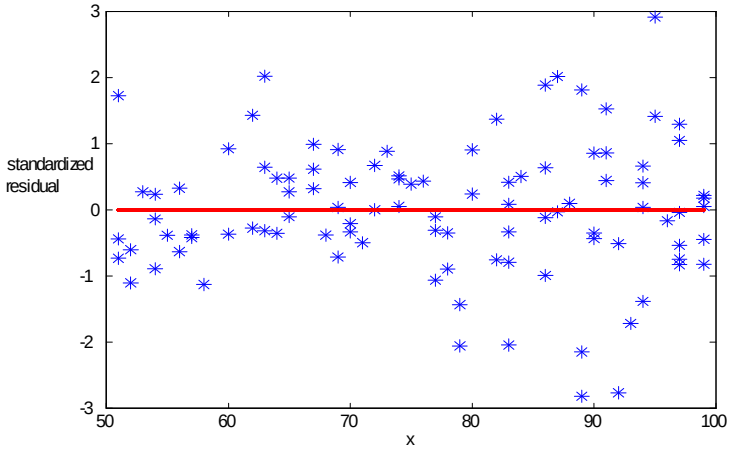


Figure 6.7: **Example of residual plot which indicates that assumption $Var(Y_i) = \sigma^2$ does not hold**

Reading these plots requires practice. You should try not to read too much into plots particularly if the plots are based on a small number of points.

Often we prefer to use standardized residuals

$$\begin{aligned} \hat{r}_i^* &= \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} \\ &= \frac{y_i - \hat{\alpha} - \hat{\beta} x_i}{s_e} \quad \text{for } i = 1, 2, \ldots, n. \end{aligned}$$

Standardized residuals were used in Figures 6.6 and 6.7. The patterns in the plots are unchanged whether we use $\hat{r}_i$ or $\hat{r}_i^*$, however the $\hat{r}_i^*$ values tend to lie in the interval $[-3, 3]$. The reason for this is that, 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 $Z \backsim G(0, 1)$, then roughly 99.73% of the observations should lie in the interval $[-3, 3]$.

Example 6.2.1 Revisited Strength of steel bolts

Figure 6.8 shows a standardized residual plot for the steel bolt data where the explanatory variate is diameter squared. No deviation from the expected pattern is observed. This is of course also evident from Figure 6.4.

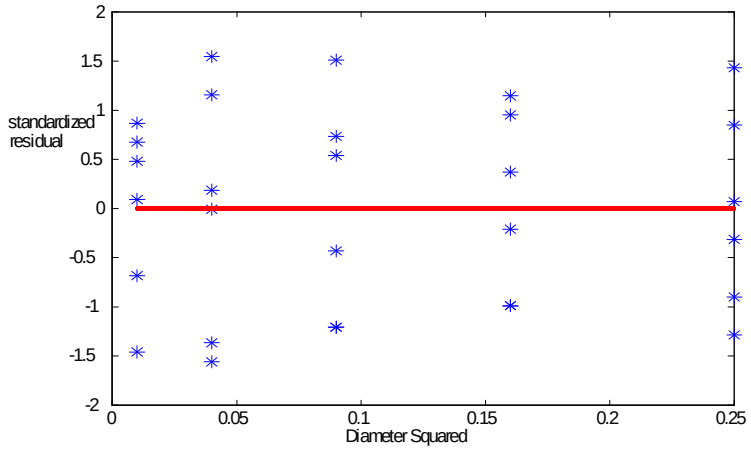


Figure 6.8: **Standard residuals versus diameter squared for bolt data**

A qqplot of the standardized residuals is given in Figure 6.9. Since the points lie reasonably along a straight line the Gaussian assumption seems reasonable. Remember that, since the quantiles of the Normal distribution change more rapidly in the tails of the distribution, we expect the points at both ends of the line to lie further from the line.

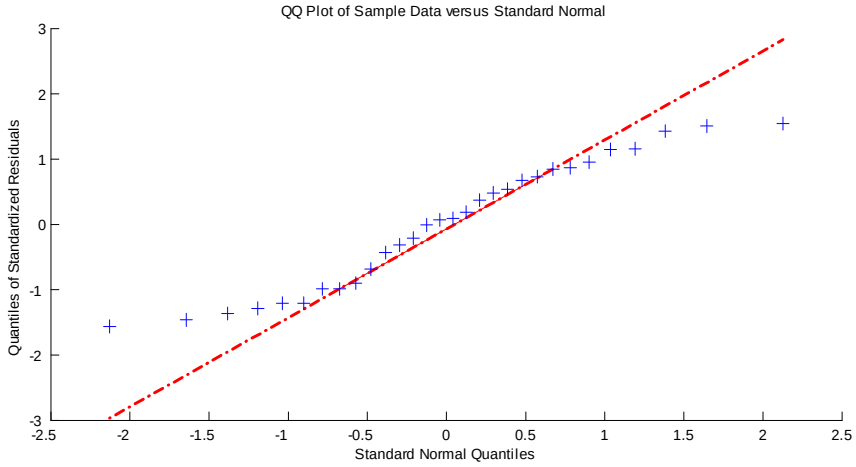


Figure 6.9: **Qqplot of standardized residuals for bolt data**

6.3 Comparing the Means of Two Populations

Two Gaussian Populations with Common Variance

Suppose $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from the $G(\mu_1, \sigma)$ distribution and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from the $G(\mu_2, \sigma)$ 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

$$L(\mu_1, \mu_2, \sigma) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(y_{ji} - \mu_j\right)^2\right].$$

Maximization of the likelihood function gives the maximum likelihood estimators:

$$\tilde{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} = \bar{Y}_1,$$
$$\tilde{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{2i} = \bar{Y}_2,$$
$$\text{and } \tilde{\sigma}^2 = \frac{1}{n_1 + n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji} - \tilde{\mu}_j)^2.$$

An estimator of the variance σ^2 (sometimes referred to as the *pooled estimator of variance*) adjusted for the degrees of freedom is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
$$= \frac{n_1 + n_2}{n_1 + n_2 - 2}\tilde{\sigma}^2$$

where

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2, \quad j = 1, 2$$

are the sample variances obtained from the individual samples. The estimator S_p^2 can be written as a *weighted average* of the estimators S_j^2. In fact

$$S_p^2 = \frac{w_1 S_1^2 + w_2 S_2^2}{w_1 + w_2} \tag{6.18}$$

where the weights are $w_j = n_j - 1$. Although you could substitute weights other than $n_j - 1$ in (6.18)[42], when you pool various estimators in order to obtain one that is better than any of those being pooled, you should do so with weights that relate to a measure of precision of the estimators. For sample variances, the number of degrees of freedom is such an indicator.

We will use the estimator S_p^2 for σ^2 rather than $\tilde{\sigma}^2$ since $E\left(S_p^2\right) = \sigma^2$.

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 the $G(\mu_1, \sigma)$ distribution and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from the $G(\mu_2, \sigma)$ distribution then*

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \backsim t\left(n_1 + n_2 - 2\right)$$

and

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2 \backsim \chi^2\left(n_1 + n_2 - 2\right)$$

Confidence intervals or tests of hypothesis for $\mu_1 - \mu_2$ and σ can be obtained by using these pivotal quantities. In particular, a $100p\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.19}$$

where $P(-a \leq T \leq a) = p$ and $T \sim t\left(n_1 + n_2 - 2\right)$.

[42] you would most likely be tempted to use $w_1 = w_2 = 1/2$.

Example 6.3.1 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

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} \backsim t\,(22)\,. \tag{6.20}$$

We have assumed[43] that the two population variances are identical, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, with σ^2 estimated by

$$s_p^2 = \frac{1}{22}\left[\sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2\right].$$

To test $H_0 : \mu_1 - \mu_2 = 0$ we use the test statistic

$$D = \frac{\left|\overline{Y}_1 - \overline{Y}_2 - 0\right|}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{\left|\overline{Y}_1 - \overline{Y}_2\right|}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}}$$

From the data above we find

$$n_1 = 12 \quad \bar{y}_1 = 10.4 \quad \sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 = 14.08 \quad s_1^2 = 1.2800$$
$$n_2 = 12 \quad \bar{y}_2 = 9.0 \quad \sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2 = 38.64 \quad s_2^2 = 3.5127.$$

This gives $\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2 = 1.4$ and $s_p^2 = 2.3964$. The observed value of the test statistic is

$$d = \frac{|\bar{y}_1 - \bar{y}_2|}{s_p\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{1.4}{\sqrt{2.3964\left(\frac{1}{6}\right)}} = 2.22$$

[43] If the sample variances differed by a great deal we would not make this assumption. Unfortunately if the variances are not assumed equal the problem becomes more difficult.

with

$$p - value = P(|T| \geq 2.22) = 2\left[1 - P\left(T \leq 2.22\right)\right] = 0.038$$

where $T \sim t\,(22)$. There is evidence based on the data against $H_0 : \mu_1 = \mu_2$.

Since $\bar{y}_1 > \bar{y}_2$, the indication is that paint A keeps its visibility better. A 95% confidence interval for $\mu_1 - \mu_2$ based on (6.20) is obtained using

$$\begin{aligned} 0.95 &= P(-2.074 \leq T \leq 2.074) \quad \text{where} \quad T \sim t\,(22) \\ &= P\left(-2.074 \leq \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} \leq 2.074)\right) \\ &= P\left(-2.074 S_p\sqrt{\frac{2}{12}} \leq \mu_1 - \mu_2 \leq 2.074 S_p\sqrt{\frac{2}{12}}\right). \end{aligned}$$

This gives the 95% confidence interval for $\mu_1 - \mu_2$ as

$$\hat{\mu}_1 - \hat{\mu}_2 \pm 2.074 s_p\sqrt{\frac{1}{12} + \frac{1}{12}} \quad \text{or} \quad [0.09,\ 2.71]\,.$$

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. (Look at $\hat{\mu}_1 = \bar{y}_1 = 14.08$ and $\hat{\mu}_2 = \bar{y}_2 = 9.0$. The relative differences are of the order of 10%).

Remark: The *R* function `t.test` will carry out the test above and will give confidence intervals for $\mu_1 - \mu_2$. This can be done with the command `t.test(y1,y2,var.equal=T)`, where y_1 and y_2 are the data vectors from 1 and 2.

Two Gaussian Populations with Unequal Variances

The procedures above assume that the two Gaussian distributions have the same standard deviations. Sometimes this isn't a reasonable assumption (it can be tested using a likelihood ratio test, but we will not do this here) and we must assume that $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from the $G(\mu_1, \sigma_1)$ distribution and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from the $G(\mu_2, \sigma_2)$ but $\sigma_1 \neq \sigma_2$. In this case there is no exact pivotal quantity which can be used to obtain a confidence interval for the difference in means $\mu_1 - \mu_2$. However the random variable

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{6.21}$$

has approximately a $G\,(0, 1)$ distribution, especially if n_1, n_2 are both large.

To illustrate its use, consider the durability of paint example, where $s_1^2 = 1.2800$ and $s_2^2 = 3.5127$. These appear quite different but they are in squared units and n_1, n_2 are

small; the standard deviations $s_1 = 1.13$ and $s_2 = 1.97$ do not provide evidence against the hypothesis that $\sigma_1 = \sigma_2$ if a likelihood ratio test is carried out. Nevertheless, let us use (6.21) to obtain a 95% confidence interval for $\mu_1 - \mu_2$. This resulting approximate 95% confidence interval is

$$\bar{y}_1 - \bar{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.22}$$

For the given data this equals 1.4 ± 1.24, or $[0.16, 2.64]$ which is not much different than the interval obtained assuming the two Gaussian distributions have the same standard deviations.

Example 6.3.2 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:

$$\begin{array}{lllll} \text{District 1:} & n_1 = 278 & \bar{y}_1 = 60.2 & s_1 = 10.16 \\ \text{District 2:} & n_2 = 345 & \bar{y}_2 = 58.1 & s_2 = 9.02 \end{array}$$

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 (6.22) to construct an approximate 95% confidence interval for $\mu_1 - \mu_2$. We obtain

$$60.2 - 58.1 \pm 1.96\sqrt{\frac{(10.16)^2}{278} + \frac{(9.02)^2}{345}} = 2.1 \pm (1.96)(0.779) \quad \text{or} \quad [0.57, 1.63]\,.$$

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 $H_0 : \mu_1 = \mu_2$, suggesting 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 and the distributions suggested for the two groups using histograms or boxplots.

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 $H_0 : \mu_1 = \mu_2$. However, it would

be a mistake[44] to conclude that the actual difference in the two distributions is very large. Unfortunately, "significant" tests like this are often used to make claims about one group or class or school is "superior" to another and such conclusions are unwarranted if, as is often the case, the assumptions of the test are not satisfied.

Comparing 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 6.3.3 Heights of males versus females [45]

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.[46]

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, so to estimate $\mu_1 - \mu_2$ the method in the preceding section should not be used since it required that we have **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 is obtained by plotting the points (Y_{1i}, Y_{2i}) in a scatterplot.

Example 6.3.4 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 litres) for the i'th car with the S and E fuels, respectively. We want to estimate $E(Y_{1i} - Y_{2i})$. The fuel

[44] We assume independence of the sample. How likely is it that marks in a class are independent of one another and no more alike than marks between two classes or two different years?

[45] See the video at www.watstat.ca called "Paired Confidence Intervals".

[46] Ask yourself "If I had (another?) brother/sister, how tall would they grow to?"

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. As in the preceding example it would not be appropriate to treat the Y_{1i}'s ($i = 1, 2, \ldots, 50$) and Y_{2i}'s ($i = 1, 2, \ldots, 50$) as two independent samples from larger populations. The observations have been paired deliberately to eliminate some factors (like driver/ car size) which might otherwise effect 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 either fuel.

Two types of Gaussian models are used to represent settings involving 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. This 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[47] (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;

$$Y_{1i} - Y_{2i} \sim N(\mu_1 - \mu_2, \sigma^2) \tag{6.23}$$

where $\sigma^2 = Var(Y_{1i}) + Var(Y_{2i}) - 2Cov(Y_{1i}, Y_{2i})$. Thus, if we are interested in estimating or testing $\mu_1 - \mu_2$, we can do this by considering the *within-pair differences* $Y_i = Y_{1i} - Y_{2i}$ and using the methods for a single Gaussian model in Section 6.2.

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 (6.23), since

$$\begin{aligned} E(Y_{1i} - Y_{2i}) &= \mu_1 - \mu_2 \quad \text{(note that the } \alpha_i\text{'s cancel)} \\ Var(Y_{1i} - Y_{2i}) &= \sigma_1^2 + \sigma_2^2 \end{aligned}$$

This model seems relevant for Example 6.3.2, where α_i refers to the i'th car type.

Thus, whenever we encounter paired data in which the variation in variables Y_{1i} and Y_{2i} is adequately modeled by Gaussian distributions, we will make inferences about $\mu_1 - \mu_2$ by working with the model (6.23).

[47] **For STAT 241:** Let $Y = (Y_1, Y_2, \ldots, Y_k)^T$ be a $k \times 1$ random vector with $E(Y_i) = \mu_i$ and $Cov(Y_i, Y_j) = \sigma_{ij}$, $i, j = 1, 2, \ldots, k$. (Note: $Cov(Y_i, Y_i) = \sigma_{ii} = Var(Y_i) = \sigma_i^2$.) Let $\mu = (\mu_1, \mu_2, \ldots, \mu_k)^T$ be the mean vector and Σ be the $k \times k$ symmetric covariance matrix whose (i, j) entry is σ_{ij}. Suppose also that Σ^{-1} exists. If the joint p.d.f. of $(Y_1, Y_2, \ldots, Y_k)$ is given by $f(y_1, \ldots, y_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]$, $y \in \Re^k$where $y = (y_1, y_2, \ldots, y_k)^T$ then Y is said to have a *Multivariate Normal distribution.* The case $k = 2$ is called the Bivariate Normal distribution.

Example 6.3.3 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

$$\bar{y} = 4.895 \text{ inches and } s^2 = \frac{1}{1400}\sum_{i=1}^{1401}(y_i - \bar{y})^2 = 6.5480 \text{ (inches)}^2.$$

Using the pivotal quantity

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

which has a $t\,(1400)$ distribution, a two-sided 95% confidence interval for $\mu = E(Y_i)$ is given by $\bar{y} \pm 1.96s/\sqrt{n}$ where $n = 1401$. (Note that $t\,(1400)$ is indistinguishable from $G(0, 1)$.) This gives the 95% confidence interval 4.895 ± 0.134 inches or $[4.76,\ 5.03]$ inches.

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. Presumably $\mu_1' = \mu_1$ (i.e. the average height of all adult males equals the average height of all adult males who also have an adult sister) and similarly $\mu_2' = \mu_2$, so $E(Y_i)$ does represent this difference. This is true provided that the males in the sibling pairs are randomly sampled from the population of all adult males, and similarly the females, but it might be worth checking.

Recall our earlier Example 1.3.2 involving the difference in the average heights of males and females in New Zealand. This gave the estimate $\hat{\mu} = \bar{y}_1 - \bar{y}_2 = 68.72 - 64.10 = 4.62$ inches, which is a little less than the difference in the example above. This is likely due to the fact that we are considering two distinct populations, but it should be noted that the New Zealand data are not paired.

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. This is the case in both Examples 6.3.3 and 6.3.4, so the pairing in these studies is a good idea.

To illustrate this further, in Example 6.3.3 the height measurement on the 1401 males gave $\bar{y}_1 = 69.720$ and $s_1^2 = 7.3861$ and the height measurements on the females gave $\bar{y}_2 = 64.825$ and $s_2^2 = 6.7832$. If the males and females were two independent samples (this is not quite right because the heights for the brother-sister combinations are not independent, but the sample means and variances are close to what we would get if we **did** have completely independent samples), then we could use (6.22) to construct an approximate 95% confidence

interval for $\mu_1 - \mu_2$. For the given data we obtain

$$69.720 - 64.825 \pm 1.96\sqrt{\frac{7.3861}{1401} + \frac{6.7832}{1401}}$$
$$\text{or} \quad [4.70, 5.09]\,.$$

We note that it is slightly wider than the 95% confidence interval [4.76, 5.03] obtained using the pairings.

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

$$\bar{Y}_1 - \bar{Y}_2$$

and we have that $E(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2$ and

$$\begin{aligned} Var(\bar{Y}_1 - \bar{Y}_2) &= Var(\bar{Y}_1) + Var(\bar{Y}_2) - 2Cov(\bar{Y}_1, \bar{Y}_2) \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\frac{\sigma_{12}}{n}, \end{aligned}$$

where $\sigma_{12} = Cov(Y_{1i}, Y_{2i})$. If $\sigma_{12} > 0$, then $Var(\bar{Y}_1 - \bar{Y}_2)$ 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 $Var(\bar{Y}_1 - \bar{Y}_2)$.

The following example involves an experimental study with pairing.

Example 6.3.5 Fibre in diet and cholesterol level[48]

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.

[48] From previous STAT 231 Course Notes of MacKay and Oldford.

Table 6.3: Cholesterol Levels on Two Diets

Subject	y_{1i}(High F)	y_{2i}(Low F)	y_i	Subject	y_{1i}(High F)	y_{2i}(Low F)	y_i
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 6.3 shows the cholesterol levels y (in mmol 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'll also assume that the differences are represented by the model

$$Y_i = Y_{1i} - Y_{2i} \sim G(\mu_1 - \mu_2, \sigma) \quad \text{for } i = 1, 2, \ldots, 20.$$

The differences y_i are also shown in Table 6.3, and from them we calculate the sample mean and standard deviation

$$\bar{y} = -0.020 \quad \text{and} \quad s = 0.411.$$

Since $P(T \leq 2.093) = 1 - 0.025 = 0.975$ where $T \sim t(19)$, a 95% confidence interval for $\mu_1 - \mu_2$ given by (6.19) is

$$\bar{y} \pm 2.093 \left(s/\sqrt{n}\right) = -0.020 \pm 2.093\,(0.411)/\sqrt{20} = -0.020 \pm 0.192 \quad \text{or} \quad [-0.212, 0.172]$$

This confidence interval includes $\mu_1 - \mu_2 = 0$, and there is clearly no evidence that the high fibre diet gives a lower cholesterol level at least in the time frame represented in this study.

Remark: The results here can be obtained using the R function `t.test`.

Exercise: Compute the p-value for the test of hypothesis $H_0 : \mu_1 - \mu_2 = 0$, 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.

6.4 More General Gaussian Response Models[49]

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:*

$$\tilde{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} \tag{6.24}$$

$$\textit{and} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \tilde{\mu}_i)^2 \quad \textit{where } \tilde{\mu}_i = \sum_{j=1}^{k} \tilde{\beta}_j x_{ij} \tag{6.25}$$

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

$$\sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = 0$$

[49] Optional

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

$$X^T(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$$
$$\text{or} \quad X^T\mathbf{y} = X^T X\boldsymbol{\beta}.$$

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[50] 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.

[50] 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

$$W = \frac{n\tilde{\sigma}^2}{\sigma^2} = \frac{(n-k)S_e^2}{\sigma^2} \tag{6.26}$$

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 (6.24) 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:[52] From Theorem 32 we can obtain confidence intervals and test hypotheses for the regression coefficients using the pivotal

$$\frac{\tilde{\beta}_j - \beta_j}{S_e\sqrt{c_j}} \sim t\,(n-k) \tag{6.27}$$

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 (6.27) 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

$$-a \leq \frac{\hat{\beta}_j - \beta_j}{s_e\sqrt{c_j}} \leq a$$

to obtain

$$\hat{\beta}_j - a s_e\sqrt{c_j} \leq \beta_j \leq \hat{\beta}_j + a s_e\sqrt{c_j}$$

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

$$\text{estimate} \ \pm a \times \text{estimated standard deviation of estimator.}$$

[52] 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.

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 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 (6.27) 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 (6.26) that

$$\frac{(n-1)S^2}{\sigma^2}$$

has a Chi-squared$(n-1)$ distribution.

6.5 Chapter 6 Problems

1. 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: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 12$.

   (a) Give 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. What is the interpretation of this interval?

   (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 $x = 50$.

2. Recall 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: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 20$ independently to these data.

   (b) Use the plots discussed in Section 6.2 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 $x = 100$?

3. Recall the steel bolt experiment in Example 6.2.1.

(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.

4. 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 6.2 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.

5. **Regression through the origin:** Consider the model $Y_i \backsim G(\beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ 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

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \backsim N\left(\beta,\ \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right).$$

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

$$\frac{\tilde{\beta} - \beta}{S_e / \sqrt{\sum_{i=1}^{n} x_i^2}} \backsim t\,(n-1)\,.$$

to construct a 95% confidence interval for β.

(e) Explain how to test the hypothesis $\beta = \beta_0$.

6. For the data in Problem 4

$$\sum_{i=1}^{20} x_i y_i = 13984.5554 \quad \sum_{i=1}^{20} x_i^2 = 14058.9097 \quad \sum_{i=1}^{20} y_i^2 = 13913.3833$$

   (a) Fit the model $Y_i \backsim G(\beta x_i, \sigma)$, $i = 1, 2, \ldots, 20$ independently to these data.

   (b) Plot a scatterplot of the data and the fitted line on the same plot. How well does the model through the origin fit the data?

   (c) Construct a 95% confidence intervals for the slope β and test the hypothesis $\beta = 1$

   (d) Let $\hat{\mu}_i = \hat{\beta} x_i$ and $\hat{r}_i^* = (y_i - \hat{\mu}_i)/s_e$. Plot the residual plots $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, 20$ and $(\hat{\mu}_i, \hat{r}_i^*)$,
   $i = 1, 2, \ldots, 20$. Plot a qqplot of the standardized residuals $\hat{r}_i^*$. Based on these plots as well as the scatterplot with the fitted line comment on how well the model fits the data.

   (e) Using the results of this analysis as well as the analysis in Problem 4 what would you conclude about using the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$ versus $Y_i \backsim G(\beta x_i, \sigma)$ for these data?

7. 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 has 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: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 46$ independently to these data.

   (b) Use the plots discussed in Section 6.2 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.

8. 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* psted 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 useful output.

```
#Import dataset skinfolddata.txt in folder S231Datasets using RStudio
x<-skinfolddata$Skinfold    # relabel Skinfold variate as x
y<-skinfolddata$BodyDensity # relabel Body Density variate as y
# run regression y = alpha+beta*x
RegModel<-lm(y~x)
# parameter estimates and p-value for test of no relationship
summary(RegModel)
n<-length(x)  # n=sample size
r<- RegModel$residuals  # get residuals
se<-sqrt(sum(r^2)/(n-2))  # estimate of sigma
se
# estimate of sigma can also be obtained using
# se<-summary(RegModel)$sigma
# Scatterplot of data with fitted line
muhat<-RegModel$fitted.values
# muhat is the vector of fitted responses
par(mfrow=c(2,2))  # creates a 2 by 2 plotting area
plot(x,y,xlab="Skinfold",ylab="Body Density")
title(main="Scatterplot with Fitted Line")
points(x,muhat,type="l")
#  Residual plots
rstar <- r/0.007877  # the standardized residuals
plot(x,rstar,xlab="Skinfold",ylab="Standardized Residual")
title(main="Residual vs Skinfold")
plot(muhat,rstar,xlab="Muhat",ylab="Standardized Residual")
title(main="Residual vs Muhat")
qqnorm(rstar,main="")
title(main="Qqplot of Residuals")
#  95% Confidence interval for slope
betahat<-RegModel$coefficients[2]  # estimate of slope
a<-qt(0.975,n-2)  # value from t table for 95% confidence interval
Sxx<-sum(x^2)-sum(x)^2/n  # value of Sxx
c(betahat-a*se/sqrt(Sxx),betahat+a*se/sqrt(Sxx))
#confidence intervals for alpha and beta can also be obtained using
# confint(RegModel,level=0.95)
```

```
#  95% Prediction interval for response at x=1.8
alphahat<-RegModel$coefficients[1]  # estimate of intercept
muhat18<-alphahat+betahat*1.8  # predicted value for x=1.8
xbar=mean(x)
pm<-a*se*sqrt(1+1/n+(1.8-xbar)^2/Sxx)
c(muhat18-pm,muhat18+pm)
# prediction interval at x=1.8 can also be obtained using
#predict(RegModel,data.frame("x"=1.8),interval="prediction",level=0.95)
```

(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 95% prediction interval for the body density of a male with skinfold measurement of 1.8?

(g) Do you think that the skinfold measurements provide a reasonable approximation to body density measurements?

9. 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.

(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		

The data are available in the file *boilingpointdata.txt* posted on the course website.

10. 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

$$Y_{1j} \backsim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 21 \text{ independently}$$

for the treatment group and independently

$$Y_{2j} \backsim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 23 \text{ independently}$$

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 Normality 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 treatmentvscontoldata.txt in folder S231Datasets using
RStudio
y<-treatmentvscontoldata$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=TRUE,
conf.level=0.95)
```

11. 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

$$Y_{1j} \backsim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 10 \text{ independently}$$

for the Type I bearings and independently

$$Y_{2j} \backsim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 10 \text{ independently}$$

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.

12. 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 x_i (for the math students) and y_i (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

$$Y_{1j} \backsim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 30 \text{ independently}$$

for the Math students and independently

$$Y_{2j} \backsim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 30 \text{ independently}$$

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.

13. 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.

14. 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.

15. 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

$$\bar{y}_2 = \tfrac{19.19}{12} = 1.599 \quad \sum_{i=1}^{12} (y_{2i} - \bar{y}_2)^2 = 0.35569$$

Analyze the data with the objective of determining whether there is any evidence that the dose of alcohol increases reaction time. Justify any models that you use.

16. 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?

17. 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

$$Y_i \backsim G(\mu, \sigma), \quad i = 1, 2, \ldots, 8 \quad \text{independently}$$

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 $H_0 : \mu = 0$.

18. **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.

19. **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 *sortingdata.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 sortingdata.txt in folder S231Datasets using RStudio
t.test(Time~Algorithm,data=sortingdata,paired=TRUE,conf.level=0.99)
t.test(Time~Algorithm,data=sortingdata,paired=FALSE,var.equal=TRUE,
conf.level=0.99)
```

20. **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.

21. **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?

22. **Challenge Problem: Least squares estimation**. 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

$$\mu_i = \mu(\mathbf{x}_i; \boldsymbol{\beta}) = \sum_{j=1}^{k} \beta_j x_{ij}.$$

23. **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 $0 < \theta_j < 1$ and $\sum_{j=1}^{k} \theta_j = 1$ so there are actually only $k-1$ unknown parameters in this model. 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 $\hat{\theta}_j = y_j/n$, $j = 1, 2, \ldots, k$.

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 $\boldsymbol{\alpha}$ is a vector of length $p < k - 1$. In other words, we assume that there are p parameters to be estimated in the hypothesized model determined by (7.3). 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. 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 $e_1, e_2, \ldots, e_k$ where $e_j = n\theta_j(\hat{\alpha})$, $j = 1, 2, \ldots, k$ 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

Recall Example 2.4.2, where people in a population are classified as being one of three blood types MM, MN, NN. The proportions of the population that are these three types are θ_1, θ_2, θ_3 respectively, with $\theta_1 + \theta_2 + \theta_3 = 1$. Genetic theory indicates, however, that the θ_j's can be expressed in terms of a single parameter α, as

$$\theta_1 = \alpha^2, \;\; \theta_2 = 2\alpha(1-\alpha), \;\; \theta_3 = (1-\alpha)^2. \tag{7.7}$$

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 that (7.7) is correct. (Note that $(Y_1, Y_2, Y_3) \sim$ Multinomial$(n; \theta_1, \theta_2, \theta_3)$ with $n = 100$.) The likelihood ratio test statistic is given by (7.5), but we have to find $\tilde{\alpha}$ and then the E_j's. The likelihood function under (7.7) is

$$\begin{aligned} L_1(\alpha) &= L(\theta_1(\alpha), \theta_2(\alpha), \theta_3(\alpha)) \\ &= c(\alpha^2)^{17}[2\alpha(1-\alpha)]^{46}[(1-\alpha)^2]^{37} \\ &= c\alpha^{80}(1-\alpha)^{120} \end{aligned}$$

where c is a constant. We easily find that $\hat{\alpha} = 0.40$. The observed expected frequencies under (7.7) are therefore $e_1 = 100\hat{\alpha}^2 = 16$, $e_2 = 100[2\hat{\alpha}(1-\hat{\alpha})] = 48$, $e_3 = 100[(1-\hat{\alpha})^2] = 36$. Clearly these are close to the observed frequencies $y_1 = 17$, $y_2 = 46$, $y_3 = 37$. 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

$$\begin{aligned} p - value &= P(\Lambda \geq 0.17; H_0) \approx P(W \geq 0.17) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq 0.41\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2(1 - 0.6591) = 0.6818 \end{aligned}$$

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 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

$$p_j(\theta) = \int_{a_{j-1}}^{a_j} f(t;\theta)dt = e^{-a_j/\theta} - e^{-a_{j-1}/\theta} \quad \text{for } j = 1, 2, ..., k \tag{7.8}$$

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$(n; p_1(\theta), p_2(\theta), \ldots, p_k(\theta))$ distribution with $n = 100$.

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

To calculate the expected frequencies we need an estimate of θ which is obtained by maximizing the likelihood function

$$L(\theta) = \prod_{j=1}^{7} [p_j(\theta)]^{y_j}.$$

It is possible to maximize $L(\theta)$ mathematically. (Hint: rewrite $L(\theta)$ in terms of the parameter $\beta = e^{-100/\theta}$ and find $\hat{\beta}$ first; then $\hat{\theta} = -100/\log\hat{\beta}$.) This gives $\hat{\theta} = 310.0$. 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 $k - 1 - p = 7 - 1 - 1 = 5$. The approximate $p - value$ is

$$p - value = P(\Lambda \geq 1.91; H_0) \approx P(W \geq 1.91) = 0.86 \quad \text{where } W \sim \chi^2(5)$$

so there is no evidence against the model (7.8).

The goodness of fit test just discussed 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 7.2.3 Goodness of fit and Poisson model

Recall the data in Example 2.6.1 collected by the physicists Rutherford and Geiger on the number of alpha particles omitted from a polonium source during 2608 time intervals each of length 1/8 minute. The data are given in Table 7.1 along with the expected frequencies calculated using the Poisson model with the mean θ estimated by the sample mean $\hat{\theta} = 3.8715$. In order to use the χ^2 approximation we have combined the last four classes so that the expected frequency in all classes is at least five.

Table 7.1: Frequency Table for Rutherford/Geiger Data

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	6	5.8
Total	2608	2607.9

The observed value of the likelihood ratio statistic (7.5) is

$$2\sum_{j=1}^{12} f_j \log\left(\frac{f_j}{e_j}\right) = 2\left[57\log\left(\frac{57}{54.3}\right) + 203\log\left(\frac{203}{210.3}\right) + \cdots + 6\log\left(\frac{6}{5.8}\right)\right] = 14.01$$

The degrees of freedom for the Chi-squared approximation equal $k-1-p = 12-1-1 = 10$. The approximate $p-value$ is

$$p-value = P(\Lambda \geq 14.01; H_0) \approx P(W \geq 14.01) = 0.17 \quad \text{where } W \sim \chi^2(10)$$

so there is no evidence against the hypothesis that a Poisson model fits these data.

The observed value of the goodness of fit statistic is

$$\sum_{j=1}^{12} \frac{(f_j - e_j)^2}{e_j} = \frac{(57-54.3)^2}{54.3} + \frac{(203-210.3)^2}{210.3} + \cdots + \frac{(6-5.9)^2}{5.9} = 12.96$$

and the approximate $p-value$ is

$$p-value = P(\Lambda \geq 12.96; H_0) \approx P(W \geq 12.96) = 0.23 \quad \text{where } W \sim \chi^2(10)$$

so again there is no evidence against the hypothesis that a Poisson model fits these data.

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 7.2.

Table 7.2: Frequency Table for Brake Pad 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

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. The degrees of freedom for the Chi-squared approximation equal $k - 1 - p = 9 - 1 - 1 = 7$. The approximate $p - value$ is

$$p - value = P(\Lambda \geq 50.36; H_0) \approx P(W \geq 50.36) \approx 0 \quad \text{where } W \sim \chi^2(7)$$

and there is very strong evidence against the hypothesis that an Exponential model fits these 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 at all. 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 $r_i = \sum_{j=1}^{b} y_{ij}$, $c_j = \sum_{i=1}^{a} y_{ij}$ 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

$$\alpha_i = P(\text{an individual is type } A_i)$$

and

$$\beta_j = P(\text{an individual is type } B_j)$$

and that (7.9) is the standard definition for independent events: $P(A_i \cap B_j) = P(A_i)P(B_j)$.

We recognize 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

$$L_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{a} \prod_{j=1}^{b} [\theta_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{y_{ij}}$$
$$= \prod_{i=1}^{a} \prod_{j=1}^{b} (\alpha_i \beta_j)^{y_{ij}}.$$

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 (7.5) for testing the hypothesis (7.9) is then

$$\lambda = 2 \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} \log \left(\frac{y_{ij}}{e_{ij}} \right).$$

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,

$$e_{11} = \frac{(244)(95)}{300} = 77.3, \quad e_{12} = \frac{244(116)}{300} = 94.4 \text{ and } e_{13} = \frac{244\,(61)}{300} = 49.6$$

The remaining expected frequencies can be obtained by subtraction and these are given in the table below in brackets next to the observed frequencies.

	O	A	B	AB	Total
Rh+	82 (77.3)	89 (94.4)	54 (49.6)	19 (22.8)	244
Rh−	13 (17.7)	27 (21.6)	7 (11.4)	9 (5.2)	56
Total	95	116	61	28	300

The degrees of freedom for the Chi-squared approximation are $(a-1)(b-1) = (3)\,(1) = 3$ 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 $\lambda = 8.52$, and the $p - value$ is approximately $P(W \geq 8.52) = 0.036$ where $W \sim \chi^2\,(3)$ so 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 get some idea about the lack of independence, or relationship, between the two classifications. We see here that the degree of dependence does not appear large.

Testing Equality of Multinomial Parameters from 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 Multinomial distributions, Multinomial$(n_i; \theta_{i1}, \theta_{i2}, \ldots, \theta_{ib})$.

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 7.3.2 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 7.3.3 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

We can think of the persons receiving aspirin and those receiving placebo as two groups, and test 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

$$e_{ij} = \frac{(y_{i+})(y_{+j})}{476} \quad \text{for } i = 1, 2.$$

This gives the expected frequencies shown in the table in brackets. The observed value of the likelihood ratio statistic is

$$2 \sum_{i=1}^{2} \sum_{j=1}^{2} y_{ij} \log \left(\frac{y_{ij}}{e_{ij}} \right) = 5.25$$

and the approximate $p-value$ is

$$\begin{aligned} p-value &= P(\Lambda \geq 5.25; H_0) \approx P(W \geq 5.25) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq 2.29\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2(1 - 0.98899) = 0.02202 \end{aligned}$$

so there is evidence against H_0 based on the 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 test can be followed up with estimates for θ_{11} and θ_{21}. Because each row of the table follows a Binomial distribution, we have

$$\hat{\theta}_{11} = \frac{y_{11}}{n_1} = \frac{64}{240} = 0.267 \text{ and } \hat{\theta}_{21} = \frac{y_{21}}{n_2} = \frac{86}{236} = 0.364.$$

We can also give individual confidence intervals for θ_{11} and θ_{21}. Based on methods derived earlier we have an approximate 95% confidence interval for θ_{11} given by

$$0.267 \pm 1.96\sqrt{\frac{(0.267)(0.733)}{240}} \text{ or } [0.211, 0.323]$$

and an approximate 95% confidence interval for θ_{11} given by

$$0.364 \pm 1.96\sqrt{\frac{(0.364)(0.636)}{240}} \text{ or } [0.303, 0.425].$$

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`.

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. The numbers 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:	64	71	42	18	4	1	0	200

Test whether a Poisson model for the number of interruptions Y on a single day is consistent with these data.

3. 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

Test the hypothesis that the number of defective items Y in a single carton has a Binomial$(12, \theta)$ distribution. Why might the Binomial not be a suitable model?

4. 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 1 had been singled out for presentation above because it gave the largest observed value of the likelihood ratio statistic Λ. What would you now do to test the hypothesis $\theta_j = 0.1$, $j = 0, 1, 2, \ldots, 9$? (Hint: Find $P(\text{largest of 6 independent } \Lambda\text{'s is} \geq \lambda)$.)

5. 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 $\theta_1 = \theta_2 = \theta_3 = \theta_4$.

6. 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

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. Construct a frequency table and test the goodness of fit of your model.

7. 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?

8. 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?

9. 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?

10. 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.

```
f<-matrix(c(18,28,19,20,51,28,12,25,9),ncol=3,byrow=TRUE)  # matrix of
observed frequencies
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)
```

11. 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.

12. 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.

8. CAUSAL RELATIONSHIPS

8.1 Establishing Causation

[54] 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 variables 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 hard 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 their risk. (Not everyone is helped by this medication.) However, on average the effect is to lower the risk of stroke. One way to measure this is by looking at the probability a randomly selected person has a stroke (say within 3 years) if they are given aspirin versus if they are not.

Therefore, a better idealized definition of causation is to say that changing x should result in a change in some attribute of the random variable Y (for example, its mean or some probability such as $P(Y > 0)$). 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

[54] See the video at www.watstat.ca called "Causation and the Flying Spaghetti monster".

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 variables may be controlled, than in observational studies. The following are brief examples.

Example 8.1.1 Strength of steel bolts

Recall Example 6.1.3 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 6.1.3, 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 variables (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 variables 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 variables 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": (i) X causes Y (meaning X has a causative effect on Y),(ii) Y causes X, and (iii) some other factor(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:

(i) Hold other possible explanatory variates fixed.

(ii) 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 variables 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.

- 84 cars of eight different types 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 two tests at each of the four 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 his or her 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 test drive. 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 Y. If these factors are also related to the explanatory variable 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 1 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 measured factors that 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 variable $Y = I($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[55] 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:

- 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.)

[55] *The Coronary Drug Research Group, New England Journal of Medicine (1980), pg. 1038.*

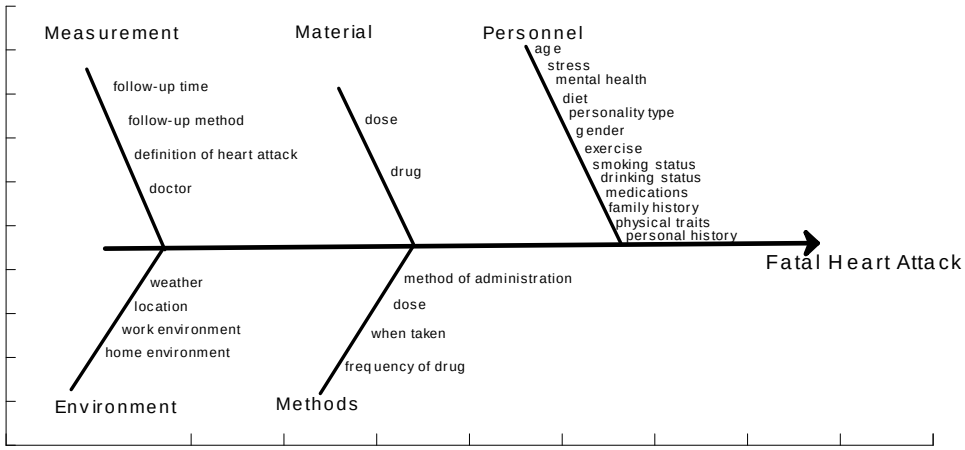


Figure 8.1: **Fishbone diagram for Chlofibrate example**

- 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:

- 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 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 $G(\mu_1, \sigma)$ and for the non-rust-proofed cars the responses were assumed to be $G(\mu_2, \sigma)$. 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.

4. In Problem 6.7 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?

5. Problem 6.9 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 a causative 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

Videos on sections: see **www.watstat.ca**

	p.f./p.d.f.	**Mean**	**Variance**	**m.g.f.**
Discrete	**p.f.**			
Binomial(n,p) $0<p<1,\ q=1-p$	$\binom{n}{y}p^y q^{n-y}$ $y=0,1,2,\ldots,n$	np	npq	$(pe^t+q)^n$
Bernoulli(p) $0<p<1,\ q=1-p$	$p^y(1-p)^{1-y}$ $y=0,1$	p	$p(1-p)$	(pe^t+q)
Negative Binomial(k,p) $0<p<1,\ q=1-p$	$\binom{y+k-1}{y}p^k q^y$ $y=0,1,2,\ldots$	$\frac{kq}{p}$	$\frac{kq}{p^2}$	$\left(\frac{p}{1-qe^t}\right)^k$ $t<-\ln q$
Geometric(p) $0<p<1,\ q=1-p$	pq^y $y=0,1,2,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{p}{1-qe^t}$ $t<-\ln q$
Hypergeometric(N,r,n) $r<N,\ n<N$	$\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$ $y=0,1,2,\ldots,\min(r,n)$	$\frac{nr}{N}$	$n\frac{r}{N}(1-\frac{r}{N})\frac{N-n}{N-1}$	intractible
Poisson(θ) $\theta>0$	$\frac{e^{-\theta}\theta^y}{y!}$ $y=0,1,\ldots$	θ	θ	$e^{\theta(e^t-1)}$
Multinomial$(n,\theta_1,\ldots\theta_k)$ $\theta_i>0,\ \sum_{i=1}^{k}\theta_i=1$	$\frac{n!}{y_1!y_2!\ldots y_k!}\theta_1^{y_1}\theta_2^{y_2}\ldots\theta_k^{y_k}$ $y_i=0,1,\ldots;\ \sum_{i=1}^{k}y_i=n$	$(n\theta_1,\ldots,n\theta_k)$	$Var(Y_i)=n\theta_i(1-\theta_i)$	
Continuous	**p.d.f.**			
Uniform(a,b)	$f(y)=\frac{1}{b-a},\ a<y<b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$ $t\neq 0$
Exponential(θ) $\theta>0$	$f(y)=\frac{1}{\theta}e^{-y/\theta},\ y>0$	θ	θ^2	$\frac{1}{1-\theta t}$ $t<1/\theta$
N(μ,σ^2) or G(μ,σ) $-\infty<\mu<\infty,\ \sigma>0$	$f(y)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(y-\mu)^2/(2\sigma^2)}$ $-\infty<y<\infty$	μ	σ^2	$e^{\mu t+\sigma^2t^2/2}$
Chi-squared(k) $k>0$	$f(y)=\frac{1}{2^{k/2}\Gamma(k/2)}y^{(k/2)-1}e^{-y/2},\ \ y>0$ where $\Gamma(a)=\int_0^\infty x^{a-1}e^{-x}dx$	k	$2k$	$(1-2t)^{-k/2}$ $t<1/2$
Student t $k>0$	$f(y)=c_k(1+\frac{y^2}{k})^{-(k+1)/2}$ $-\infty<y<\infty$ where $c_k=\Gamma(\frac{k+1}{2})/\sqrt{k\pi}\,\Gamma(\frac{k}{2})$	0 if $k>1$	$\frac{k}{k-2}$ if $k>2$	undefined

Formulae

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$
$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$	$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$
$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$	$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

Pivotals/Test Statistics

Random variable	Distribution	Mean or df	Standard Deviation
$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$	Gaussian	0	1
$\frac{(n-1)S^2}{\sigma^2}$	Chi-squared	df = $n-1$	
$\frac{\bar{Y}-\mu}{S/\sqrt{n}}$	Student t	df = $n-1$	
$\tilde{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i$	Gaussian	β	$\sigma\left(\frac{1}{S_{xx}}\right)^{1/2}$
$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}}$	Student t	df = $n-2$	
$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$	Gaussian	α	$\sigma\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]^{1/2}$
$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$	Gaussian	$\mu(x) = \alpha + \beta x$	$\sigma\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}$
$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$	Student t	df = $n-2$	
$Y - \tilde{\mu}(x)$	Gaussian	0	$\sigma\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}$
$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$	Student t	df = $n-2$	
$\frac{(n-2)S_e^2}{\sigma^2}$	Chi-squared	df = $n-2$	
$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	Student t	df = n_1+n_2-2	
$\frac{(n_1+n_2-2)S_p^2}{\sigma^2}$	Chi-squared	df = n_1+n_2-2	

Approximate Pivotals

$\frac{\tilde{\theta}-\theta}{\sqrt{\tilde{\theta}(1-\tilde{\theta})/n}} \backsim \mathrm{N}(0,1)$ approximately if $\tilde{\theta} = Y/n$ and $Y \backsim \mathrm{Binomial}(n,\theta)$

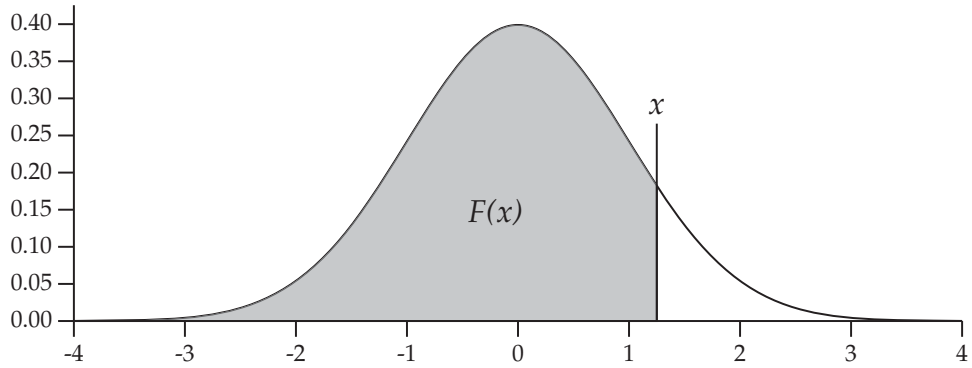
$\frac{\bar{Y}-\theta}{\sqrt{\bar{Y}/n}} \backsim \mathrm{N}(0,1)$ approximately for a random sample from $\mathrm{Poisson}(\theta)$ distribution

$\Lambda = -2\log[R(\theta)] = 2[\ell(\tilde{\theta}) - \ell(\theta)] \sim$ approximately $\chi^2(1)$, if n is large

$\Lambda = 2\sum_{j=1}^{k} Y_j \log(Y_j/E_j) \sim$ approximately $\chi^2(\mathrm{df})$ where df = $(k-1)$ - (no. of parameters estimated under H_0)

$D = \sum_{j=1}^{k}(Y_j - E_j)^2/E_j \sim$ approximately $\chi^2(\mathrm{df})$ where df = $(k-1)$ - (no. of parameters estimated under H_0)

Probabilities for Standard Normal *N(0,1)* Distribution



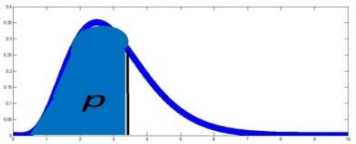
This table gives the values of $F(x)$ for $x \geq 0$

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	0.56750	0.57142	0.57534
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	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

This table gives the values of $F^{-1}(p)$ for $p \geq 0.50$

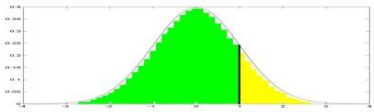
p	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.2019	0.2275
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4677	0.4959
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7722	0.8064
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1750	1.2265
0.9	1.2816	1.3408	1.4051	1.4758	1.5548	1.6449	1.7507	1.8808	2.0537	2.3263

CHI-SQUARED DISTRIBUTION QUANTILES



df\p	0.005	0.01	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.842	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.446	0.713	1.022	1.386	1.833	2.408	3.219	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.005	1.424	1.869	2.366	2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.649	2.195	2.753	3.357	4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.146	1.610	2.343	3.000	3.656	4.352	5.132	6.064	7.289	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.070	3.828	4.570	5.348	6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	3.822	4.671	5.493	6.346	7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	4.594	5.527	6.423	7.344	8.351	9.525	11.030	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.380	6.393	7.357	8.343	9.414	10.656	12.242	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.179	7.267	8.296	9.342	10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188
11	2.603	3.054	3.816	4.575	5.578	6.989	8.148	9.237	10.341	11.530	12.899	14.631	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	7.807	9.034	10.182	11.340	12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	8.634	9.926	11.129	12.340	13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	9.467	10.821	12.078	13.339	14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	10.307	11.721	13.030	14.339	15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.152	12.624	13.983	15.338	16.780	18.418	20.465	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.002	13.531	14.937	16.338	17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.391	10.865	12.857	14.440	15.893	17.338	18.868	20.601	22.760	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	13.716	15.352	16.850	18.338	19.910	21.689	23.900	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	14.578	16.266	17.809	19.337	20.951	22.775	25.038	28.412	31.410	34.170	37.566	39.997
25	10.520	11.524	13.120	14.611	16.473	18.940	20.867	22.616	24.337	26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928
30	13.787	14.953	16.791	18.493	20.599	23.364	25.508	27.442	29.336	31.316	33.530	36.250	40.256	43.773	46.979	50.892	53.672
35	17.192	18.509	20.569	22.465	24.797	27.836	30.178	32.282	34.336	36.475	38.859	41.778	46.059	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	29.051	32.345	34.872	37.134	39.335	41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766
45	24.311	25.901	28.366	30.612	33.350	36.884	39.585	41.995	44.335	46.761	49.452	52.729	57.505	61.656	65.410	69.957	73.166
50	27.991	29.707	32.357	34.764	37.689	41.449	44.313	46.864	49.335	51.892	54.723	58.164	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	50.641	53.809	56.620	59.335	62.135	65.227	68.972	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	59.898	63.346	66.396	69.334	72.358	75.689	79.715	85.527	90.531	95.023	100.430	104.210
80	51.172	53.540	57.153	60.391	64.278	69.207	72.915	76.188	79.334	82.566	86.120	90.405	96.578	101.880	106.630	112.330	116.320
90	59.196	61.754	65.647	69.126	73.291	78.558	82.511	85.993	89.334	92.761	96.524	101.050	107.570	113.150	118.140	124.120	128.300
100	67.328	70.065	74.222	77.929	82.358	87.945	92.129	95.808	99.334	102.950	106.910	111.670	118.500	124.340	129.560	135.810	140.170

Student t Quantiles



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

APPENDIX A: ANSWERS TO END OF CHAPTER PROBLEMS

Chapter 1

1.1 (a) The new mean is

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n}(a + by_i) = \frac{1}{n}\left(na + b\sum_{i=1}^{n} y_i\right) = a + b\bar{y}$$

and the new median is $\hat{m}_u = a + b\hat{m}$.

(b) There is no general result for the sample mean but if all $y_i \geq 0$ and n is an odd number then the new median is $\hat{m}^2$.

(c)

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

(d) Since

$$a(y_0) = \frac{1}{n+1}\left(y_0 + \sum_{i=1}^{n} y_i\right) = \frac{n\bar{y} + y_0}{n+1}$$

therefore

$$\lim_{y_0 \to \infty} a(y_0) = \lim_{y_0 \to \infty} \frac{n\bar{y} + y_0}{n+1} = \infty$$

and

$$\lim_{y_0 \to -\infty} a(y_0) = \lim_{y_0 \to -\infty} \frac{n\bar{y} + y_0}{n+1} = -\infty$$

This means that an additional very large (or very small) observation has a large effect on the sample mean.

(e) Case 1: If n is odd then $\hat{m} = \hat{m}(y_1, y_2, \ldots, y_n) = y_{\left(\frac{n+1}{2}\right)}$.
If $y_0 > y_{\left(\frac{n+1}{2}\right)+1}$ then there are now an even number of observations and the new median is $\hat{m}(y_0) = \frac{1}{2}\left(y_{\left(\frac{n+1}{2}\right)} + y_{\left(\frac{n+1}{2}\right)+1}\right)$. If $y_{\left(\frac{n+1}{2}\right)}$ and $y_{\left(\frac{n+1}{2}\right)+1}$ are close in value then the median will change by very little and the change does not depend on the value of y_0.
If $y_0 < y_{\left(\frac{n+1}{2}\right)-1}$ then there are now an even number of observations and the new

median is $\hat{m}(y_0) = \frac{1}{2}\left(y_{\left(\frac{n+1}{2}\right)-1} + y_{\left(\frac{n+1}{2}\right)}\right)$. If $y_{\left(\frac{n+1}{2}\right)-1}$ and $y_{\left(\frac{n+1}{2}\right)}$ are close in value then the median will change by very little and the change does not depend on the value of y_0.

Case 2: If n is even then $\hat{m} = \hat{m}(y_1, y_2, \ldots, y_n) = \frac{1}{2}\left(y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}\right)+1}\right)$.
If $y_0 > y_{\left(\frac{n}{2}+1\right)}$ then there are now an even number of observations and the new median is $\hat{m}(y_0) = y_{\left(\frac{n}{2}\right)+1}$. If $y_{\left(\frac{n}{2}\right)}$ and $y_{\left(\frac{n}{2}\right)+1}$ are close in value then the median will change by very little and the change does not depend on the value of y_0.
If $y_0 < y_{\left(\frac{n}{2}\right)}$ then there are now an even number of observations and the new median is $\hat{m}(y_0) = y_{\left(\frac{n}{2}\right)}$. If $y_{\left(\frac{n}{2}\right)}$ and $y_{\left(\frac{n}{2}\right)+1}$ are close in value then the median will change by very little and the change does not depend on the value of y_0.

(f) Unlike the sample mean the sample median is not affected by outliers (very large y_0 or very small y_0) so it is a more robust numerical summary of location. In many countries there are usually a few people with very large incomes. The mean income is affected by these few very large incomes so reporting the mean income rather than the median income would give the false impression that people are doing well in general with respect to income.

(g)

$$\frac{d}{d\mu}V(\mu) = -2\sum_{i=1}^{n}(y_i - \mu) = -2n(\bar{y} - \mu) = 0 \text{ if } \mu = \bar{y}$$

and by the First Derivative Test, $V(\mu)$ is minimized at $\mu = \bar{y}$.

1.2 (a)

$$\begin{aligned} s_u^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[a + by_i - (a + b\bar{y})]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(by_i - b\bar{y})^2 = \frac{b^2}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = b^2 s^2 \\ s_u &= |b|\, s \end{aligned}$$

$$IQR(u_1, \ldots, u_n) = |b|\, IQR$$

(b)

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

(c)

$$\begin{aligned}
s^2(y_0) &= \frac{1}{n}\left[\sum_{i=1}^{n} y_i^2 + y_0^2 - (n+1)\left(\frac{n\bar{y}+y_0}{n+1}\right)^2\right] \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n} y_i^2 + y_0^2 - \frac{1}{(n+1)}\left[n^2(\bar{y})^2 + 2n\bar{y}y_0 + y_0^2\right]\right\} \\
&= \frac{1}{n(n+1)}\left\{(n+1)\sum_{i=1}^{n} y_i^2 + (n+1)y_0^2 - n^2(\bar{y})^2 - 2n\bar{y}y_0 - y_0^2\right\} \\
&= \frac{1}{n(n+1)}\left\{n\left[\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2\right] + \sum_{i=1}^{n} y_i^2 + ny_0^2 - 2n\bar{y}y_0\right\} \\
&= \frac{(n-1)}{(n+1)}s^2 + \frac{1}{n(n+1)}\sum_{i=1}^{n} y_i^2 + \frac{1}{(n+1)}y_0(y_0 - 2\bar{y})
\end{aligned}$$

Therefore

$$\begin{aligned}
\lim_{y_0\to\pm\infty} s(y_0) &= \lim_{y_0\to\pm\infty}\left[\frac{(n-1)}{(n+1)}s^2 + \frac{1}{n(n+1)}\sum_{i=1}^{n} y_i^2 + \frac{1}{(n+1)}y_0(y_0 - 2\bar{y})\right]^{1/2} \\
&= \left[\frac{(n-1)}{(n+1)}s^2 + \frac{1}{n(n+1)}\sum_{i=1}^{n} y_i^2 + \frac{1}{(n+1)}\lim_{y_0\to\pm\infty} y_0(y_0 - 2\bar{y})\right]^{1/2} = \infty
\end{aligned}$$

This means that an additional very large (or very small) observation has a large effect on the sample standard deviation.

(d) Once y_0 is larger than $q(0.75)$ or smaller than $q(0.25)$, then y_0 has little effect on the interquartile range as y_0 increases or decreases.

1.3 Since

$$\begin{aligned}
\frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2\right]^{3/2}} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(by_i - b\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(by_i - b\bar{y})^2\right]^{3/2}} \\
&= \frac{b^3}{(b^2)^{3/2}}\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}} \\
&= \left[\frac{b}{|b|}\right]^3 g_1
\end{aligned}$$

Therefore $g_1(u_1, \ldots, u_n) = g_1$ if $b > 0$ and $g_1(u_1, \ldots, u_n) = -g_1$ if $b < 0$. In summary the magnitude of the sample skewness remains unchanged but the sample skewness changes sign if $b < 0$.

Since

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^2\right]^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(by_i-b\bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(by_i-b\bar{y})^2\right]^2}$$

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

therefore the sample kurtosis is the same for both data sets.

1.4 For the revenues:
sample mean $= (-7)(2500) + 1000 = -16500$
sample standard deviation $= |-7|(5500) = 38500$
sample median $= (-7)(2600) + 1000 = -17200$
sample skewness $= (-1)(1.2) = -1.2$
sample kurtosis $= 3.9$
range $= (7)(7500) = 52500$

1.5 (a) The relative frequency histogram of the piston diameters is given in Figure 12.1.

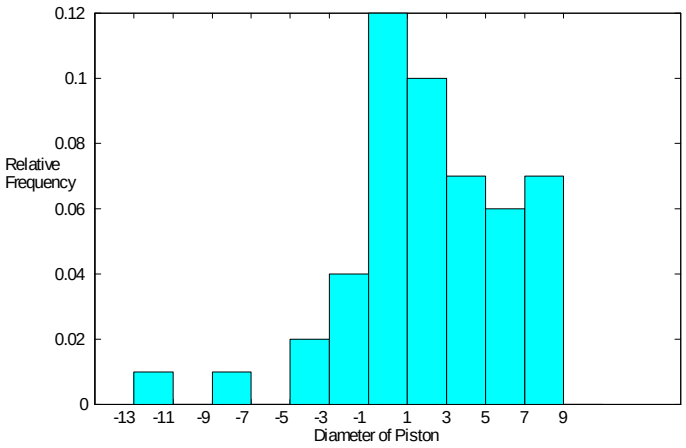


Figure 12.1: Histogram of Piston Diameters

(b) $\bar{y} = 100.7/50 = 2.014$, $\hat{m} = q(0.5) = \frac{1}{2}\left(y_{(25)} + y_{(26)}\right) = \frac{1}{2}(2.1 + 2.5) = 2.3$

(c) $s^2 = \frac{1}{49}\left[1110.79 - 50(2.014)^2\right] = 18.5302$, $s = 4.3047$,
$q(0.25) = \frac{1}{2}\left(y_{(12)} + y_{(13)}\right) = \frac{1}{2}[-0.7 + (-0.6)] = -0.65$
$q(0.75) = \frac{1}{2}\left(y_{(38)} + y_{(39)}\right) = \frac{1}{2}[5.1 + 5.4] = 5.25$
$IQR = 5.25 - (-0.65) = 5.9$

(d) The five number summary is: -12.8, -0.65, 2.3, 5.25, 8.9

(e) $Ppk = 0.6184$

(f) If $\bar{y} \approx \pm 10$ then $Ppk \approx 0$. Values of $\bar{y}$ less than -10 or bigger than $+10$ indicate that performance is poor. If $\bar{y} \approx 0$ then $Ppk \approx 10/3s$. Recall that for Normal data we would expect approximately 99% of the observed data to lie between $\mu - 3\sigma \approx \bar{y} - 3s$ and $\mu + 3\sigma \approx \bar{y} + 3s$. Therefore if $\bar{y} \approx 0$ and $3s \approx 10$ or $10/3s \approx 1$ then this indicates that performance is good. Therefore $Ppk \approx 10/3s = 1$ indicates good performance.

(g) Let $Y \backsim G(2.014, 4.3047)$ then

$$P(\text{diameters out of specification}) = 1 - P(-10 < Y < 10) = 0.03408$$

1.6

1.7 The data from smallest to largest are: 1.1 3.9 4.3 4.5 5.2 6.3 7.2 7.6 8.5 14.0

$$\begin{aligned} q(0.25) &= \frac{1}{2}\left(y_{(3)} + y_{(4)}\right) = \frac{1}{2}(4.3 + 4.5) = 4.4 \\ q(0.5) &= \frac{1}{2}\left(y_{(5)} + y_{(6)}\right) = \frac{1}{2}(5.2 + 6.3) = 5.75 \\ q(0.75) &= \frac{1}{2}\left(y_{(8)} + y_{(9)}\right) = \frac{1}{2}(7.6 + 8.5) = 8.05 \\ IQR &= q(0.75) - q(0.25) = 8.05 - 4.4 = 3.65 \\ q(0.25) - 1.5 \times IQR &= 4.4 - 1.5(3.65) = -1.075 \\ q(0.25) + 1.5 \times IQR &= 4.4 + 1.5(3.65) = 13.525 \end{aligned}$$

The top of the box is at 8.05, the bottom is at 4.4 and the line inside the box is at 5.75. The upper whisker is at 8.5 and the lower whisker is at 1.1. There is one outlier at 14.0.

The empirical c.d.f. is a step function which jumps a height of 0.1 at each of the points: 1.1 3.9 4.3 4.5 5.2 6.3 7.2 7.6 8.5 14.0

1.9 (a) The relative frequency histograms are given in Figure 12.2.

(b) Five number summary for female coyotes: 71.0 85.5 89.75 93.25 102.5
Five number summary for male coyotes: 78.0 87.0 92.0 96.0 105.0

(c) The boxplots are shown in Figure 12.3.

(d) Female coyotes: $\bar{x} = 89.24$, $s_1 = 6.5482$
Male coyotes: $\bar{y} = 92.06$, $s_2 = 6.6960$
In Figure 12.2 we note that the Normal p.d.f. fits the female lengths better than the male lengths.

(e) We note that the e.c.d.f. for the male coyotes is to the right of the e.c.d.f. for the female coyotes which would indicate that male lengths of coyotes are generally larger than females as you might expect.

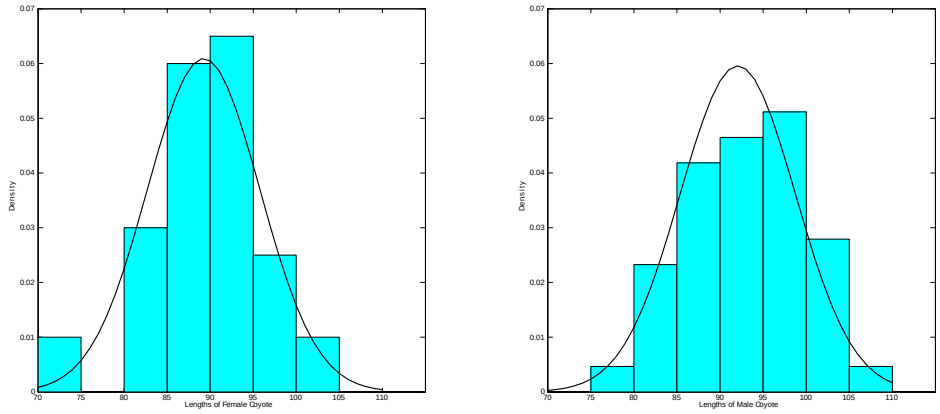


Figure 12.2: Histograms for lengths of female and male coyotes

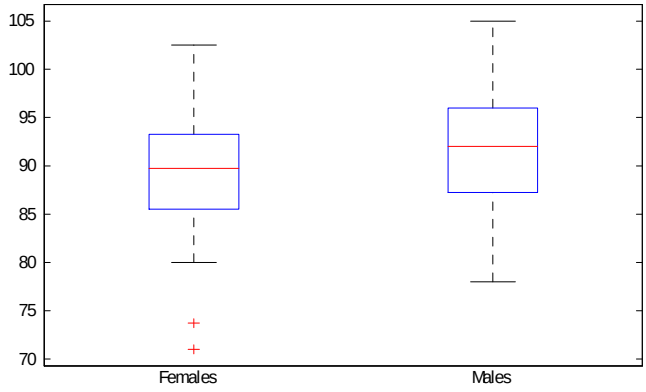


Figure 12.3: Boxplots for lengths of female and male coyotes

1.10 (a) The two variates are Value (x) and Gross (y), where Value is the average amount the actor's movies have made (in millions of U.S. dollars), and Gross is the amount of the highest grossing movie in which the actor played as a major character (in millions of U.S. dollars). Since the goal is to study the effect of an actor's value (x) on the amount grossed in a movie (y), we choose x as the explanatory variate and y as the response variate.

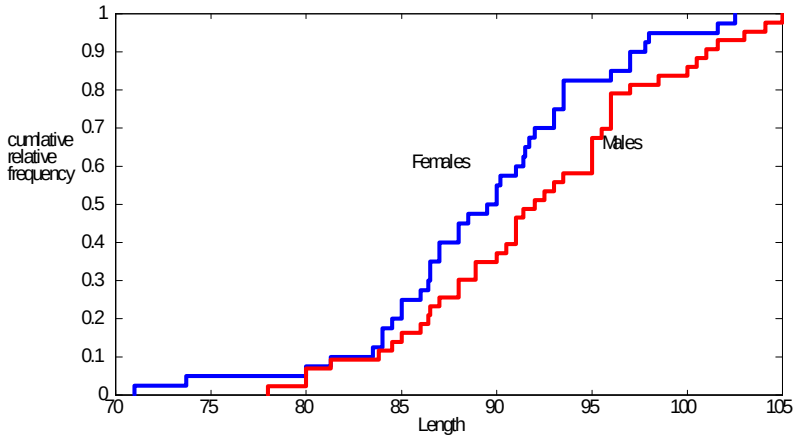


Figure 12.4: E.c.d.f.'s for lengths of female and male coyotes

(b) A scatterplot of the data is given in Figure 12.5.

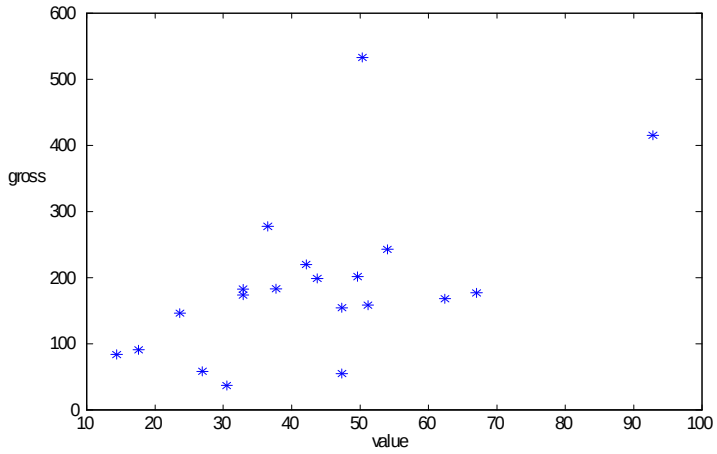


Figure 12.5: Scatterplot of gross versus value

(c) The sample correlation is

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{184540.93 - 20(860.6/20)(3759.5/20)}{\left[43315.04 - 20\,(860.6/20)^2\right]^{1/2}\left[971560.19 - 20\,(3759.5/20)^2\right]^{1/2}} \\ &= 0.558 \end{aligned}$$

There is a moderately strong positive linear relationship between x and y.

(d) In this example we do not have enough evidence to conclude that a causal relationship exists. Another plausible explanation for the observed data is that there is a third variate such as "the talent of the actor" that affects both the Value (x) and Gross(y) (of course it is very difficult to measure the variate "talent"). Consequently, x and y are expected to be positively correlated, and this is what we observe in this data set.

1.11 (a) The two-way table is:

	Cold	No Cold	Total
Vitamin C	20	80	100
Placebo	30	70	100
Total	50	150	200

(b) The relative risk of a cold in the vitamin C group as compared to the placebo group equals $\frac{20/(20+80)}{30/(30+70)} = \frac{2}{3}$.

(c) The group taking Vitamin C are only two-thirds as likely to catch a cold as compared to the placebo group which might suggest that taking Vitamin C is associated with fewer colds. (More on this in Chapter 7.)

1.12 (a) Since the n people are selected at random from a large population it is reasonable to assume that the people are independent and that the probability a randomly chosen person has blood type A is equal to θ. Therefore we have a sequence of n independent trials (people) with two outcomes on each trial (success = person has blood type A, failure = person does not have blood type A) and $P(\text{Success}) = \theta$. The probability function for the random variable Y = the number of people with blood type A = number of successes in n Bernoulli trials is given by the Binomial distribution:

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

(b) Since $Y \backsim \text{Binomial}(n, \theta)$ then $E(Y) = n\theta$ and $Var(Y) = n\theta(1-\theta)$.

(c)

$$P(Y = 20) = \binom{n}{20} \theta^{20} (1-\theta)^{30} \quad 0 < \theta < 1.$$

(d) A reasonable estimate of θ is given by the proportion of observed successes in $n = 50$ trials which is $20/50 = 0.4$. An estimate of the probability that in a sample of $n = 10$ there will be at least one person with blood type A is given by

$$\begin{aligned} & 1 - \binom{10}{0} (0.4)^0 (0.6)^{10} = 1 - (0.6)^{10} \\ = \; & 0.9940 \end{aligned}$$

(e) If y successes are observed in n Bernoulli trials then a reasonable estimate of θ is given by the (sample) proportion of successes, that is, a reasonable estimate of θ is y/n.

(f)

$$\begin{aligned} E\left(\frac{Y}{n}\right) &= \frac{1}{n}E(Y) = \frac{1}{n}(n\theta) = \theta \\ Var\left(\frac{Y}{n}\right) &= \left(\frac{1}{n}\right)^2 Var(Y) = \left(\frac{1}{n}\right)^2 [n\theta(1-\theta)] = \frac{\theta(1-\theta)}{n} \to 0 \text{ as } n \to \infty. \end{aligned}$$

For large values of n, Y/n should be close to θ. By the Central Limit Theorem

$$\begin{aligned} & P\left(\frac{Y}{n} - 1.96\sqrt{\frac{\theta(1-\theta)}{n}} \le \theta \le \frac{Y}{n} + 1.96\sqrt{\frac{\theta(1-\theta)}{n}}\right) \\ = & P\left(\left|\frac{\frac{Y}{n} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right| \le 1.96\right) \approx P(|Z| \le 1.96) \quad \text{where } Z \backsim N(0,1) \\ = & 2P(Z \le 1.96) - 1 = 2(0.975) - 1 = 0.95 \end{aligned}$$

(g) Since there are now four possible outcomes on each independent trial the joint distribution of $Y_1 =$ no. of A types, $Y_2 =$ no. of B types, $Y_3 =$ no. of AB types, $Y_4 =$ no. of O types is given by the Multinomial distribution.

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4) = \frac{n!}{y_1!y_2!y_3!y_4!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}\theta_4^{y_4}$$

$$\begin{aligned} \text{for } y_i &= 0, 1, \ldots, n; \quad i = 1,2,3,4 \quad \sum_{i=1}^{4} y_i = n \\ \text{and } 0 &< \theta_i < 1; \quad i = 1,2,3,4 \quad \sum_{i=1}^{4} \theta_i = 1. \end{aligned}$$

(h) Since we observe outcome A, y_1 times in a sample of n people a reasonable estimate of $\theta_1 =$ proportion of type A in the large population is given by the sample proportion y_1/n. Similarly a reasonable estimate of θ_i is y_i/n for $i = 2,3,4$.

1.13 (a) Since $Y \backsim G(\mu, \sigma)$ the probability density function of Y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \text{ for } y \in \Re,\ \mu \in \Re,\ \sigma > 0,$$

(b) Since $Y \backsim G(\mu, \sigma)$, $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

(c) A reasonable estimate of the mean μ is the sample mean

$$\bar{y} = \frac{1916}{16} = 119.75.$$

A reasonable estimate of the variance σ^2 is the sample variance

$$s^2 = \frac{1}{15}\left[231618 - 16\left(119.75\right)^2\right] = 145.1\dot{3}.$$

An estimate of the probability that a random chosen UWaterloo Math student will have an IQ greater than 120 is given by

$$\begin{aligned} P\left(Y \geq 120\right) \quad \text{where } Y &\backsim N\left(119.75, 145.1\dot{3}\right) \\ &= P\left(Z \geq \frac{120 - 119.75}{\sqrt{145.1\dot{3}}}\right) \quad \text{where } Z \backsim N\left(0,1\right) \\ &= P\left(Z \geq 0.0208\right) \approx P\left(Z \geq 0.02\right) = 1 - 0.50798 \\ &= 0.49202. \end{aligned}$$

(i) The distribution of a linear combination of Gaussian (Normal) random variables has a Gaussian (Normal) distribution. Since

$$\begin{aligned} E\left(\bar{Y}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\left(n\mu\right) = \mu \\ \text{and } \; Var\left(\bar{Y}\right) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var\left(Y_i\right) \quad \text{since } Y_i's \text{ are independent r.v.'s} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\sigma^2 = \left(\frac{1}{n}\right)^2\left(n\sigma^2\right) = \frac{\sigma^2}{n} \end{aligned}$$

therefore $\bar{Y} \backsim G\left(\mu, \sigma/\sqrt{n}\right)$.

$$Var\left(\bar{Y}\right) = \frac{\sigma^2}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

For large values of n, the sample mean $\bar{Y}$ should be close to the mean μ.

(ii)

$$\begin{aligned} & P\left(\bar{Y} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{Y} + 1.96\sigma/\sqrt{n}\right) \\ = \; & P\left(\left|\bar{Y} - \mu\right| \leq 1.96\sigma/\sqrt{n}\right) = P\left(|Z| \leq 1.96\right) \quad \text{where } Z \backsim G\left(0,1\right) \\ = \; & 2P\left(Z \leq 1.96\right) - 1 \\ = \; & 2\left(0.975\right) - 1 = 0.95 \end{aligned}$$

(iii) We want $P\left(\left|\bar{Y} - \mu\right| \leq 1.0\right) \geq 0.95$ where $\bar{Y} \backsim G\left(\mu, 12/\sqrt{n}\right)$ or

$$\begin{aligned} P\left(\left|\bar{Y} - \mu\right| \leq 1.0\right) &= P\left(\frac{\left|\bar{Y} - \mu\right|}{12/\sqrt{n}} \leq \frac{1.0}{12/\sqrt{n}}\right) \\ &= P\left(|Z| \leq \frac{\sqrt{n}}{12}\right) \geq 0.95 \quad \text{where } Z \backsim G\left(0,1\right). \end{aligned}$$

Since $P(|Z| \leq 1.96) = 0.95$ we want $\sqrt{n}/12 \geq 1.96$ or
$n \geq (1.96)^2 (144) = 553.2$. Therefore $n = 554$.

1.14 (a) Since $Y \backsim$ Exponential(θ) the probability density function of Y is

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

(b) Since $Y \backsim$ Exponential(θ), $E(Y) = \theta$ and $Var(Y) = \theta^2$.

(c) A reasonable estimate of the mean θ is the sample mean $\bar{y} = 1325.1/20 = 66.255$. An estimate of

$$P(Y > 100) = \int_{100}^{\infty} \frac{1}{\theta} e^{-y/\theta} dy = e^{-100/\theta} \quad \text{is} \quad e^{-100/66.255} = 0.2211.$$

(i)

$$\begin{aligned} E\left(\bar{Y}\right) &= \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \sum_{i=1}^{n} \theta = \frac{1}{n} (n\theta) = \theta \\ \text{and} \quad Var\left(\bar{Y}\right) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(Y_i) \quad \text{since } Y_i's \text{ are independent r.v.'s} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \theta^2 = \left(\frac{1}{n}\right)^2 \left(n\theta^2\right) = \frac{\theta^2}{n} \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

For large values of n, the sample mean $\bar{Y}$ should be close to the mean θ.

(ii) By the Central Limit Theorem

$$\begin{aligned} & P\left(\bar{Y} - 1.6449\theta/\sqrt{n} \leq \theta \leq \bar{Y} + 1.6449\theta/\sqrt{n}\right) \\ = \; & P\left(\left|\frac{\bar{Y} - \theta}{\theta/\sqrt{n}}\right| \leq 1.6449\right) \approx P(|Z| \leq 1.6449) \quad \text{where } Z \backsim N(0,1) \\ = \; & 2P(Z \leq 1.6449) - 1 = 2(0.95) - 1 = 0.9 \end{aligned}$$

1.15 (a) Since $Y \backsim$ Poisson(θ) the probability density function of Y is

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

(b) Since $Y \backsim$ Poisson(θ), $E(Y) = \theta$ and $Var(Y) = \theta$.

(c) Let $Y_i =$ no. of accidents on day i, $i = 1, 2, \ldots, 6$. Then

$$\begin{aligned} & P(Y_1 = 0, Y_2 = 2, Y_3 = 0, Y_4 = 1, Y_5 = 3, Y_6 = 1) \\ = \; & P(Y_1 = 0) P(Y_2 = 2) P(Y_3 = 0) P(Y_4 = 1) P(Y_5 = 3) P(Y_6 = 1) \\ = \; & \left(\frac{\theta^0 e^{-\theta}}{0!}\right)\left(\frac{\theta^2 e^{-\theta}}{2!}\right)\left(\frac{\theta^0 e^{-\theta}}{0!}\right)\left(\frac{\theta^1 e^{-\theta}}{1!}\right)\left(\frac{\theta^3 e^{-\theta}}{3!}\right)\left(\frac{\theta^1 e^{-\theta}}{1!}\right) \\ = \; & \frac{\theta^7 e^{-6\theta}}{12} \quad \text{for} \quad \theta > 0. \end{aligned}$$

A reasonable estimate of the mean θ is the sample mean $\bar{y} = 7/6$. An estimate of the probability that there is at least one accident at this intersection next Wednesday is given by

$$1 - \frac{(7/6)^0 e^{-7/6}}{0!} = 1 - e^{-7/6} = 0.6886.$$

(i)

$$\begin{aligned} E\left(\bar{Y}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\theta = \frac{1}{n}\left(n\theta\right) = \theta \\ \text{and } Var\left(\bar{Y}\right) &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var\left(Y_i\right) \text{ since } Y_i's \text{ are independent r.v.'s} \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\theta = \left(\frac{1}{n}\right)^2 (n\theta) = \frac{\theta}{n} \to 0 \text{ as } n \to \infty. \end{aligned}$$

For large values of n, the sample mean $\bar{Y}$ should be close to the mean θ.

(ii) By the Central Limit Theorem

$$\begin{aligned} & P\left(\bar{Y} - 1.96\sqrt{\theta/n} \le \theta \le \bar{Y} + 1.96\sqrt{\theta/n}\right) \\ = \; & P\left(\left|\frac{\bar{Y}-\theta}{\sqrt{\theta/n}}\right| \le 1.96\right) \approx P\left(|Z| \le 1.96\right) \text{ where } Z \backsim N\left(0,1\right) \\ = \; & 2P\left(Z \le 1.96\right) - 1 \\ = \; & 2\left(0.975\right) - 1 \\ = \; & 0.95 \end{aligned}$$

(a) $E\left(Y_i^2\right) = Var\left(Y_i\right) + \left[E\left(Y_i\right)\right]^2 = \sigma^2 + \mu^2$.

(b)

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\left(n\mu\right) = \mu$$

Since the $Y_i's$ are independent random variables

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var\left(Y_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\sigma^2 \\ &= \left(\frac{1}{n}\right)^2 \left(n\sigma^2\right) = \frac{\sigma^2}{n} \\ E\left[(\bar{Y})^2\right] &= \left[E\left(\bar{Y}\right)\right]^2 + Var\left(\bar{Y}\right) = \mu^2 + \frac{\sigma^2}{n} \end{aligned}$$

(c)

$$\begin{aligned}
E\left(S^2\right) &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left(Y_i^2\right) - nE\left[\left(\bar{Y}\right)^2\right]\right\} \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right) - n\left(\mu^2+\frac{\sigma^2}{n}\right)\right] \\
&= \frac{1}{n-1}\left[n\left(\mu^2+\sigma^2\right) - n\mu^2 - \sigma^2\right] \\
&= \frac{1}{n-1}\left[\left(n-1\right)\sigma^2\right] = \sigma^2
\end{aligned}$$

Chapter 2

2.1 (a)

$$\begin{aligned} G(\theta) &= \theta^a (1-\theta)^b, \quad 0 < \theta < 1 \\ g(\theta) &= \log G(\theta) = a \log \theta + b \log (1-\theta), \quad 0 < \theta < 1 \\ g'(\theta) &= \frac{a}{\theta} - \frac{b}{1-\theta} = \frac{a(1-\theta) - b\theta}{\theta(1-\theta)} = \frac{a-(a+b)\theta}{\theta(1-\theta)} \\ g'(\theta) &= 0 \text{ if } \theta = \frac{a}{a+b} \end{aligned}$$

Since $g'(\theta) > 0$ for $0 < \theta < \frac{a}{a+b}$ and $g'(\theta) < 0$ for $1 > \theta > \frac{a}{a+b}$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = \frac{a}{a+b}$.

(b)

$$\begin{aligned} G(\theta) &= \theta^{-a} e^{-b/\theta}, \quad \theta > 0 \\ g(\theta) &= \log G(\theta) = -a \log \theta - \frac{b}{\theta}, \quad \theta > 0 \\ g'(\theta) &= \frac{-a}{\theta} + \frac{b}{\theta^2} = \frac{-a\theta + b}{\theta^2} \\ g'(\theta) &= 0 \text{ if } \theta = \frac{b}{a} \end{aligned}$$

Since $g'(\theta) > 0$ for $0 < \theta < \frac{b}{a}$ and $g'(\theta) < 0$ for $\theta > \frac{b}{a}$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = \frac{b}{a}$.

(c)

$$\begin{aligned} G(\theta) &= \theta^a e^{-b\theta}, \quad \theta > 0 \\ g(\theta) &= \log G(\theta) = a \log \theta - b\theta, \quad \theta > 0 \\ g'(\theta) &= \frac{a}{\theta} - b = \frac{a - b\theta}{\theta} \\ g'(\theta) &= 0 \text{ if } \theta = \frac{a}{b} \end{aligned}$$

Since $g'(\theta) > 0$ for $0 < \theta < \frac{a}{b}$ and $g'(\theta) < 0$ for $\theta > \frac{a}{b}$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = \frac{a}{b}$.

(d)

$$\begin{aligned} G(\theta) &= e^{-a(\theta-b)^2}, \quad \theta \in \Re \\ g(\theta) &= \log G(\theta) = -a(\theta-b)^2, \quad \theta \in \Re \\ g'(\theta) &= -2a(\theta-b) \\ g'(\theta) &= 0 \text{ if } \theta = b \end{aligned}$$

Since $g'(\theta) > 0$ for $\theta < b$ and $g'(\theta) < 0$ for $\theta > b$ then by the First Derivative Test $g(\theta)$ has a maximum value at $\theta = b$.

2.2 (a) The probability of the observed results for Experiment 1 is

$$\begin{aligned} & P\left(\text{total number of individuals examined } = 100; \theta\right) \\ = & \binom{99}{9} \theta^{10} \left(1-\theta\right)^{90} \quad \text{for } 0 < \theta < 1. \end{aligned}$$

The probability of the observed results for Experiment 2 is

$$\begin{aligned} & P\left(10 \text{ individuals with blood type B } ; \theta\right) \\ = & \binom{100}{10} \theta^{10} \left(1-\theta\right)^{90} \quad \text{for } 0 < \theta < 1. \end{aligned}$$

(b) The likelihood function in both cases simplifies to

$$L\left(\theta\right) = \theta^{10} \left(1-\theta\right)^{90} \quad \text{for } 0 < \theta < 1.$$

if we ignore constants with respect to θ. The log likelihood function is

$$l\left(\theta\right) = 10 \log \theta + 90 \log\left(1-\theta\right) \quad \text{for } 0 < \theta < 1.$$

Now

$$\begin{aligned} l'\left(\theta\right) &= \frac{10}{\theta} - \frac{90}{1-\theta} \\ &= \frac{10 - 100\theta}{\theta\left(1-\theta\right)} = 0 \text{ if } \theta = \frac{10}{100} = 0.1 \end{aligned}$$

and the maximum likelihood estimate of θ is $\hat{\theta} = 0.1$.

(c) Let Y = the number of donors with blood type B. Then $Y \backsim \text{Binomial}(n, 0.1)$ and $E\left(Y\right) = 0.1n$ and $Var\left(Y\right) = 0.1\left(0.9\right)n = 0.09n$. We went to find n such that $P\left(Y \geq 10\right) \geq 0.90$. By the Normal approximation to the Binomial we have

$$P\left(Y \geq 10\right) \approx P\left(Z \geq \frac{9.5 - 0.1n}{\sqrt{0.09n}}\right) \text{ where } Z \backsim N\left(0,1\right).$$

Since $P\left(Z \geq -1.2816\right) = 0.90$ we solve

$$\frac{9.5 - 0.1n}{\sqrt{0.09n}} = -1.2816$$

or

$$n^2 - 204.78n + 9025 = 0$$

which gives $n = 140.6$ so we take $n = 141$. Using *gbinom()* or *pbinom()* we find $n = 140$.

2.3 (a) The likelihood function is

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

and the log likelihood is

$$l(\theta) = n(\bar{y}-1)\log\theta + n\log(1-\theta) \quad \text{for} \quad 0 < \theta < 1.$$

Solving

$$l'(\theta) = \frac{n(\bar{y}-1)}{\theta} - \frac{n}{1-\theta} = \frac{n(\bar{y}-1)(1-\theta) - n\theta}{\theta(1-\theta)} = 0$$

gives the maximum likelihood estimate

$$\hat{\theta} = \frac{\bar{y}-1}{\bar{y}}.$$

(b) The relative likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{n(\bar{y}-1)}(1-\theta)^n}{\hat{\theta}^{n(\bar{y}-1)}\left(1-\hat{\theta}\right)^n} = \left[\left(\frac{\theta}{\hat{\theta}}\right)^{(\bar{y}-1)}\left(\frac{1-\theta}{1-\hat{\theta}}\right)\right]^n \quad \text{for} \quad 0 < \theta < 1.$$

If $n = 200$ and $\sum_{i=1}^{200} y_i = 400$ then

$$\begin{aligned} \bar{y} &= \frac{400}{200} = 2, \quad \hat{\theta} = \frac{2-1}{2} = 0.5 \\ \text{and} \quad R(\theta) &= \left[\left(\frac{\theta}{0.5}\right)\left(\frac{1-\theta}{0.5}\right)\right]^{200} \quad \text{for} \quad 0 < \theta < 1. \end{aligned}$$

A graph of $R(\theta)$ is given in Figure 12.6.

(c) Since $p = P(Y = 1; \theta) = (1-\theta)$ then by the Invariance property of maximum likelihood estimates the maximum likelihood estimate of p is $\hat{p} = \left(1-\hat{\theta}\right) = 1 - 0.5 = 0.5$.

2.4 (a) Since $t = 1$, the likelihood function is

$$L(\theta) = \prod_{i=1}^{10} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \left(\prod_{i=1}^{10} y_i!\right)^{-1} \theta^{41} e^{-10\theta} \quad \text{for } \theta > 0$$

or more simply

$$L(\theta) = \theta^{41} e^{-10\theta} \quad \text{for } \theta > 0.$$

The log likelihood function is

$$l(\theta) = 41\log\theta - 10\theta \text{ for } \theta > 0.$$

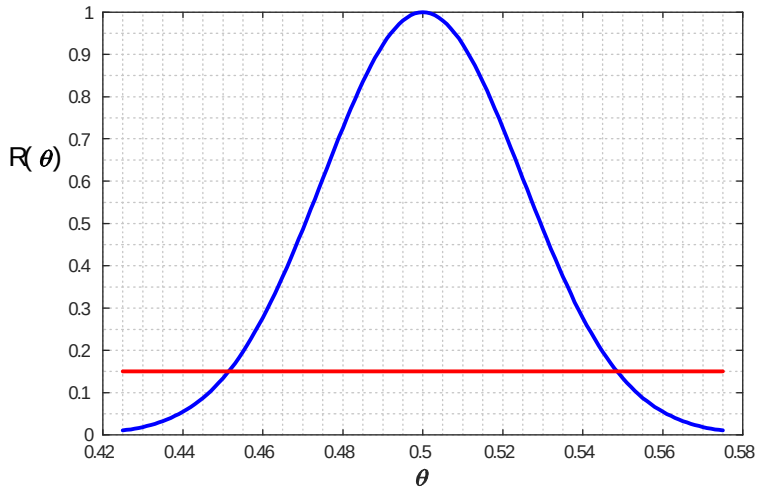


Figure 12.6: Relative likelihood function for fracture data

Solving

$$l'(\theta) = \frac{41}{\theta} - 10 = 0$$

gives the maximum likelihood estimate $\hat{\theta} = 4.1$.

(b) Since

$$p = P(\text{no transactions in a two minute interval }; \theta) = \frac{(2\theta)^0 e^{-2\theta}}{0!} = e^{-2\theta}$$

then by the invariance property of maximum likelihood estimates the maximum likelihood estimate of p is $\hat{p} = e^{-2\hat{\theta}} = 0.000275$.

2.5 (a) The joint p.d.f. of the observations $y_1, y_2, ..., y_n$ is given by

$$\begin{aligned}\prod_{i=1}^{n} f(y_i;\theta) &= \prod_{i=1}^{n} \frac{2y_i}{\theta} e^{-y_i^2/\theta} \quad \text{for } \theta > 0 \\ &= 2^n (\prod_{i=1}^{n} y_i) \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2\right).\end{aligned}$$

The likelihood function is

$$L(\theta) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2\right) \quad \theta > 0$$

and the log likelihood is

$$l(\theta) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} y_i^2 \quad \theta > 0.$$

Solving

$$l'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} y_i^2 = \frac{1}{\theta^2}\left(\sum_{i=1}^{n} y_i^2 - n\theta\right) = 0$$

gives the maximum likelihood estimate

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

(b) The relative likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \left(\frac{\hat{\theta}}{\theta}\right)^n e^{n\left(1-\hat{\theta}/\theta\right)} = \left[\frac{\hat{\theta}}{\theta}e^{\left(1-\hat{\theta}/\theta\right)}\right]^n \quad \theta > 0.$$

If $n = 20$ and $\sum_{i=1}^{20} y_i^2 = 72$ then $\hat{\theta} = 72/20 = 3.6$. A graph of $R(\theta)$ is given in Figure 12.7.

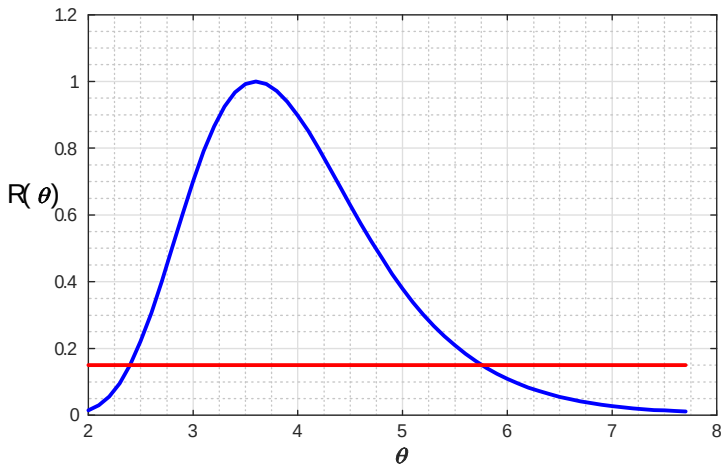


Figure 12.7: $R(\theta)$ for Problem 2.5

2.6 (a) If σ is known then the likelihood function of μ is

$$\begin{aligned}
L(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\
&= (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\
&= (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(y_i-\bar{y})+(\bar{y}-\mu)\right]^2\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right]\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right]
\end{aligned}$$

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

Ignoring constants with respect to μ we have

$$L(\mu) = \exp\left[-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right] \quad \text{for } \mu \in \Re.$$

The log likelihood function is

$$l(\mu) = -\frac{n}{2\sigma^2}(\bar{y} - \mu)^2 \quad \text{for } \mu \in \Re$$

and

$$l'(\mu) = \frac{n}{\sigma^2}(\bar{y} - \mu) = 0 \quad \text{if} \quad \mu = \bar{y}$$

and therefore the maximum likelihood estimate of μ is $\hat{\mu} = \bar{y}$ which does not depend on σ.

(b) If μ is known then the likelihood function of σ is

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

or more simply

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

The log likelihood function is

$$l(\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \quad \text{for } \sigma > 0$$

and

$$l'(\sigma) = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)^2 = \frac{1}{\sigma^3}\left[-n\sigma^2 + \sum_{i=1}^{n}(y_i - \mu)^2\right] = 0$$

if

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Therefore the maximum likelihood estimate of σ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2}$$

which does depend on μ.

2.7 (a) The likelihood function

$$L(\theta) = \prod_{i=1}^{n} (\theta+1) y_i^{\theta} = (\theta+1)^n \left(\prod_{i=1}^{n} y_i \right)^{\theta} \quad \text{for } \theta > -1$$

The log likelihood function is

$$l(\theta) = n \log(\theta+1) + \theta \sum_{i=1}^{n} \log(y_i) \quad \text{for } \theta > -1$$

Solving

$$\frac{d}{d\theta} l(\theta) = \frac{n}{1+\theta} + \sum_{i=1}^{n} \log(y_i) = 0$$

gives

$$\hat{\theta} = \frac{n}{-\sum_{i=1}^{n} \log(y_i)} - 1.$$

(b) The log relative likelihood function is

$$r(\theta) = l(\theta) - l(\hat{\theta}) = n \log\left(\frac{\theta+1}{\hat{\theta}+1} \right) + (\theta - \hat{\theta}) \sum_{i=1}^{n} \log(y_i) \quad \text{for } \theta > -1.$$

If $n = 15$ and $\sum_{i=1}^{15} \log(y_i) = -34.5$ then $\hat{\theta} = \left(\frac{15}{34.5}\right) - 1 = -0.5652$. The graph of $r(\theta)$ is given in Figure 12.8.

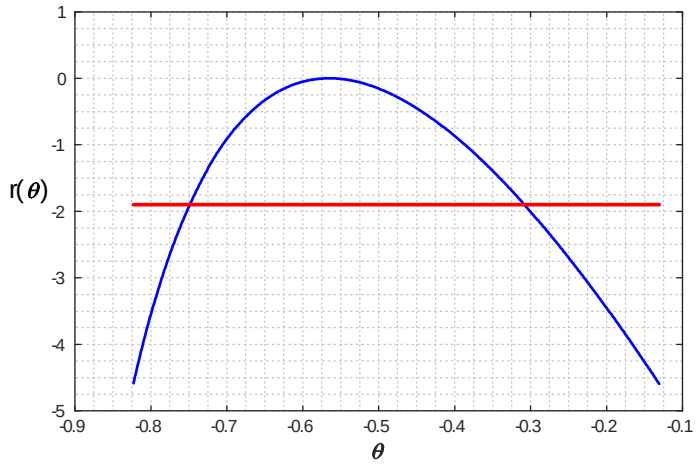


Figure 12.8: $r(\theta)$ for Problem 2.6

2.8 (a)

$$\begin{aligned} P(MM) &= P(FF) = \left(\frac{1}{2}\right)\alpha + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(1-\alpha) = \frac{1+\alpha}{4} \\ P(MF) &= 1 - 2\left(\frac{1+\alpha}{4}\right) = \frac{1-\alpha}{2} \end{aligned}$$

where α = probability the pair is identical.

(b)

$$L(\alpha) = \frac{n!}{n_1!n_2!n_3!}\left(\frac{1+\alpha}{4}\right)^{n_1}\left(\frac{1+\alpha}{4}\right)^{n_2}\left(\frac{1-\alpha}{2}\right)^{n_3} \quad \text{where } n = n_1+n_2+n_3$$

or more simply

$$L(\alpha) = (1+\alpha)^{n_1+n_2}(1-\alpha)^{n_3}.$$

Maximizing $L(\alpha)$ gives $\hat{\alpha} = (n_1+n_2-n_3)/n$. For $n_1 = 16$, $n_2 = 16$ and $n_3 = 18$, $\hat{\alpha} = 0.28$.

2.9 (a) The likelihood function based on the Poisson model and the frequency table is

$$\begin{aligned} L(\theta) &= \frac{696!}{69!155!171!143!79!57!14!6!2!0!} \\ &\quad \times \left(\frac{\theta^0 e^{-\theta}}{0!}\right)^{69}\left(\frac{\theta^1 e^{-\theta}}{1!}\right)^{155}\left(\frac{\theta^2 e^{-\theta}}{2!}\right)^{171}\left(\frac{\theta^3 e^{-\theta}}{3!}\right)^{143}\left(\frac{\theta^4 e^{-\theta}}{4!}\right)^{79} \\ &\quad \times \left(\frac{\theta^5 e^{-\theta}}{5!}\right)^{57}\left(\frac{\theta^6 e^{-\theta}}{6!}\right)^{14}\left(\frac{\theta^7 e^{-\theta}}{7!}\right)^{6}\left(\frac{\theta^8 e^{-\theta}}{8!}\right)^{2}\left(\sum_{y=9}^{\infty}\frac{\theta^y e^{-\theta}}{y!}\right)^{0} \end{aligned}$$

or more simply

$$\begin{aligned} L(\theta) &= \theta^{0(69)+1(155)+2(171)+3(143)+4(79)+5(57)+6(14)+7(6)+8(2)} \\ &\quad \times e^{-(69+155+171+143+79+57+14+6+2)\theta} \\ &= \theta^{1669}e^{-696\theta}, \quad \theta > 0. \end{aligned}$$

(b) The log likelihood function is

$$l(\theta) = 1669\log\theta - 696\theta, \quad \theta > 0$$

and

$$l'(\theta) = \frac{1669}{\theta} - 696 = \frac{1669-696\theta}{\theta} = 0 \text{ if } \theta = \frac{1669}{696} \approx 2.3980.$$

The maximum likelihood estimate of θ is $\hat{\theta} = 1669/696$.

(c) The expected frequencies are calculated using

$$e_y = 696 \cdot \frac{\left(\frac{1669}{696}\right)^y e^{-1669/696}}{y!}, \quad y = 0, 1, \ldots, 8$$

and are given in the table below:

Number of Points in a Game: y	Observed Number of Games with y points: f_y	Expected Number of Games with y points: e_y
0	69	63.27
1	155	151.71
2	171	181.90
3	143	145.40
4	79	87.17
5	57	41.81
6	14	16.71
7	6	5.72
8	2	1.72
≥ 9	0	0.60
Total	696	696.01

Their is quite good agreement between the observed and expected frequencies which indicates the Poisson model is very reasonable. Recall the homogeneity assumption for the Poisson process. Since a Poisson model fits the data well this suggests that Wayne was a very consistent player when he played with the Edmonton Oilers.

2.10 (a) Since $P(Y = 1; \theta) = \theta$, the parameter θ represents the probability a randomly chosen family has one child.

(b) Let $F_y =$ the number of families with y children. The probability of observing the data

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

is

$$\frac{n!}{f_0!f_1!\cdots f_{\max}!0!}\left(\frac{1-2\theta}{1-\theta}\right)^{f_0}(\theta)^{f_2}\left(\theta^2\right)^{f_2}\cdots\left(\theta^{y_{\max}}\right)^{f_{\max}}\left(\sum_{y=y_{\max}+1}^{\infty}\theta^y\right)^0$$

$$\frac{n!}{f_0!f_1!\cdots f_{\max}!}\left(\frac{1-2\theta}{1-\theta}\right)^{f_0}\prod_{y=1}^{y_{\max}}\theta^{yf_y}.$$

If we ignore constants with respect to θ, the likelihood function is

$$\left(\frac{1-2\theta}{1-\theta}\right)^{f_0}\prod_{y=1}^{y_{\max}}\theta^{yf_y}$$

and the log likelihood is

$$l(\theta) = f_0 \log\left(\frac{1-2\theta}{1-\theta}\right) + \left(\sum_{y=1}^{y_{\max}} y f_y\right) \log\theta \quad \text{for } 0 < \theta \leq 0.5$$

$$= f_0 \log(1-2\theta) - f_0 \log(1-\theta) + T\log\theta \quad \text{where } T = \sum_{y=1}^{y_{\max}} y f_y.$$

Now

$$\begin{aligned} l'(\theta) &= \frac{-2f_0}{1-2\theta} + \frac{f_0}{1-\theta} + \frac{1}{\theta}T \\ &= \frac{1}{\theta(1-\theta)(1-2\theta)}\left[2T\theta^2 - (f_0+3T)\theta + T\right] \end{aligned}$$

and $l'(\theta) = 0$ if

$$\theta = \frac{(f_0+3T) \pm [(f_0+3T)^2 - 8T^2]^{1/2}}{4T}$$

and since

$$\frac{(f_0+3T) + [(f_0+3T)^2 - 8T^2]^{1/2}}{4T} \geq \frac{f_0+3T}{4T} \geq \frac{3}{4} > 0.5$$

therefore

$$\hat{\theta} = \frac{(f_0+3T) - [(f_0+3T)^2 - 8T^2]^{1/2}}{4T}$$

(c) The probability that a randomly selected family has x children is θ^x, $x = 1, 2, \ldots$. Suppose for simplicity there are N different families where N is very large. Then the number of families that have x children is $N \times$ (probability a family has x children) $= N\theta^x$ for $x = 1, 2, \ldots$ and there is a total of $xN\theta^x$ children in families of x children and a total of $\sum_{x=1}^{\infty} xN\theta^x$ children altogether. Therefore the probability a randomly chosen child is in a family of x children is:

$$\frac{xN\theta^x}{\sum_{x=1}^{\infty} xN\theta^x} = cx\theta^x, \qquad x = 1, 2, \ldots$$

Note that $\sum_{x=0}^{\infty} \theta^x = \frac{1}{1-\theta}$. Therefore taking derivatives

$$\sum_{x=1}^{\infty} x\theta^{x-1} = \frac{1}{(1-\theta)^2} \text{ and } \sum_{x=1}^{\infty} x\theta^x = \frac{\theta}{(1-\theta)^2}$$

Solving

$$\sum_{x=1}^{\infty} cx\theta^x = 1$$

gives $c = (1-\theta)^2/\theta$ and

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

(d) The probability of observing the given data for model (c) is

$$\frac{33!}{22!7!3!1!}\left[(1-\theta)^2\right]^{22}\left[2(1-\theta)^2\theta\right]^{7}\left[3(1-\theta)^2\theta^2\right]^{3}\left[4(1-\theta)^2\theta^3\right] \quad \text{for } 0<\theta\leq\frac{1}{2}.$$

The likelihood function is

$$\begin{aligned} L(\theta) &= (1-\theta)^{2(22+7+3+1)}\theta^{7+2(3)+3} \\ &= \theta^{16}(1-\theta)^{66} \quad \text{for } 0<\theta\leq\frac{1}{2} \end{aligned}$$

which is maximized for $\theta = 16/(16+66) = 16/82 = 8/41 = 0.1951$.

Since the probability a family has no children is

$$P(Y=0;\theta) = \frac{1-2\theta}{1-\theta} = g(\theta)$$

then by the Invariance Property of maximum likelihood estimates the maximum likelihood of $g(\theta)$ is

$$g(\hat{\theta}) = \frac{1-2\hat{\theta}}{1-\hat{\theta}} = \frac{1-2(0.1951)}{1-0.1951} = 0.7576$$

(e) For these data $f_0 = 0$, $T = 49.$ and $l'(\theta) = 49/\theta = 0$ has no solution. Since $l'(\theta) = 49/\theta > 0$ for all $0 < \theta \leq 0.5$, therefore $l(\theta)$ is an increasing function on this interval. Thus the maximum value of $l(\theta)$ occurs at the endpoint $\theta = 0.5$ and therefore $\hat{\theta} = 0.5$.

2.11 (a) Let Y_i = the number of particles emitted in time interval i of length t_i, $i = 1, 2, \ldots, n$. We assume that the Y_i's are independent random variables. The likelihood function is the probability of observing the data $y_1, y_2, \ldots, y_n$ which is

$$\begin{aligned} L(\theta) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n; \theta) \\ &= \prod_{i=1}^{n} P(Y_i = y_i; \theta) = \prod_{i=1}^{n} \frac{(\theta t_i)^{y_i} e^{-\theta t_i}}{y_i!} \\ &= \prod_{i=1}^{n} \frac{(t_i)^{y_i}}{y_i!} \prod_{i=1}^{n} \theta^{-y_i} e^{-\theta t_i} \end{aligned}$$

or more simply

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i} e^{-\theta t_i} = \theta^{\sum_{i=1}^{n} y_i} e^{-\theta \sum_{i=1}^{n} t_i}$$

The log likelihood function is

$$l(\theta) = \left(\sum_{i=1}^{n} y_i\right)\log\theta - \theta\sum_{i=1}^{n} t_i$$

and

$$l'(\theta) = \frac{\sum_{i=1}^{n} y_i}{\theta} - \sum_{i=1}^{n} t_i = \frac{1}{\theta}\left(\sum_{i=1}^{n} y_i - \theta \sum_{i=1}^{n} t_i\right) = 0$$

if

$$\theta = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} t_i}$$

so $\hat{\theta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} t_i$ is the maximum likelihood estimate of θ.

(b) Let X = number of intervals of length t with no particles emitted. Then $X \backsim$ Binomial(n, p) where

$$p = P(X = 0; \theta) = \frac{(\theta t)^0 e^{-\theta t}}{0!} = e^{-\theta t}$$

Suppose that x intervals were observed with no particles. Since $X \backsim$ Binomial(n, p) the maximum likelihood estimate of p is $\hat{p} = x/n$. Since $p = e^{-\theta t}$ implies $\theta = -(\log p)/t$ then by the Invariance Property of maximum likelihood estimates $\hat{\theta} = -(\log \hat{p})/t$.

2.13 (a) The five number summary is: 3 16.25 19.5 22 30

(b) $\bar{y} = 19.14$ and $s = 4.47$

(c) The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s] = [14.68, 23.61]$ is $71/100 = 0.71$. If $Y \backsim G(\mu, \sigma)$ then

$$\begin{aligned} P(Y \in [\mu - \sigma, \mu + \sigma]) &= P(|Y - \mu| \leq \sigma) = P\left(\frac{|Y - \mu|}{\sigma} \leq 1\right) \\ &= P(|Z| \leq 1) = 2P(Z \leq 1) - 1 \quad \text{where} \quad Z \backsim N(0, 1) \\ &= 2(0.84134) - 1 = 0.68268 \\ &\approx 0.68 \end{aligned}$$

The proportion of observations in the interval (0.71) is slightly higher than what would be expected for Normal data (0.68).

(d)

$$IQR = 22 - 16.25 = 5.75.$$

To show that for Normally distributed data that IQR $= 1.349\sigma$ we need to solve

$$0.5 = P(|Y - \mu| \leq c\sigma) = P\left(\frac{|Y - \mu|}{\sigma} \leq c\right) \quad \text{for } c \text{ if } Y \backsim G(\mu, \sigma).$$

From $N(0,1)$ tables $P(|Z| \leq c) = 2P(Z \leq c) - 1 = 0.5$ holds if $c = 0.6745$. Therefore

$$IQR = 2(0.6745)\sigma = 1.349\sigma$$

for Normally distributed data.

(e) For Normally distributed data we expect the skewness to be close to zero and the sample mean and median to be approximately equal. For these data the sample skewness $= -0.50$ and the sample median $= 19.5 >$ mean $= 19.14$. Both of these results indicate that the data are not symmetric but **slightly** skewed to the left. This is also evident in the boxplot in which neither the box nor the whiskers are divided approximately in half by the median.
For Normally distributed data we expect the sample kurtosis to be close to 3. The sample kurtosis for these data equals 4.30 which indicates that there are more observations in the tails then would be expected for Normally distributed data.
In the list of observations as well as the boxplot we observe two extreme observations, 3 and 5, which are also evident in the qqplot (see lower left hand corner of graph). These extremes have a large influence on the sample mean as well as on the sample skewness and sample kurtosis. If the sample mean, median, skewness and kurtosis were recalculated with these observations removed then the values of these numerical summaries would be more in agreement with what we expect to see for Normally distributed data.
Except for the outliers, the points in the qqplot lie quite well along a straight line. For Normally distributed data we expect the points to lie reasonably along a straight line although the points at both ends may lie further from the straight line since the quantiles of the Normal distribution change more rapidly in both tails of the distribution.
The proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$ is slightly higher than we would expect for Normally distributed data. This also agrees with the sample kurtosis value of 4.3 being larger than 3.
Overall, except for the two outliers, it seems reasonable to to assume that the data are approximately Normally distributed. It would be a good idea to do any formal analyses of the data with and without the outliers to determine the effect of these outliers on the conclusions of these analyses.

2.14 (a) The sample mean $\bar{y} = 159.77$ and the sample standard deviation $s = 6.03$ for the data.

(b) The number of observations in the interval $[\bar{y} - s, \bar{y} + s] = [153.75, 165.80]$ is 244 or 69.5% and actual number of observations in the interval $[\bar{y} - 2s, \bar{y} + 2s] =$

$[147.72, 171.83]$ is 334 or 95.2%. If $Y \backsim G(\mu, \sigma)$ then

$$\begin{aligned} P(Y \in [\mu - \sigma, \mu + \sigma]) &= P(|Y - \mu| \leq \sigma) = P\left(\frac{|Y - \mu|}{\sigma} \leq 1\right) \\ &= P(|Z| \leq 1) = 2P(Z \leq 1) - 1 \text{ where } Z \backsim N(0,1) \\ &= 2(0.84134) - 1 = 0.68268. \end{aligned}$$

Similarly $P(Y \in [\mu - 2\sigma, \mu + 2\sigma]) = 0.9545$. The observed and expected proportions and very close to what one would expect if the data were Normally distributed.

(c) The sample skewness for these data is 0.13 while for Normally distributed data we expect a sample skewness close to 0. The sample kurtosis for these data is 3.16 while for Normally distributed data we expect a sample kurtosis close to 3. Both the sample skewness and the sample kurtosis are reasonably close to what we expect for Normally distributed data.

(d) The five-number summary for the data is given by $y_{(1)}$, $q(0.25)$, $q(0.5)$, $q(0.75)$, $y_{(n)} = 142$, 156, 160, 164, 178

(e) $IQR = q(0.75) - q(0.25) = 164 - 156 = 8$. For Normally distributed data we expect $IQR = 1.349\sigma$. For these data $IQR = 1.33s$ so this relationship is almost exact.

(f) The frequency histogram and superimposed Gaussian probability density function are given in the top left graph in Figure 12.9.

(g) The empirical cumulative distribution function and superimposed Gaussian cumulative distribution function are given in the top right graph in Figure 12.9.

(h) The boxplot is given in the bottom left graph in Figure 12.9.

(i) The qqplot is given in the bottom right graph in Figure 12.9. The "steplike" behaviour of the plot is due to the rounding of the data to the nearest centimeter.

(j) All the numerical summaries indicate good agreement with the Gaussian assumption. The relative frequency histogram has the shape of a Gaussian probability density function. The empirical cumulative distribution function and the Gaussian cumulative distribution function also have similar shapes. The boxplot is consistent with Gaussian data and the points in the qqplot lie reasonably along a straight line also indicating good agreement with a Gaussian model. A Gaussian distribution seems very reasonable for these data.

2.15 (a) $\hat{\mu} = 1.744$, $\hat{\sigma} = 0.0664$ (M) $\hat{\mu} = 1.618$, $\hat{\sigma} = 0.0636$ (F)

(b) 1.659 and 1.829 (M) 1.536 and 1.670 (F)

(c) 0.098 (M) and 0.0004 (F)

(d) $11/50 = 0.073$ (M) 0 (F)

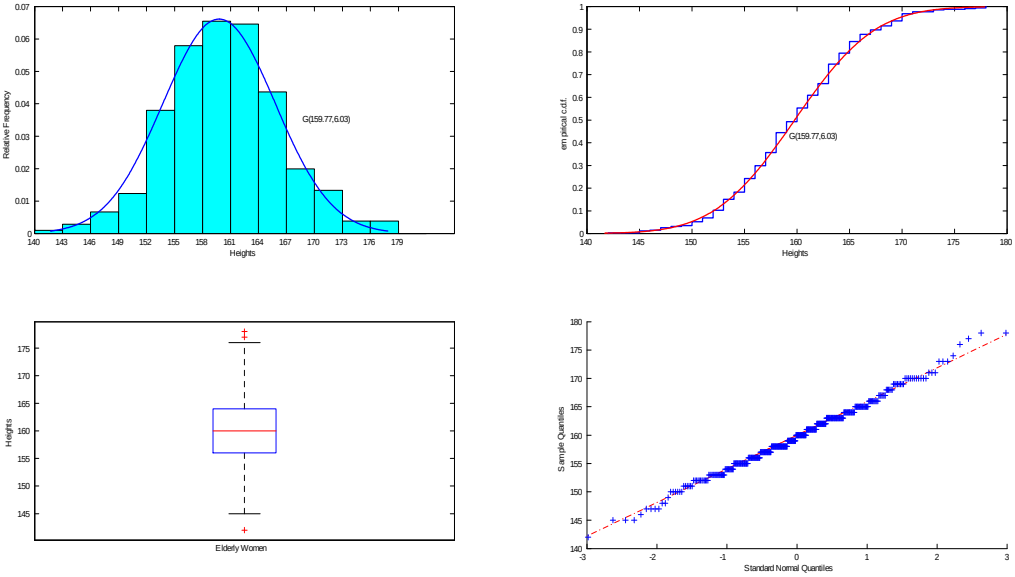


Figure 12.9: Plots for Heights of Elderly Women

2.16 See Figure 12.10. Note that the qqplot for the $\log y_i$'s is far more linear than for the y_i's indicating that the Normal model is more reasonable for the transformed data.

2.17 (a) If they are independent $P(S \text{ and } H) = P(S)P(H) = \alpha\beta$. The others are similar.

(b) The Multinomial probability function evaluated at the observed values is

$$L(\alpha, \beta) = \frac{100!}{20!15!22!43!} (\alpha\beta)^{20} [\alpha (1-\beta)]^{15} [(1-\alpha)\beta]^{22} [(1-\alpha)(1-\beta)]^{43}$$

and the log likelihood (ignoring constants) is

$$l(\alpha, \beta) = 35 \log(\alpha) + 65 \log(1-\alpha) + 42 \log(\beta) + 58 \log(1-\beta).$$

Setting the derivatives to zero gives the maximum likelihood estimates,

(c) The expected frequencies are

$$100\hat{\alpha}\hat{\beta},\ 100\hat{\alpha}\left(1-\hat{\beta}\right),\ 100\left(1-\hat{\alpha}\right)\hat{\beta},\ 100\left(1-\hat{\alpha}\right)\left(1-\hat{\beta}\right) \quad \text{respectively}$$

or

$$\left(\frac{35(42)}{100}, \frac{35(58)}{100}, \frac{65(42)}{100}, \frac{65(58)}{100}\right) = (14.7, 20.3, 27.3, 37.7)$$

which can be compared with $20, 15, 22, 43$. The observed and expected frequencies do not appear to be very close. In Chapter 7 we will see how to construct a formal test of the model.

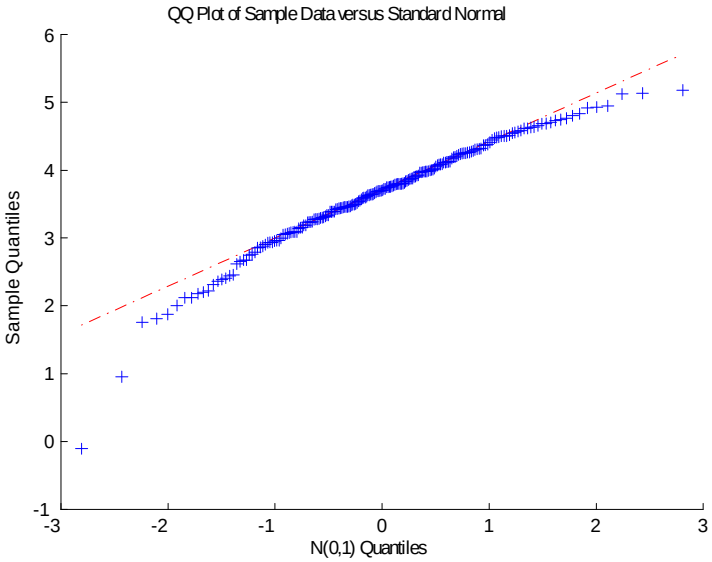


Figure 12.10: Qqplot of log brake pad lifetimes

2.19 (a) The median of the $N(0,1)$ distribution is $m = 0$. Reading from the qqplot the sample quantile on the $y-$axis which corresponds to 0 on the $x-$axis is approximately equal to 1.0 so the sample median is approximately 1.0.

(b) To determine $q(0.25)$ for the these data we note that $P(Z \leq -0.6745) = 0.25$ if $Z \backsim N(0,1)$. Reading from the qqplot the sample quantile on the $y-$axis which corresponds to -0.67 on the $x-$axis is approximately equal to 0.4 so $q(0.25)$ is approximately 0.4. To determine $q(0.75)$ for the these data we note that $P(Z \leq -0.6745) = 0.25$ if $Z \backsim N(0,1)$. Reading from the qqplot the sample quantile on the $y-$axis which corresponds to 0.67 on the $x-$axis is approximately equal to 1.5 so $q(0.75)$ is approximately 1.5. The IQR for these data is approximately $1.5 - 0.4 = 1.1$.

(c) The frequency histogram of the data would be approximately symmetric about the sample mean.

(d) The frequency histogram would most resemble a Uniform probability density function.

2.20 (a) If there is adequate mixing of the tagged animals, the number of tagged animals caught in the second round is a random sample selected without replacement so follows a hypergeometric distribution (see the STAT 230 Course Notes).

(b)
$$\frac{L(N+1)}{L(N)} = \frac{(N+1-k)(N+1-n)}{(N+1-k-n+y)(N+1)}$$
and $L(N)$ reaches its maximum within an integer of kn/y.

(c) The model requires sufficient mixing between captures that the second stage is a random sample. If they are herd animals this model will not fit well.

2.21

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i;\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \quad \text{if } \theta \geq y_i \;\; i = 1, 2, \ldots, n \\ &= \frac{1}{\theta^n} \quad \text{if } \theta \geq y_{(n)} = \max(y_1, y_2, \ldots, y_n) \end{aligned}$$

where θ^{-n} is a decreasing function of θ. Note also that $L(\theta) = 0$ for $0 < \theta < y_{(n)}$. Therefore the maximum value of $L(\theta)$ occurs at $\theta = y_{(n)}$ and therefore the maximum likelihood estimate of θ is $\hat{\theta} = y_{(n)}$.

2.22 (a)

$$P(Y > c;\theta) = \int_c^{\infty} \frac{1}{\theta} e^{-y/\theta} dy = e^{-c/\theta}.$$

(b) For the i'th piece that failed at time $y_i < c$, the contribution to the likelihood is $\frac{1}{\theta} e^{-y_i/\theta}$. For those pieces that survive past time c, the contribution to the likelihood is the probability of the event, $P(Y > c;\theta) = e^{-c/\theta}$. Therefore the likelihood is

$$L(\theta) = \left(\prod_{i=1}^{k} \frac{1}{\theta} e^{-y_i/\theta} \right) \left(e^{-c/\theta} \right)^{n-k}$$

$$l(\theta) = -k \log(\theta) - \frac{1}{\theta} \sum_{i=1}^{k} y_i - (n-k)\frac{c}{\theta}$$

and solving $l'(\theta) = 0$ we obtain the maximum likelihood estimate,

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

(c) When $k = 0$ and $c > 0$ the maximum likelihood estimator is $\hat{\theta} = \infty$. In this case there are no failures in the time interval $[0, c]$ and this is more likely to happen when $E(Y) = \theta$ is very large.

2.23 The likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{[\theta(x_i)]^{y_i}}{y_i!} e^{-\theta(x_i)}$$

and, ignoring the terms $y_i!$ which do not contain the parameters, the log likelihood is

$$l(\alpha, \beta) = \sum_{i=1}^{n} \left[y_i(\alpha + \beta x_i) - e^{(\alpha + \beta x_i)} \right]$$

To maximize we set the derivatives equal to zero and solve

$$\frac{\partial}{\partial \alpha} l(\alpha, \beta) = \sum_{i=1}^{n} \left[y_i - e^{(\alpha + \beta x_i)} \right] = 0$$
$$\frac{\partial}{\partial \beta} l(\alpha, \beta) = \sum_{i=1}^{n} x_i \left[y_i - e^{(\alpha + \beta x_i)} \right] = 0$$

For a given set of data we can solve this system of equations numerically but not explicitly.

Chapter 3

3.1 (a) The Problem is to determine the proportion of eligible voters who plan to vote and, of those, the proportion who plan to support the party. This is a descriptive Problem since the aim of the study is to determine the attributes just mentioned for a population of eligible voters.

(b) The target population is all eligible voters. This would include those eligible voters in all regions and those with/without telephone numbers on the list.

(c) One variate is whether or not an eligible voter plans to vote which is a categorical variate. Another variate is whether or not an eligible voter supports the party which is also a categorical variate.

(d) There are two attributes of interest in the target population. One attribute is the proportion of the target population who plan to vote. Another attribute is the proportion of the target population who plan to vote who also plan to support the party.

(e) The study population is all eligible voters on the list.

(f) The sample is the 1104 eligible voters who responded to the questions.

(g) A possible source of study error is that the polling firmed only called eligible voters in urban areas. Urban eligible voters may have different views that rural eligible voters – this is a difference between the target and study populations. Eligible voters with phones may have different views than those without.

(h) A possible source of sample error is that many of the people called refused to participate in the survey. People who refuse to participate may have different voting preferences as compared to people who participate. For example, people who refuse to participate in the survey may also be less likely to vote.

(i) An estimate of the proportion of the study population who plan to vote based on the data is 732/1104.
An estimate of the proportion of the study population who plan to vote who also plan to support the party based on the data is 351/732.

3.2 (a) This study is an experimental study since the researchers are in control of which schools received the regular curriculum and which schools are using the JUMP program.

(b) The Problem is to compare the performance in math of students at Ontario schools using the current provincial curriculum as compared to the performance in math of students at Ontario schools using the JUMP math program.

(c) This is a causative problem since the researcher are interested in whether the JUMP program causes better student performance in math.

(d) The target population is all elementary students in Ontario public schools at the time the study or all elementary students in Ontario public schools at the time the study and into the future.

(e) One variate of interest is whether a student the student receives the standard curriculum or the standard curriculum plus the JUMP program which is a categorical study. Another variate of interest is classroom test scores which is a discrete variate since scores only take on a finite number of countable values.

(f) The study population is all Ontario elementary students in Grades 2 and 5 in public schools at the time the study.

(g) The sampling protocol was to select the schools in one school board in Ontario. The researchers did not indicate how this school board was chosen.

(h) A possible source of study error is that the ability of students in Grades 2 and 5 to learn math skills might be different than students in other grades.

(i) A possible source of sample error is that the schools in the chosen school board may not be representative of all the elementary schools in Ontario. For example, the schools in the chosen board may have larger class sizes compared to other schools. Student in larger classes may not receive as much help to improve their math skills as students in smaller classes. Another example is that the chosen school board might be in a low income area of a city. Students from low income families may respond differently to changes in the Math curriculum as compared to students from middle class families.

(j) It is unclear from the article what type of classroom tests will be used or how they will be graded. So depending on how this is done it could lead to measurement error. For example, different schools may use different grading criteria for the same test.

(k) Randomization ensures that the difference in the learning outcome is only due to different teaching programs, and not due to other potential confounders (e.g., class size, parents' education level, parents' social economic status, etc.).

3.3 (a) This is an experimental study because the researchers controlled, using randomization, which students were assigned to the racing-type game and which students were assigned to the game of solitaire.

(b) The Problem is to determine whether playing racing games makes players more likely to take risks in a simulated driving test.

(c) This is a causative type Problem because the researchers were interested in whether playing racing games as compared to playing a game like solitaire caused players to take more risks in the driving test.

(d) A suitable target population for this study is young adults living in China at the time of the study OR students attending university in China at the time of the study.

(e) One important variate is whether the student played the racing-type driving game or the game of solitaire. This is a categorical variate. The other important variate was how long, in seconds, the student waited to hit the "stop" key in the Vienna Risk-Taking Test. This is a continuous variate.

(f) The variate of interest in the target population is the mean difference in the time to hit the "stop" key in the Vienna Risk-Taking Test between young adults who play racing games compared to young adults who play neutral games.

(g) A suitable study population for this study is students attending Xi'an Jiatong University at the time of the study.

(h) From the article it appears that the researchers recruited volunteers for the study. The article does not indicate how these volunteers were obtained.

(i) If the target population is young adults living in China and the study population is students attending university in China at the time of the study then a possible source of study error is that students who attend university are more educated and more intelligent (on average) and therefore possibly different in their levels of risk-taking as compared to young adults in China not attending university.

(j) Since the sample consisted of volunteers and not a random sample of students from the Xi'an Jiatong University then a possible source of study error is that students who volunteer for such studies are more likely to take risks than non-volunteers who might be more conservative. The risk-taking habits of the volunteers (on average) may be different than the risk-taking habits of all students at the Xi'an Jiatong University.

(k) An estimate based on the data of the mean difference in the time to hit the "stop" key in the Vienna Risk-Taking Test between young adults who play racing games compared to young adults who play neutral games in the study population is $12 - 10 = 2$ seconds.

Chapter 4

4.1 (a) If $n = 1000$ and $\theta = 0.5$ then

$$\begin{aligned} P\left(-0.03 \le \tilde{\theta} - \theta \le 0.03\right) &= P\left(-0.03 \le \frac{Y}{1000} - 0.5 \le 0.03\right) \\ &= P\left(\frac{-0.03}{\sqrt{\frac{(0.5)(0.5)}{1000}}} \le \frac{\frac{Y}{1000} - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}} \le \frac{0.03}{\sqrt{\frac{(0.5)(0.5)}{1000}}}\right) \\ &\approx P\left(-1.90 \le Z \le 1.90\right) \quad \text{where } Z \sim N(0,1) \\ &= 2P\left(Z \le 1.90\right) - 1 = 2\left(0.97128\right) - 1 \\ &= 0.94256 \end{aligned}$$

(b)

$$\begin{aligned} P\left(-0.03 \le \tilde{\theta} - \theta \le 0.03\right) &= P\left(-0.03 \le \frac{Y}{n} - 0.5 \le 0.03\right) \\ &= P\left(\frac{-0.03}{\sqrt{\frac{(0.5)(0.5)}{n}}} \le \frac{\frac{Y}{n} - 0.5}{\sqrt{\frac{(0.5)(0.5)}{n}}} \le \frac{0.03}{\sqrt{\frac{(0.5)(0.5)}{n}}}\right) \\ &\approx P\left(-0.06\sqrt{n} \le Z \le 0.06\sqrt{n}\right) \end{aligned}$$

where $Z \sim N(0,1)$. Since $P\left(-1.96 \le Z \le 1.96\right) = 0.95$, we need $0.06\sqrt{n} \ge 1.96$ or $n \ge \left(1.96/0.06\right)^2 = 1067.1$. Therefore n should be at least 1068.

(c)

$$\begin{aligned} P\left(-0.03 \le \tilde{\theta} - \theta \le 0.03\right) &= P\left(-0.03 \le \frac{Y}{n} - \theta \le 0.03\right) \\ &= P\left(\frac{-0.03}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le \frac{\frac{Y}{n} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \le \frac{0.03}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right) \\ &\approx P\left(-\frac{0.03\sqrt{n}}{\sqrt{\theta\left(1-\theta\right)}} \le Z \le \frac{0.03\sqrt{n}}{\sqrt{\theta\left(1-\theta\right)}}\right) \end{aligned}$$

where $Z \sim N(0,1)$. Since $P\left(-1.96 \le Z \le 1.96\right) = 0.95$, we need

$$\frac{0.03\sqrt{n}}{\sqrt{\theta\left(1-\theta\right)}} \ge 1.96$$

or

$$n \ge \left(\frac{1.96}{0.03}\right)^2 \theta\left(1-\theta\right).$$

Since θ is unknown we take $\theta = 0.5$ so the inequality is true for all $0 < \theta < 1$. Thus

$$n \ge \left(\frac{1.96}{0.03}\right)^2 (0.5)^2 = 1067.1$$

and n should be at least 1068.

4.2 (b) For samples of size $n = 30$ the histogram of simulated means should be centred very close to $\mu = 2.326$, the variability of the sample means should be smaller compared to the variability for samples of size $n = 15$ since $sd\left(\bar{Y}\right) \approx \sigma/\sqrt{n}$ and the histogram should be more symmetric. The estimate of $P\left(\left|\bar{Y} - 2.326\right| \leq 0.5\right)$ should increase since $P\left(\left|\bar{Y} - 2.326\right| \leq 0.5\right)$ increases as n increases.

(c) Since $E\left(\bar{Y}\right) = \mu$, the mean of the original population will affect the location of the center of the histogram of simulated means.
Since $sd\left(\bar{Y}\right) \approx \sigma/\sqrt{n}$, the standard deviation of the original population will affect the spread of the histogram. Larger values of σ will result in histograms with more spread.
From the Central Limit Theorem we know that if the original population is very symmetric then the distribution of $\bar{Y}$ approaches a Normal distribution more rapidly as n increases as compared to the case in which the original distribution is very non-symmetric. Therefore if the original distribution is reasonably symmetric then the histogram will be very symmetric even if the sample size n is not large. If the original distribution is not symmetric then you would not expect the histogram to be reasonably symmetric unless n is large.

(d) Since $sd\left(\bar{Y}\right) \approx \sigma/\sqrt{n}$, the spread of the histogram will be affected by the sample size n. Larger sample sizes will result in histograms which are more concentrated around the mean μ which intuitively makes sense.

4.6 (a) Suppose the experiment which was used to estimate μ was conducted a large number of times and each time a 95% confidence interval for μ was constructed using the observed data. Then, approximately 95% of these constructed intervals would contain the true, but unknown value of μ. Since we only have one interval $[42.8, 47.8]$ we do not know whether it contains the true value of μ or not. We can only say that we are 95% confident that the given interval $[42.8, 47.8]$ contains the true value of μ since we are told it is a 95% confidence interval. In other words, we hope we were one of the "lucky" 95% who constructed an interval containing the true value of μ. **Warning:** You cannot say that the probability that the interval $[42.8, 47.8]$ contains the true value of is 0.95!!!

(b) An approximate 95% confidence interval for the proportion of Canadians whose mobile phone is a smartphone is

$$\begin{aligned} \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} &= 0.45 \pm 1.96\sqrt{\frac{0.45(0.55)}{1000}} = 0.45 \pm 0.03083 \\ &= [0.4192, 0.4808]. \end{aligned}$$

(c) We need n such that

$$n \geq \left(\frac{1.96}{0.02}\right)^2 (0.5)^2 = 2401.$$

A sample size of 2401 or larger should be used.

4.7 Let Y = number of women who tested positive. Assume that model $Y \backsim \text{Binomial}(n, \theta)$. Since $P(-2.58 \leq Z \leq 2.58) = 0.99$, an approximate 99% confidence interval is given by:

$$\begin{aligned} \hat{\theta} \pm 2.58\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} &= \frac{64}{29000} \pm 2.58\sqrt{\frac{\frac{64}{29000}(\frac{28936}{29000})}{29000}} = 0.0022 \pm 0.0007 \\ &= [0.0015, 0.0029]. \end{aligned}$$

The Binomial model assumes that the 29,000 women represented 29,000 independent trials and that the probability that a randomly chosen women is HIV positive is equal to θ. The women may not represent independent trials and the probability that a randomly chosen women is HIV positive may be higher among certain high risk women such as women who are intravenous drug users.

4.8 (a) If Y is the number who support this information then $Y \backsim \text{Binomial}(n, \theta)$. An approximate 95% confidence interval is given by

$$\begin{aligned} 0.7 \pm 1.96\sqrt{\frac{0.7(0.3)}{200}} &= 0.7 \pm 0.6351 \\ &= [0.6365,\ 0.7635]. \end{aligned}$$

(b) The Binomial model assumes that the 200 people represent 200 independent trials. If 100 of the people interviewed were 50 married couples then the two people in a couple are probably not independent with respect to their views.

4.9 From Problem 2.3 we have

$$R(\theta) = \left[\left(\frac{\theta}{0.5}\right)\left(\frac{1-\theta}{0.5}\right)\right]^{200} = [4\theta(1-\theta)]^{200} \quad \text{for} \quad 0 < \theta < 1.$$

The 15% likelihood interval can be obtained from the graph of $R(\theta)$ given in Figure 12.6 or by using the R command:
`uniroot(function(x)((4*x*(1-x))^200-0.15),lower=0.4,upper=0.5)`.
The 15% likelihood interval is $[0.45, 0.55]$.

4.10 From Problem 2.5 we have

$$R(\theta) = \left[\frac{3.6}{\theta}e^{(1-3.6/\theta)}\right]^{20} \qquad \theta > 0.$$

The 15% likelihood interval can be obtained from the graph of $R(\theta)$ given in Figure 12.7 or by using the R command:
`uniroot(function(x)(((3.6/x)*exp(1-3.6/x))^20-0.15),lower=2.0,upper=3.0)`.
The 15% likelihood interval is $[2.40, 5.76]$.

4.11 From Problem 2.7 we have

$$r(\theta) = 15\log[2.3(\theta+1)] - 34.5(\theta+1) + 15 \quad \text{for } \theta > -1.$$

The 15% likelihood interval can be obtained from the graph of $r(\theta)$ given in Figure 12.8 or by using the R command:
`uniroot(function(x)(15*log(2.3*(x+1))-34.5*(x+1)+15-log(0.15)), lower=-0.8,upper=-0.7)`.
The 15% likelihood interval is $[-0.75, -0.31]$.

4.12 (a) For the data $n_1 = 16$, $n_2 = 16$ and $n_3 = 18$, $\hat{\alpha} = 0.28$ and

$$R(\alpha) = \frac{(1+\alpha)^{32}(1-\alpha)^{18}}{(1+0.28)^{32}(1-0.28)^{18}}, \quad 0 < \alpha < 1$$

Looking at Figure 12.11 we can see that $R(\alpha) = 0.1$ corresponds to α between 0.5 to 0.6. We use the following command in R:
`uniroot(function(x)(((1+x)/1.28)^32*((1-x)/0.72)^18-0.1), lower=0.5,upper=0.6)` to obtain the answer 0.55. Therefore the 10% likelihood interval is $[0, 0.55]$. Since the 10% likelihood interval is very wide this indicates that α is not very accurately determined.

(b) For the data for which 17 identical pairs were found, $\hat{\alpha} = 17/50 = 0.34$ and the relative likelihood function is

$$R(\alpha) = \frac{\alpha^{17}(1-\alpha)^{33}}{(0.34)^{17}(1-0.34)^{33}}, \quad 0 < \alpha < 1$$

We use
`uniroot(function(x)((x/0.34)^17*((1-x)/0.66)^33-0.1),lower=0,upper=0.3)`
to obtain the 10% likelihood interval $[0.21, 0.49]$. This interval is much narrower than the interval in (a) which indicates that α is more accurately determined by the second model.

4.13 From Problem 2.10 we have

$$L(\theta) = \theta^{16}(1-\theta)^{66} \quad \text{for } 0 < \theta \leq \frac{1}{2},$$

$\hat{\theta} = 16/82 = 8/41$ and

$$R(\theta) = \frac{\theta^{16}(1-\theta)^{66}}{(8/41)^{16}(33/41)^{66}} \quad \text{for } 0 < \theta \leq \frac{1}{2}.$$

A graph of $R(\theta)$ is given in Figure 12.12.The 15% likelihood interval can be obtained from the graph of $R(\theta)$ given in Figure 12.12 or by using the R command:
`uniroot(function(x)(((41*x)/8)^16*(41*(1-x)/33)^66-0.15), lower=0.1,upper=0.15)`. The 15% likelihood interval is $[0.12, 0.29]$.

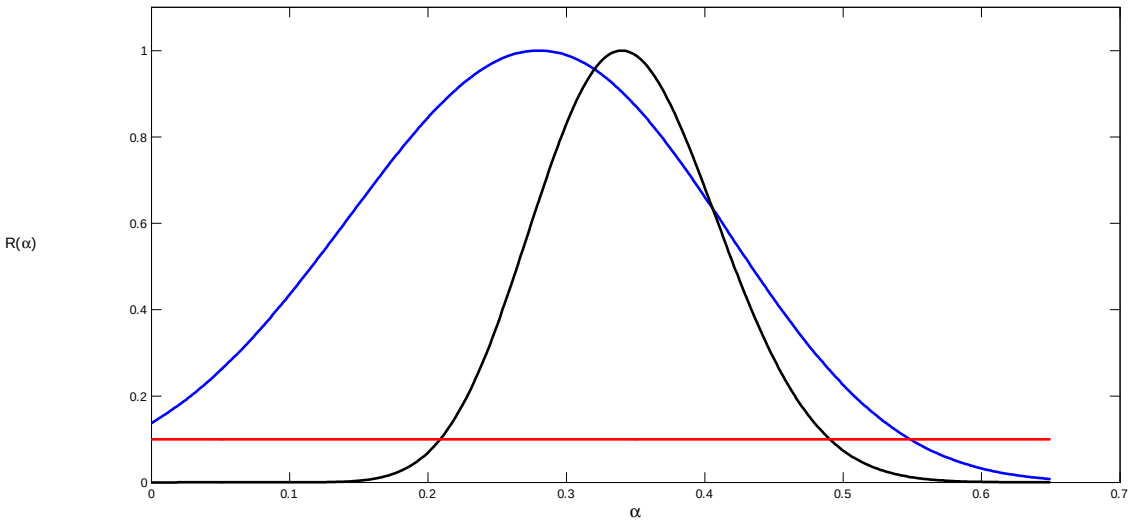


Figure 12.11: Relative likelihood functions for Twin Data

4.14 (a) The probability a group tests negative is $p = (1-\theta)^k$. The probability that y out of n groups test negative is

$$\binom{n}{y} p^y (1-p)^{n-y} \qquad y = 0, 1, ..., n.$$

We are assuming that the nk people represent independent trials and that θ does not vary across subpopulations of the population of interest.

(b) Since $L(p) = p^y(1-p)^{n-y}$ is the usual Binomial likelihood we know $\hat{p} = y/n$. Solving $p = (1-\theta)^k$ for θ we obtain $\theta = 1 - p^{1/k}$. Therefore by the Invariance Property of maximum likelihood estimates, the maximum likelihood estimate of θ is

$$\hat{\theta} = 1 - (\hat{p})^{1/k} = 1 - (y/n)^{1/k}.$$

(c) For $n = 100$, $k = 10$ and $y = 89$ we have $\hat{p} = 89/100 = 0.89$ and $\hat{\theta} = 1 - (89/100)^{1/10} = 0.0116$. A 10% likelihood interval for p is found by using: `uniroot(function(x)((x/0.89)^89*((1-x)/0.11)^11-0.1),lower=0.5,upper=0.9)` which gives the interval $[0.8113, 0.9451]$ for p. The 10% likelihood interval for θ is

$$\left[1 - (0.9451)^{1/10}, 1 - (0.8113)^{1/10}\right] = [0.0056,\ 0.0207]\,.$$

4.15

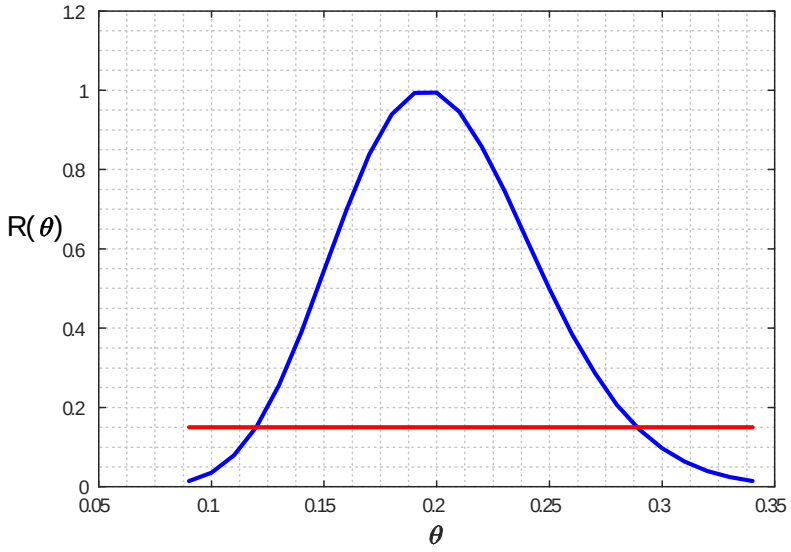


Figure 12.12: Relative likelihood function for size of family data

(a) Since

$$\begin{aligned} E\left(Y^k\right) &= \int_0^\infty y^k \frac{y}{\theta^2} e^{-y/\theta} dy = \int_0^\infty \frac{y^{k+1}}{\theta^2} e^{-y/\theta} dy \quad \text{let } x = y/\theta \\ &= \frac{1}{\theta^2} \int_0^\infty (x\theta)^{k+1} e^{-x} \theta dx = \theta^k \int_0^\infty x^{k+1} e^{-x} dx = \theta^k \Gamma(k+2) \end{aligned}$$

therefore

$$\begin{aligned} E(Y) &= \theta\Gamma(3) = 2\theta, \quad E\left(Y^2\right) = \theta^2\Gamma(4) = 6\theta^2 \\ Var(Y) &= E\left(Y^2\right) - [E(Y)]^2 = 6\theta^2 - (2\theta)^2 = 2\theta^2 \end{aligned}$$

as required.

(b) The likelihood function is

$$L(\theta) = \prod_{i=1}^n \frac{y_i}{\theta^2} e^{-y_i/\theta} = \left(\prod_{i=1}^n y_i\right) \theta^{-2n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^n y_i\right), \quad \theta > 0$$

or more simply

$$L(\theta) = \theta^{-2n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^n y_i\right), \quad \theta > 0.$$

The log likelihood function is

$$l(\theta) = -2n \log\theta - \frac{1}{\theta}\sum_{i=1}^n y_i, \quad \theta > 0$$

and

$$l'(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} y_i = \frac{1}{\theta^2}\left(\sum_{i=1}^{n} y_i - 2n\theta\right), \quad \theta > 0.$$

Now $l'(\theta) = 0$ if

$$\theta = \frac{1}{2n}\sum_{i=1}^{n} y_i = \frac{1}{2}\bar{y}.$$

(Note a First Derivative Test could be used to confirm that $l(\theta)$ has an absolute maximum at $\theta = \bar{y}/2$.) The maximum likelihood estimate of θ is

$$\hat{\theta} = \bar{y}/2.$$

(c)

$$E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} 2\theta = \frac{1}{n}\left(2n\theta\right) = 2\theta$$

and

$$Var\left(\bar{Y}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} 2\theta^2 = \frac{1}{n^2}\left(2n\theta^2\right) = \frac{2\theta^2}{n}.$$

(d) Since $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables then by the Central Limit Theorem

$$\frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \text{ has approximately a } \mathrm{N}(0,1) \text{ distribution.}$$

If $Z \backsim N(0,1)$

$$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95.$$

Therefore

$$P\left(-1.96 \leq \frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \leq 1.96\right) \approx 0.95.$$

(e) Since

$$\begin{aligned} 0.95 &\approx P\left(-1.96 \leq \frac{\bar{Y} - 2\theta}{\theta\sqrt{2/n}} \leq 1.96\right) \\ &= P\left(\bar{Y} - 1.96\theta\sqrt{2/n} \leq 2\theta \leq \bar{Y} + 1.96\theta\sqrt{2/n}\right) \\ &= P\left(\bar{Y}/2 - 0.98\theta\sqrt{2/n} \leq \theta \leq \bar{Y}/2 + 0.98\theta\sqrt{2/n}\right) \end{aligned}$$

an approximate 95% confidence interval for θ is

$$\left[\hat{\theta} - 0.98\hat{\theta}\sqrt{2/n},\ \hat{\theta} + 0.98\hat{\theta}\sqrt{2/n}\right]$$

where $\hat{\theta} = \bar{y}/2$.

(f) For these data the maximum likelihood estimate of θ is

$$\hat{\theta} = \bar{y}/2 = 88.92/\left(2 \times 18\right) = 2.47$$

and the approximate 95% confidence interval for θ is

$$2.47 \pm 0.98\left(2.47\right)\sqrt{\frac{2}{18}} = \left[1.66,\ 3.28\right].$$

4.16

(a)

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{2}\theta^3 t_i^2 \exp\left(-\theta t_i\right) = \left(\frac{1}{2^n}\prod_{i=1}^{n} t_i^2\right)\theta^{3n}\exp\left(-\theta\sum_{i=1}^{n} t_i\right)$$

or more simply

$$L(\theta) = \theta^{3n}\exp\left(-\theta\sum_{i=1}^{n} t_i\right), \quad \theta > 0.$$

The log likelihood function is

$$l(\theta) = 3n\log\theta - \theta\sum_{i=1}^{n} t_i \qquad \frac{dl}{d\theta} = \frac{3n}{\theta} - \sum_{i=1}^{n} t_i.$$

Solving $l(\theta) = 0$, we obtain the maximum likelihood estimate $\hat{\theta} = 3n/\sum_{i=1}^{n} t_i$.
The relative likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \left(\frac{\theta}{\hat{\theta}}\right)^{3n}\exp\left[3n\left(1 - \frac{\theta}{\hat{\theta}}\right)\right], \quad \theta > 0.$$

(b) Since $n = 20$ and $\sum_{i=1}^{20} t_i = 996$, therefore $\hat{\theta} = 3\left(20\right)/996 = 0.06024$. Reading from the graph in Figure 12.13 or by solving $R(\theta) = 0.15$ using the `uniroot` function in R, we obtain the 15% likelihood interval $[0.0463,\ 0.0768]$ which is an approximate 95% confidence interval for θ.

(c)

$$\begin{aligned} E\left(T\right) &= \frac{1}{2}\int_0^\infty \theta^3 t^3 e^{-\theta t} dt = \frac{1}{2}\int_0^\infty (\theta t)^3 e^{-(\theta t)} dt \\ &= \frac{1}{2\theta}\int_0^\infty x^3 e^{-x} dx \quad \text{(by letting } x = \theta t\text{)} \\ &= \frac{1}{2\theta}\Gamma(4) = \frac{1}{2\theta}3! = \frac{3}{\theta} \end{aligned}$$

and a 95% approximate confidence interval for $E\left(T\right) = 3/\theta$ is

$$\left[\frac{3}{0.0463}, \frac{3}{0.0768}\right] = \left[39.1,\ 64.8\right].$$

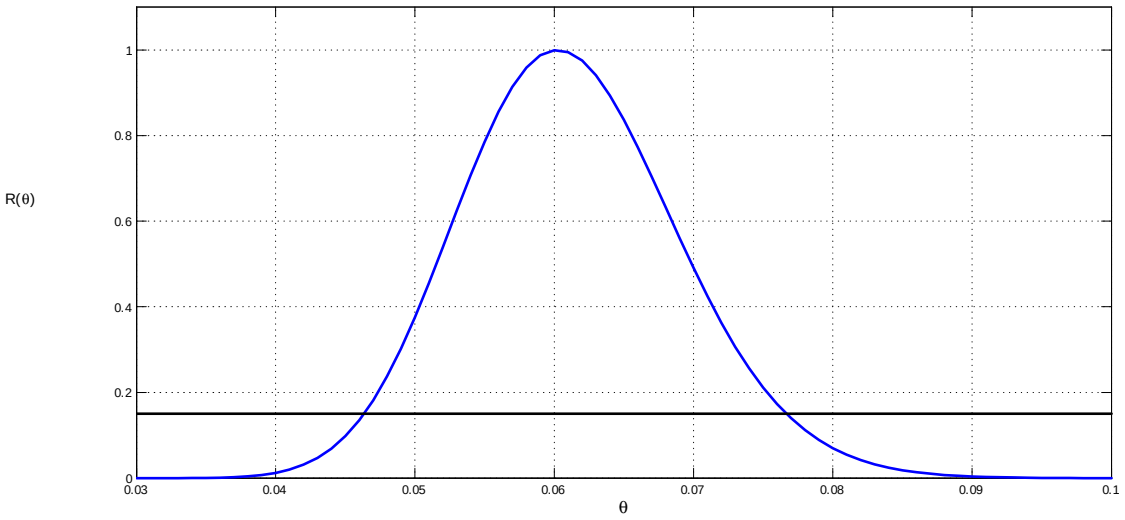


Figure 12.13: Relative Likelihood for Light Bulb Data

(d)

$$\begin{aligned} p(\theta) &= P(T \leq 50) = \frac{\theta^3}{2}\int_0^{50} t^2 e^{-\theta t} dt \\ &= \frac{\theta^3}{2}\left[\frac{-2500}{\theta}e^{-50\theta} - \frac{100}{\theta^2}e^{-50\theta} + \frac{2}{\theta^2}\left(-\frac{1}{\theta}e^{-50\theta} + \frac{1}{\theta}\right)\right] \\ &= 1 - \left(1250\theta^2 + 50\theta + 1\right)e^{-50\theta}. \end{aligned}$$

Since

$$p\left(0.0463\right) = 1 - \left[1250\left(0.0463\right)^2 + 50\left(0.0463\right) + 1\right]e^{-50(0.0463)} = 0.408$$

and

$$p\left(0.0768\right) = 1 - \left[1250\left(0.0768\right)^2 + 50\left(0.0768\right) + 1\right]e^{-50(0.0768)} = 0.738$$

the confidence intervals for $p(\theta)$ using the model is $[0.408,\ 0.738]$.
The confidence interval using the Binomial model is

$$\begin{aligned} \hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} &= \frac{11}{20} \pm 1.96\sqrt{\frac{\left(11/20\right)\left(9/20\right)}{20}} = 0.55 \pm 0.218 \\ &= [0.332,\ 0.768]\,. \end{aligned}$$

The Binomial model involves fewer model assumptions but gives a less precise (wider) interval.

4.17

(a) (i) If $X \backsim \chi^2(10)$ then $P(X \leq 2.6) \approx P(X \leq 2.558) = 0.01$ and $P(X > 16) \approx 1 - P(X \leq 15.987) = 1 - 0.9 = 0.1$.

(ii) If $X \backsim \chi^2(4)$ then $P(X > 15) \approx 1 - P(X \leq 14.86) = 1 - 0.995 = 0.005$.

(iii) If $X \backsim \chi^2(40)$ then $P(X \leq 24.4) \approx P(X \leq 24.433) = 0.025$ and $P(X \leq 55.8) \approx P(X \leq 55.758) = 0.95$.
If $Y \backsim N(40, 80)$ then

$$\begin{aligned} P(Y \leq 24.4) &= P\left(Z \leq \frac{24.4 - 40}{\sqrt{80}}\right) \text{ where } Z \backsim N(0,1) \\ &= P(Z \leq -1.74) = 1 - P(Z \leq 1.74) \\ &= 1 - 0.95907 = 0.04093 \approx 0.041 \end{aligned}$$

and

$$\begin{aligned} P(Y \leq 55.8) &= P\left(Z \leq \frac{55.8 - 40}{\sqrt{80}}\right) \text{ where } Z \backsim N(0,1) \\ &= P(Z \leq 1.77) = 0.96164 \approx 0.96 \end{aligned}$$

If $X \backsim \chi^2(40)$ then the graph of the probability density function of X will be fairly symmetric about the mean $E(X) = 40$ and very similar to the graph of the probability density function of a $N(40, 80)$ random variable. We note that $P(X \leq 55.8) = 0.95$ is close to $P(Y \leq 55.8) = 0.96$ while $P(X \leq 24.4) = 0.025$ and $P(Y \leq 24.4) = 0.041$ are not as close.

(iv) If $X \backsim \chi^2(25)$ then solving $P(X \leq a) = 0.025$ and $P(X > b) = 0.025$ gives $a = 13.120$ and $b = 40.646$.

(v) If $X \backsim \chi^2(12)$ then solving $P(X \leq a) = 0.05$ and $P(X > b) = 0.05$ gives $a = 5.226$ and $b = 21.026$.

(b)

(i) $P(X \leq 2.6) = pchisq(2.6, 10) = 0.01621621$,
$P(X > 16) = 1 - P(X \leq 16) = 1 - pchisq(16, 10) = 0.0996324$.

(ii) $P(X > 15) = 1 - P(X \leq 15) = 1 - pchisq(15, 4) = 0.004701217$.

(iii) $P(X \leq 24.4) = pchisq(24.4, 40) = 0.02469984$ and $P(X \leq 55.8) = pchisq(55.8, 40) = 0.950383$.

(iv) $a = qchisq(0.025, 25) = 13.11972$ and $b = qchisq(0.975, 25) = 40.64647$

(v) $a = qchisq(0.05, 12) = 5.226029$ and $b = qchisq(0.95, 12) = 21.02607$

(c)

(i) If $X \backsim \chi^2(1)$ then

$$\begin{aligned} P(X \leq 2) &= P\left(|Z| \leq \sqrt{2}\right) \text{ where } Z \backsim N(0,1) \\ &= 2P(Z \leq 1.41) - 1 = 2(0.92073) - 1 = 0.84146 \end{aligned}$$

and

$$\begin{aligned} P(X > 1.4) &= 1 - P\left(|Z| \leq \sqrt{1.4}\right) \text{ where } Z \backsim N(0,1) \\ &= 2\left[1 - P(Z \leq 1.41)\right] = 2(1 - 0.88100) = 0.23800 \end{aligned}$$

(ii) If $X \backsim \chi^2(2) = \text{Exponential}(2)$ then

$$P(X \leq 2) = 1 - e^{-2/2} = 1 - e^{-1} \approx 0.632$$

and

$$P(X > 3) = e^{-3/2} = e^{-1.5} \approx 0.223$$

(d) If $X \backsim G(3,2)$ then $\left(\frac{X-3}{2}\right)^2 \backsim \chi^2(1)$. Since $Y_i \backsim \textit{Exponential}(2)$, $i = 1, 2, \ldots, 5$ independently and $\textit{Exponential}(2)$ is the same distribution as $\chi^2(2)$,therefore $W = \sum_{i=1}^{5} Y_i + \left(\frac{X-3}{2}\right)^2 \backsim \chi^2(10+1)$ or $\chi^2(11)$.

(e) If $X_i \backsim \chi^2(i)$, $i = 1, 2, \ldots, 10$ independently then $\sum_{i=1}^{10} X_i \backsim \chi^2\left(\sum_{i=1}^{10} i\right)$ or $\chi^2(55)$.

4.18 (a)

$$\begin{aligned} \int_0^\infty \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{-\frac{y}{2}} dy &= \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^\infty \left(\frac{y}{2}\right)^{\frac{k}{2}-1} e^{-\frac{y}{2}} \frac{dy}{2} \quad \text{let } x = \frac{y}{2} \\ &= \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^\infty x^{\frac{k}{2}-1} e^{-x} dx \\ &= \frac{1}{\Gamma\left(\frac{k}{2}\right)} \Gamma\left(\frac{k}{2}\right) \quad \text{since } \int_0^\infty x^{\alpha-1} e^{-x} dx = \Gamma(\alpha) \\ &= 1 \end{aligned}$$

(b) See Figure 12.14. As k increases the probability density function becomes more symmetric about the line $y = k$.

(c)

$$\begin{aligned} M(t) &= E\left(e^{Yt}\right) = \int_0^\infty \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} y^{\frac{k}{2}-1} e^{-\frac{y}{2}} e^{yt} dy \\ &= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^\infty y^{\frac{k}{2}-1} e^{-(\frac{1}{2}-t)y} dy \quad \text{converges for } t < \frac{1}{2} \\ &= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})(\frac{1}{2}-t)^{\frac{k}{2}}} \int_0^\infty x^{\frac{k}{2}-1} e^{-x} dx \quad \text{by letting } x = \left(\frac{1}{2} - t\right) y \\ &= \left[2^{\frac{k}{2}}(\frac{1}{2}-t)^{\frac{k}{2}}\right]^{-1} = (1-2t)^{-\frac{k}{2}} \quad \text{for } t < \frac{1}{2} \end{aligned}$$

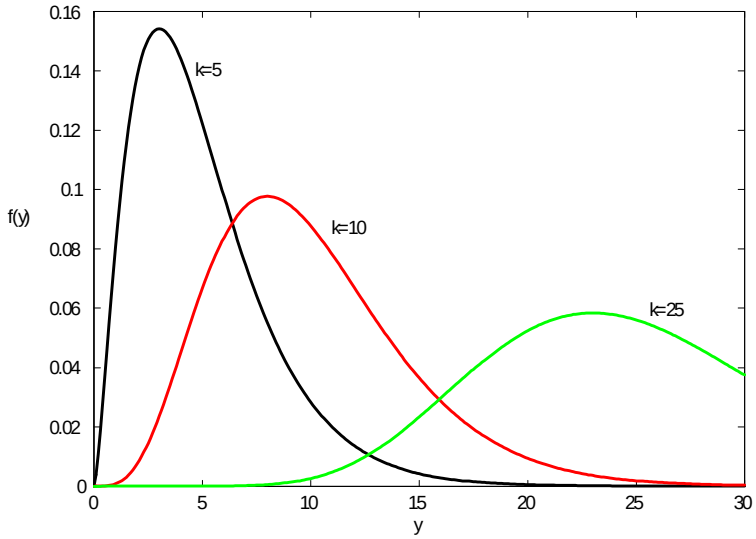


Figure 12.14: Chi-squared probability density functions for $k = 5, 10, 25$

Therefore

$$\begin{aligned} M^{'}(0) &= E(Y) = -\frac{k}{2}(1-2t)^{-\frac{k}{2}-1}(-2)|_{t=0} = k \\ M^{''}(0) &= E\left(Y^2\right) = -\frac{k}{2}\left[-\left(\frac{k}{2}+1\right)\right](1-2t)^{-\frac{k}{2}-2}(-2\times -2)|_{t=0} = k^2+2k \\ Var(Y) &= k^2+2k-k^2 = 2k \end{aligned}$$

(d) $W_i \backsim \chi^2(k_i)$ has moment generating function $M_i(t) = (1-2t)^{-k_i/2}$. Therefore $S = \sum_{i=1}^{n} W_i$ has moment generating function

$$M_s(t) = \prod_{i=1}^{n} M_i(t) = (1-2t)^{-\sum_{i=1}^{n} k_i/2}$$

which is the moment generating function of a χ^2 distribution with degrees of freedom equal to $\sum_{i=1}^{n} k_i$. Therefore $S \backsim \chi^2\left(\sum_{i=1}^{n} k_i\right)$ as required.

4.19

(a) Since

$$W_i = Y_i - \bar{Y} = Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i = \left(1-\frac{1}{n}\right)Y_i - \frac{1}{n}\sum_{j\neq i} Y_j \quad i = 1, 2, \ldots, n$$

therefore W_i is a linear combination of $Y_1, Y_2, \ldots, Y_n$ and therefore a linear combination of independent Normal random variables.

(b)

$$E(W_i) = E\left(Y_i - \bar{Y}\right) = E(Y_i) - E\left(\bar{Y}\right) = \mu - \mu = 0 \quad i = 1, 2, \ldots, n$$

Now $Cov(Y_i, Y_j) = 0$ if $i \neq j$ (since the $Y_i's$ are independent random variables) and $Cov(Y_i, Y_j) = \sigma^2$ if $i = j$ (since $Cov(Y_i, Y_i) = Var(Y_i) = \sigma^2$). This implies

$$\begin{aligned} Cov\left(Y_i, \bar{Y}\right) &= Cov\left(Y_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n} Cov\left(Y_i, \sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n} Cov(Y_i, Y_i) = \frac{1}{n} Var(Y_i) = \frac{\sigma^2}{n}. \end{aligned}$$

Therefore

$$\begin{aligned} Var(W_i) &= Var\left(Y_i - \bar{Y}\right) = Var(Y_i) + Var\left(\bar{Y}\right) - 2Cov\left(Y_i, \bar{Y}\right) \\ &= \sigma^2 + \frac{\sigma^2}{n} - 2\left(\frac{\sigma^2}{n}\right) = \sigma^2\left(1 - \frac{1}{n}.\right) \end{aligned}$$

(c)

$$\begin{aligned} Cov(W_i, W_j) &= Cov\left(Y_i - \bar{Y}, Y_j - \bar{Y}\right) \quad i \neq j \\ &= Cov(Y_i, Y_j) - Cov\left(Y_i, \bar{Y}\right) - Cov\left(\bar{Y}, Y_j\right) + Cov\left(\bar{Y}, \bar{Y}\right) \\ &= 0 - \frac{\sigma^2}{n} - \frac{\sigma^2}{n} + Var\left(\bar{Y}\right) \\ &= -\frac{2\sigma^2}{n} + \frac{\sigma^2}{n} = -\frac{\sigma^2}{n} \end{aligned}$$

4.20 (a) The graph is given in Figure 12.15. As k increases the graphs become more and more like the graph of the $N(0,1)$ probability density function and for $k = 25$ there is little difference between the $t(25)$ probability density function and the $N(0,1)$ probability density function.

(b)

$$\begin{aligned} \frac{d}{dt} f(t;k) &= \frac{d}{dt} c_k \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} = c_k \left(-\frac{k+1}{2}\right)\left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}-1} \frac{2t}{k} \\ &= t \cdot c_k \left(-\frac{k+1}{k}\right)\left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}-1} = 0 \text{ if } t = 0 \end{aligned}$$

Since $\frac{d}{dt} f(t;k) > 0$ if $t < 0$ and $\frac{d}{dt} f(t;k) < 0$ if $t > 0$ then by the First Derivative Test $f(t;k)$ has a global maximum at $t = 0$.

(c)

$$\begin{aligned} E(T) &= E\left(\frac{Z}{\sqrt{\frac{U}{k}}}\right) = E(Z)\, E\left(\frac{1}{\sqrt{\frac{U}{k}}}\right) \\ &\qquad \text{since } Z \text{ and } U \text{ are independent random variables} \\ &= 0 \quad \text{since } E(Z) = 0. \end{aligned}$$

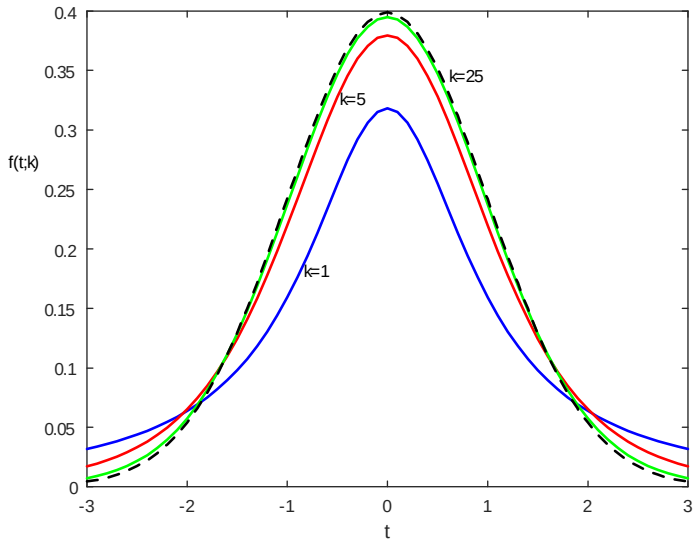


Figure 12.15: Graphs of the $t\,(k)$ pd.f. for $k = 1, 5, 25$ and the $N\,(0,1)$ p.d.f. (dashed line)

(d)

(i) If $T \backsim t(10)$ then $P(T \leq 0.88) = 0.8$,

$$P(T \leq -0.88) = P(T > 0.88) \approx 1 - P(T \leq 0.8791) \approx 1 - 0.8 = 0.2$$

and

$$\begin{aligned} P(|T| &\leq 0.88) = P\left(-0.88 \leq T \leq 0.88\right) = P(T \leq 0.88) - P(T \leq -0.88) \\ &= P(T \leq 0.88) - [1 - P(T \leq 0.88)] = 2P(T \leq 0.88) - 1 \\ &\approx 2P(T \leq 0.8791) - 1 = 2\,(0.8) - 1 = 0.6. \end{aligned}$$

(ii) If $T \backsim t(17)$ then

$$\begin{aligned} P(|T| &\geq 2.90) = 2P(T \geq 2.90) \text{ by symmetry} \\ &= 2\left[1 - P(T \leq 2.90)\right] \approx 2\left[1 - P(T \leq 2.8982)\right] = 2\,(1 - 0.995) = 0.01 \end{aligned}$$

(iii) If $T \backsim t(30)$ then

$$\begin{aligned} P(T &\leq -2.04) = P(T \geq 2.04) = 1 - P(T \leq 2.04) \\ &\approx 1 - P(T \leq 2.0423) = 1 - 0.975 = 0.025 \end{aligned}$$

and if $Z \backsim N\,(0,1)$ then

$$\begin{aligned} P(Z &\leq -2.04) = 1 - P(Z \leq 2.04) \\ &= 1 - 0.97932 = 0.02068 \end{aligned}$$

and these values are close.

If $T \backsim t(30)$ then $P(T \leq 0.26) \approx P(T \leq 0.2556) = 0.6$ which is close to $P(Z \leq 0.26) = 0.60257$ if $Z \backsim N(0,1)$.

(iv) If $T \backsim t(18)$ then $P(T \leq 2.1009) = 0.975$ so $P(T \geq 2.1009) = 0.025$ and by symmetry $P(T \leq -2.1009) = 0.025$. Therefore $a = -2.1009$ and $b = 2.1009$.

(v) If $T \backsim t(13)$ then $P(T \leq 1.7709) = 0.95$ so $P(T \geq 1.7709) = 0.05$ and by symmetry $P(T \leq -1.7709) = 0.05$. Therefore $a = -1.7709$ and $b = 1.7709$.

(i) $P(T \leq -0.88) = pt(-0.88, 10) = 0.1997567$ and $P(|T| \leq 0.88) = 2P(T \leq 0.88) - 1 = 2 \times pt(0.88, 10) - 1 = 0.6004867$

(ii) $P(|T| \geq 2.90) = 2P(T \geq 2.90) = 2[1 - P(T \leq 2.90)] = 2[1 - pt(2.90, 17)] = 0.009962573$

(iii) $P(T \leq -2.04) = pt(-2.04, 30) = 0.02511979$ and $P(T \leq 0.26) = pt(0.26, 30) = 0.60168$

(iv) $a = qt(0.025, 18) = -2.100922$ and $b = qt(0.975, 18) = 2.100922$

(v) $a = qt(0.05, 13) = -1.770933$ and $b = qt(0.95, 13) = 1.770933$

4.21

$$\begin{aligned}
\lim_{k\to\infty} f(t;k) &= \lim_{k\to\infty} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \\
&= \lim_{k\to\infty} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k}{2}} \left(1 + \frac{t^2}{k}\right)^{-\frac{1}{2}} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \quad \text{for } t \in \Re
\end{aligned}$$

since

$$\begin{aligned}
\lim_{k\to\infty} c_k &= \frac{1}{\sqrt{2\pi}} \\
\lim_{k\to\infty} \left(1 + \frac{t^2}{k}\right)^{-\frac{1}{2}} &= 1 \text{ and} \\
\lim_{k\to\infty} \left(1 + \frac{t^2}{k}\right)^{-\frac{k}{2}} &= \exp\left(-\frac{1}{2}t^2\right) \text{ since } \lim_{y\to\infty} \left(1 + \frac{a}{n}\right)^{bn} = e^{ab}
\end{aligned}$$

4.22 (a) From Example 2.3.2

$$L(\theta) = \theta^{-n} e^{-n\bar{y}/\theta} \quad \text{for } \theta > 0 \quad \text{and} \quad \hat{\theta} = \bar{y}.$$

Therefore

$$\begin{aligned}
R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} = \frac{L(\theta)}{L(\bar{y})} = \frac{\theta^{-n} e^{-n\bar{y}/\theta}}{(\bar{y})^{-n} e^{-n}} = \left(\frac{\bar{y}}{\theta}\right)^n e^{n(1-\bar{y}/\theta)} \\
&= \left[\frac{\bar{y}}{\theta} e^{(1-\bar{y}/\theta)}\right]^n \quad \text{for } \theta > 0.
\end{aligned}$$

For the given data $n = 30$ and $\hat{\theta} = \frac{1}{30}(11400) = 380$ and

$$R(\theta) = \left[\frac{380}{\theta} e^{(1-380/\theta)}\right]^{30} \quad \text{for } \theta > 0.$$

From Inverse Normal Tables

$$\begin{aligned} 0.90 &= P(|Z| \leq 1.6449) \quad \text{where } Z \backsim N(0,1) \\ &= P\left(W \leq (1.6449)^2\right) \quad \text{where } W \backsim \chi^2(1) \\ &= P(W \leq 2.7057). \end{aligned}$$

Since (see Section 4.6) $\{\theta : \Lambda(\theta) \leq 2.7057\} = \left\{\theta : 2l(\hat{\theta}) - 2l(\theta) \leq 2.7057\right\}$ is an approximate 90% confidence interval. Therefore

$$\begin{aligned} & \left\{\theta : 2l(\hat{\theta}) - 2l(\theta) \leq 2.7057\right\} \\ = & \left\{\theta : R(\theta) \geq e^{-2.7057/2}\right\} \\ = & \{\theta : R(\theta) \geq 0.2585\} \end{aligned}$$

which implies that a 26% likelihood interval is an approximate 90% confidence interval.
Using the uniroot function in R and

$$R(\theta) = \left[\frac{380}{\theta} e^{(1-380/\theta)}\right]^{30} \quad \text{for } \theta > 0$$

we obtain the interval as [285.5, 521.3]. Alternatively the likelihood interval can be determined approximately from a graph of the relative likelihood function. See Figure 12.16.

(b) Since $P(X \leq m) = 1 - e^{-m/\theta} = 0.5$, therefore $m = -\theta \log(0.5) = \theta \log 2$ and the confidence interval for m is $[285.5 \log 2,\ 521.3 \log 2] = [197.9,\ 361.3]$ by using the confidence interval for θ obtained in (a).

4.23 (a) Let $F(y) = P(Y \leq y)$ be the cumulative distribution function of Y. For $w > 0$,

$$G(w) = P(W \leq w) = P\left(\frac{2Y}{\theta} \leq w\right) = P\left(Y \leq \frac{\theta w}{2}\right) = F\left(\frac{\theta w}{2}\right).$$

Taking the derivative with respect to w gives the probability density function of W as

$$\begin{aligned} g(w) &= \frac{d}{dw} G(w) = \frac{d}{dw} F\left(\frac{\theta w}{2}\right) \\ &= f\left(\frac{\theta w}{2}\right) \frac{d}{dw}\left(\frac{\theta w}{2}\right) = \frac{1}{\theta} e^{-\left(\frac{\theta w}{2}\right)/\theta}\left(\frac{\theta}{2}\right) \\ &= \frac{1}{2} e^{-\frac{w}{2}} \quad \text{for } w > 0 \end{aligned}$$

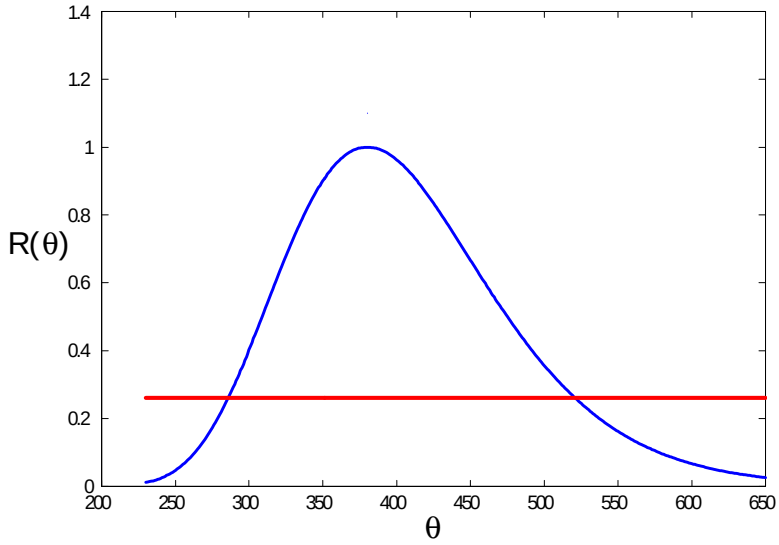


Figure 12.16: Relative likelihood function for survival times for AIDS patients

which can be easily verified as the probability density function of a $\chi^2(2)$ random variable.

(b) Let $W_i = 2Y_i/\theta \backsim \chi^2(2)$, $i = 1, 2, \ldots, n$ independently. Then by Theorem 29,

$$U = \sum_{i=1}^{n} W_i = \sum_{i=1}^{n} \frac{2Y_i}{\theta} \backsim \chi^2(2n).$$

(c) Since

$$0.9 = P(43.19 \leq W \leq 79.08) \quad \text{where } W \backsim \chi^2\,(60)$$

therefore

$$\begin{aligned} 0.9 &= P\left(43.19 \leq \frac{2}{\theta}\sum_{i=1}^{n} Y_i \leq 79.08\right) \\ &= P\left(\frac{2}{79.08}\sum_{i=1}^{n} Y_i \leq \theta \leq \frac{2}{43.19}\sum_{i=1}^{n} Y_i\right) \end{aligned}$$

and thus

$$\left[\frac{2}{79.08}\sum_{i=1}^{n} y_i,\ \ \frac{2}{43.19}\sum_{i=1}^{n} y_i\right]$$

is a 90% confidence interval for θ. Substituting $\sum_{i=1}^{30} y_i = 11400$, we obtain the 90% confidence interval for θ as $[288.3,\ 527.9]$ which is very close to the approximate 90% likelihood-based confidence interval $[285.5,\ 521.3]$. The intervals are close since $n = 30$ is reasonably large.

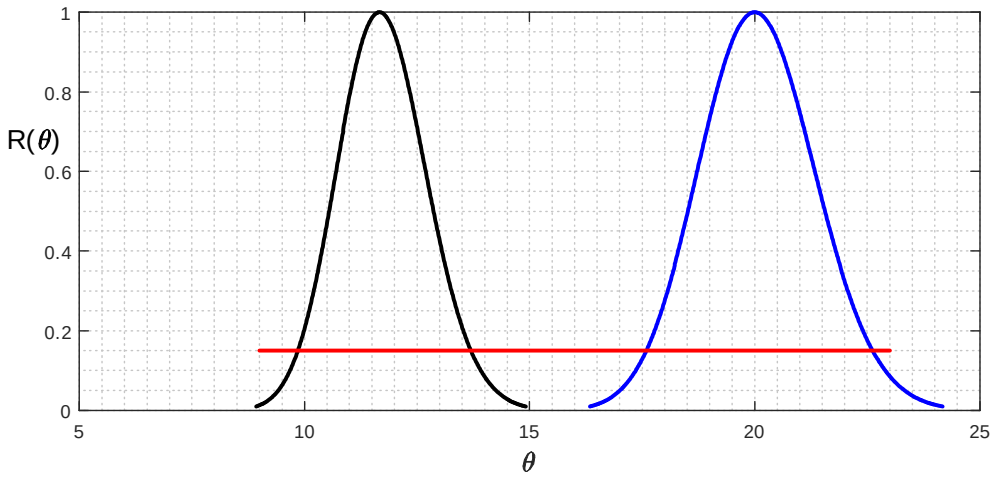


Figure 12.17: Relative Likelihood Functions for Company A and Company B Photocopiers

4.24 (a) From Example 2.2.2 the likelihood function for Poisson data is

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

with corresponding maximum likelihood estimate $\hat{\theta} = \bar{y}$. For Company A, $n = 12$ and $\hat{\theta} = 20$ and the relative likelihood function is

$$R(\theta) = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\bar{y}^{n\bar{y}} e^{-n\bar{y}}} \quad \text{for } \theta > 0.$$

See Figure 12.17 for a graph of the relative likelihood function (graph on the right).

(b) For Company B, $n = 12$ and $\hat{\theta} = 11.67$. See Figure 12.17 for a graph of the relative likelihood function (graph on the left).

(c) The 15% likelihood interval for Company A is: [17.59, 22.62] and the 15% likelihood interval for Company B is: [9.84, 13.71]. It is clear from these approximate 95% confidence intervals that the mean number of service calls for Company A is much larger than for Company B which implies the decision to go with Company B is a good one.

(d) The assumptions of the Poisson process (individuality, independence and homogeneity) would need to hold.

(e) Since

$$\begin{aligned} 0.95 &\approx P\left(-1.96 \leq \frac{\bar{Y} - \theta}{\sqrt{\bar{Y}/n}} \leq 1.96\right) \\ &= P\left(\bar{Y} - 1.96\sqrt{\bar{Y}/n} \leq \theta \leq \bar{Y} + 1.96\sqrt{\bar{Y}/n}\right) \end{aligned}$$

therefore the interval $\left[\bar{y} - 1.96\sqrt{\bar{y}/n},\ \bar{y} + 1.96\sqrt{\bar{y}/n}\right]$ is an approximate 95% confidence interval for θ. For Company A this interval is $[17.5, 22.5]$ and for Company B this interval is $[9.73, 13.60]$. These intervals are similar but not identical to the intervals in (c) since $n = 12$ is small. The intervals would be more similar for a larger value of n.

4.25 (a) Since the points in the qqplot in Figure 12.18 lie reasonably along a straight line the Gaussian model seems reasonable for these data.

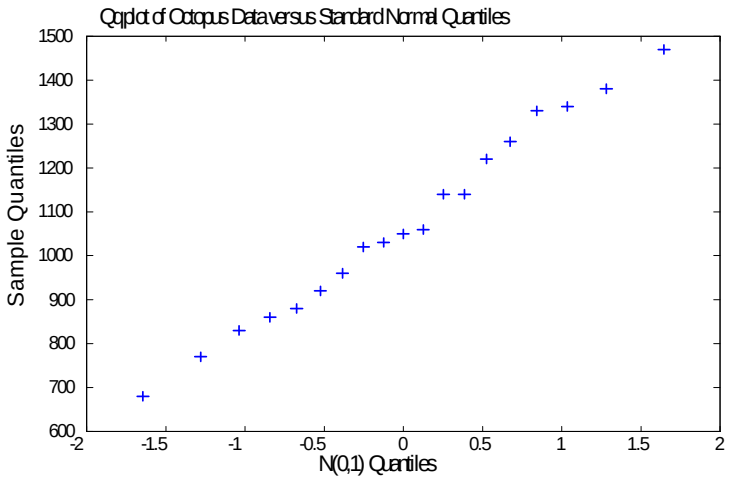


Figure 12.18: Qqplot for octopus data

(b) A suitable study population for this study would be common octopi in the Ria de Vigo. The parameter μ represents the mean weight in grams of common octopi in the Ria de Vigo. The parameter σ represents the standard deviation of the weights in grams of common octopi in the Ria de Vigo.

(c) Since $P(T \leq 2.1009) = (1 + 0.95)/2 = 0.975$,

$$\hat{\mu} = \bar{y} = \frac{20340}{19} = 1070.526 \quad \text{and} \quad s = \left[\frac{1}{18}(884095)\right]^{1/2} = 221.62$$

therefore a 95% confidence interval for μ is

$$\begin{aligned} & 1070.526 \pm 2.1009\,(221.62)/\sqrt{19} = 1070.526 \pm 106.817 \\ = \ & [963.709, 1177.343] \end{aligned}$$

Since the value $\mu = 1100$ grams is well within this interval then the researchers could conclude that based on these data the octopi in the Ria de Vigo are reasonably healthy based on their mean weight.

(d) Since $P(W \leq 9.391) = 0.05 = P(W \geq 28.869)$ where $W \backsim \chi^2(18)$ a 90% confidence interval for σ for the given data is

$$\left[\left(\frac{884095}{28.869}\right)^{1/2}, \left(\frac{884095}{9.391}\right)^{1/2}\right] = \left[(306.24)^{1/2}, (941.42)^{1/2}\right] = [175.00, 306.83].$$

4.26 (a) Qqplots of the weights for females and males separately are shown in Figures 12.19 and 12.20. In both cases the points lie reasonably along a straight line so it is reasonable to assume a Normal model for each data set.

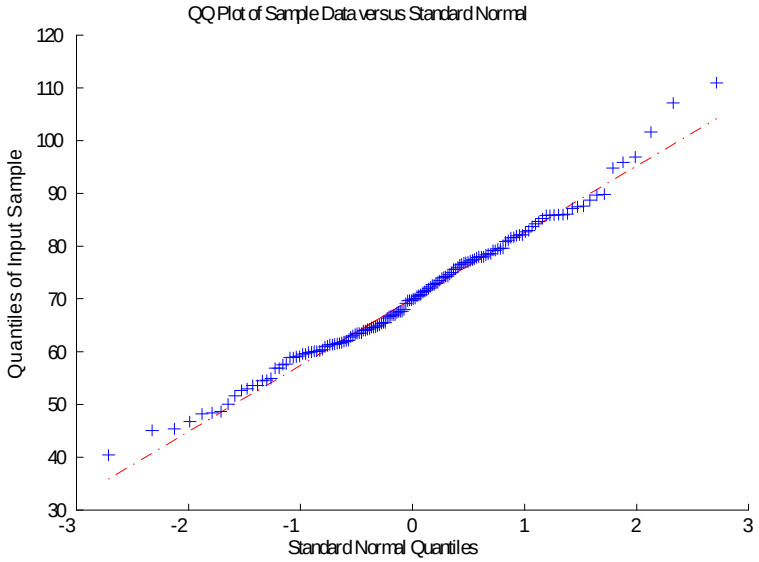


Figure 12.19: Qqplot of female weights

(b) A 95% confidence interval for the mean weight of females is

$$\begin{aligned} & \left[\bar{y}_f - 1.9647 s_f/\sqrt{150}, \bar{y}_f + 1.9647 s_f/\sqrt{150}\right] \\ = & \left[70.4432 - (1.9647)(12.5092)/\sqrt{150}, 70.4432 + (1.9647)(12.5092)/\sqrt{150}\right] \\ = & [68.4365, 72.4499]. \end{aligned}$$

A 95% confidence interval for the mean weight of males is

$$\begin{aligned} & \left[\bar{y}_m - 1.9647 s_m/\sqrt{150}, \bar{y}_m + 1.9647 s_m/\sqrt{150}\right] \\ = & \left[82.5919 - (1.9647)(12.8536)/\sqrt{150}, 82.5919 + (1.9647)(12.8536)/\sqrt{150}\right] \\ = & [80.5300, 84.6539]. \end{aligned}$$

Note that since the value for $t(149)$ is not available in the t-tables we used $P(T \leq 1.9647) = (1 + 0.95)/2 = 0.975$ where $T \backsim t(100)$. Using R we obtain

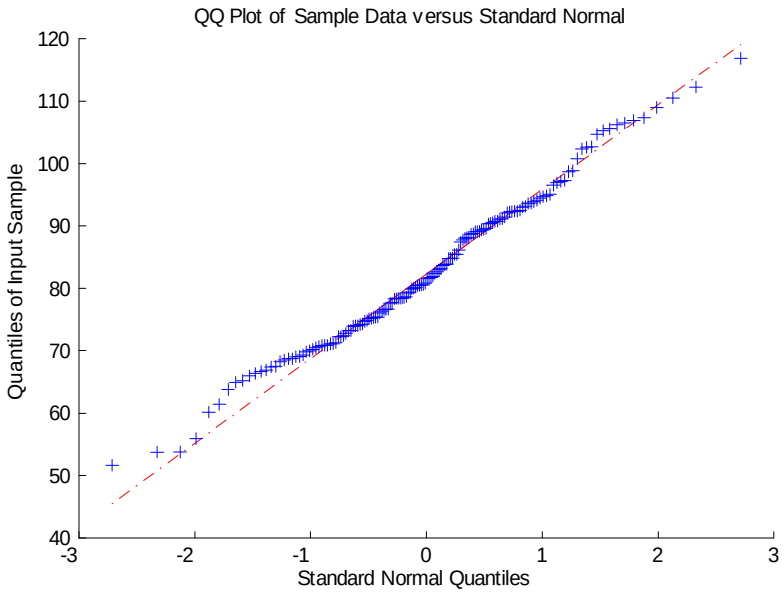


Figure 12.20: Qqplot of male weights

$P(T \leq 1.976) = 0.975$ where $T \backsim t(149)$. The intervals will not change substantially.

We note that the interval for females and the interval for males have no values in common. The mean weight for males is higher than the mean weight for females.

(c) To obtain confidence intervals for the standard deviations we note that the pivotal quantity $(n-1) S^2/\sigma^2 = 149S^2/\sigma^2$ has a $\chi^2(149)$ distribution and the Chi-squared tables stop at degrees of freedom = 100. Since $E\left(149S^2/\sigma^2\right) = 149$ and $Var\left(149S^2/\sigma^2\right) = 2(149) = 298$ we use $149S^2/\sigma^2 \backsim N(149, 298)$ approximately to construct an approximate 95% confidence interval given by

$$\left[\sqrt{\frac{149s^2}{149+1.96\sqrt{298}}}, \sqrt{\frac{149s^2}{149-1.96\sqrt{298}}}\right]$$
$$= \left[\sqrt{\frac{149s^2}{182.8348}}, \sqrt{\frac{149s^2}{115.1652}}\right].$$

For the females we obtain

$$\left[\sqrt{\frac{149\,(156.4806)}{182.8348}}, \sqrt{\frac{149\,(156.4806)}{115.1652}}\right]$$
$$= \left[\sqrt{127.5228}, \sqrt{202.4536}\right] = [11.2926, 14.2286].$$

For the males we obtain

$$\begin{aligned} & \left[\sqrt{\frac{149\,(165.2162)}{182.8348}}, \sqrt{\frac{149\,(165.2162)}{115.1652}}\right] \\ = & \left[\sqrt{134.6418}, \sqrt{213.7558}\right] = [11.6035, 14.6204]\,. \end{aligned}$$

These intervals are quite similar.

4.27 (a) A suitable study population consists of the detergent packages produced by this particular detergent packaging machine. The parameter μ corresponds to the mean weight of the detergent packages produced by this detergent packaging machine. The parameter σ is the standard deviation of the weights of the detergent packages produced by this detergent packaging machine.

(b) For these data

$$\begin{aligned} \bar{y} &= \frac{4803}{16} = 300.1875 \\ s^2 &= \frac{1}{15}\left[1442369 - 16\,(300.1875)^2\right] = 37.89583 \\ s &= 6.155959. \end{aligned}$$

Since $P\,(T \leq 2.1314) = (1 + 0.95)/2 = 0.975$ where $T \backsim t\,(15)$, a 95% confidence interval for μ is

$$\begin{aligned} 300.1875 \pm (2.1314)\,(6.155959)\,/\sqrt{16} &= 300.1875 \pm 3.2803 \\ &= [296.91,\ 303.47]\,. \end{aligned}$$

Since $P\,(W \leq 6.262) = (1 - 0.95)\,/2 = 0.025$ and $P\,(W \leq 27.488) = (1 + 0.95)\,/2 = 0.975$, a 95% confidence interval for σ

$$\left[\sqrt{\frac{(15)\,(37.89583)}{27.488}},\ \sqrt{\frac{(15)\,(37.89583)}{6.262}}\right] = [4.55,\ 9.53]$$

(c) Since $P\,(T \leq 2.1314) = (1 + 0.95)/2 = 0.975$ where $T \backsim t\,(15)$, a 95% prediction interval for Y is

$$\begin{aligned} & 300.1875 \pm (2.1314)\,(6.155959)\sqrt{1 + \frac{1}{16}} \\ = & \ 300.1875 \pm 13.5249 \\ = & \ [286.7,\ 313.7]\,. \end{aligned}$$

4.28 For the radon data

$$n = 12, \quad \bar{y} = 104.1333 \ \text{ and } \ s = \left[\frac{1}{11}\sum_{i=1}^{12}(y_i - \bar{y})^2\right]^{1/2} = 9.3974.$$

From t tables, $P(T \leq 2.20) = (1 + 0.95)/2 = 0.975$ where $T \backsim t(11)$. Therefore a 95% prediction interval for Y, the reading for the new radon detector exposed to 105 picocuries per liter of radon over 3 days, is

$$\left[104.1333 - 2.20\,(9.3974)\left(1 + \frac{1}{12}\right)^{1/2},\ 104.1333 + 2.20\,(9.3974)\left(1 + \frac{1}{12}\right)^{1/2}\right]$$
$$= [104.1333 - 21.5185,\ 104.1333 + 21.5185]$$
$$= [82.6148,\ 125.6519]\,.$$

4.29 Use $\sigma^2 \approx s^2 = \frac{45}{9} = 5$ and $d = 0.5$. Hence, $n \approx \left(\frac{1.96\sigma}{d}\right)^2 = \left(\frac{1.96}{0.5}\right)^2 5 = 76.832$. Since 10 observations have already been taken, the manufacturer should be advised to take at least $77 - 10 = 67$ additional measurements. We note that this calculation depends on an estimate of σ from a small sample ($n = 10$) and the value 1.96 is from the Normal tables rather than the t tables so the manufacturer should be advised to take more than 67 additional measurements. If we round 1.96 to 2 to account for the the fact that we don't actually know σ, and note that $\left(\frac{2}{0.5}\right)^2 5 = 80$ then this would suggest that, to be safe, the manufacturer should take an additional $80 - 10 = 70$ measurements.

4.30 (a) The combined likelihood function for μ is

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2}(x_i - \mu)^2\right] \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}(y_i - \mu)^2\right] \\ &= (2\pi)^{-(n+m)/2}\,\sigma_1^{-m}\sigma_2^{-n} \exp\left\{-\frac{1}{2\sigma_1^2}\left[\sum_{i=1}^{m}(x_i - \bar{x})^2 + m(\bar{x} - \mu)^2\right]\right\} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_2^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right]\right\} \end{aligned}$$

or more simply ignoring constants

$$L(\mu) = \exp\left[\frac{-m}{2\sigma_1^2}(\bar{x} - \mu)^2 - \frac{n}{2\sigma_2^2}(\bar{y} - \mu)^2\right] \quad \text{for } \mu \in \Re$$

since σ_1^2 and σ_2^2 are known. The log likelihood function is

$$l(\mu) = -\frac{m}{2\sigma_1^2}(\bar{x} - \mu)^2 - \frac{n}{2\sigma_2^2}(\bar{y} - \mu)^2\,.$$

Solving

$$l'(\mu) = \frac{m}{\sigma_1^2}(\bar{x} - \mu) + \frac{n}{\sigma_2^2}(\bar{y} - \mu) = \frac{\left(m\sigma_2^2\bar{x} + n\sigma_1^2\bar{y}\right) - \left(m\sigma_2^2 + n\sigma_1^2\right)\mu}{\sigma_1^2\sigma_2^2} = 0$$

gives the maximum likelihood estimate for μ as

$$\begin{aligned} \hat{\mu} &= \frac{m\sigma_2^2\bar{x} + n\sigma_1^2\bar{y}}{m\sigma_2^2 + n\sigma_1^2} = \frac{\left(m/\sigma_1^2\right)\bar{x} + \left(n/\sigma_2^2\right)\bar{y}}{m/\sigma_1^2 + n/\sigma_2^2} \\ &= \frac{w_1\bar{x} + w_2\bar{y}}{w_1 + w_2} \end{aligned}$$

where $w_1 = m/\sigma_1^2$ and $w_2 = n/\sigma_2^2$. We first note that both $\bar{x}$ and $\bar{y}$ are both estimates of μ and it makes sense to take a weighted average of the two estimates to get a better estimate of μ. If the sample sizes n and m are not equal it makes sense to weight the estimate that is a function of more observations. It also makes sense that the mean of the observations that come from a distribution with smaller variance is a better estimate of μ and should be given more weight. By examining the weights w_1 and w_2 we can see that the estimate $\hat{\mu}$ does satisfies both of these requirements.

(b) Since the observations in $\bar{x}$ are observations from a distribution with larger variability then we don't want to take just an average of $\bar{x}$ and $\bar{y}$. We would choose an estimate that weights $\bar{y}$ more that $\bar{x}$ since $\bar{y}$ is a better estimate.

(c)

$$\begin{aligned} Var(\tilde{\mu}) &= Var\left(\frac{\overline{X}+4\overline{Y}}{5}\right) = \frac{1}{25}\left[Var(\overline{X}) + 16Var(\bar{Y})\right] \\ &= \frac{1}{25}\left[\frac{1}{10} + 16\left(\frac{0.25}{10}\right)\right] = 0.02. \end{aligned}$$

and $\sqrt{Var(\tilde{\mu})} = 0.1414$.

$$Var\left(\frac{\overline{X}+\overline{Y}}{2}\right) = \frac{1}{4}\left[Var(\overline{X}) + Var(\overline{Y})\right] = \frac{1}{4}\left(\frac{1}{10} + \frac{0.25}{10}\right) = 0.03125.$$

and $\sqrt{Var\left(\frac{\overline{X}+\overline{Y}}{2}\right)} = 0.1768$. We can clearly see now that $\tilde{\mu}$ has a smaller standard deviation than the estimator $\left(\overline{X}+\overline{Y}\right)/2$.

Chapter 5

5.1. (a) The model $Y \backsim \text{Binomial}(n, \theta)$ is appropriate in the case in which the experiment consists of a sequence of n independent trials with two outcomes on each trial (Success and Failure) and $P(\text{Success}) = \theta$ is the same on each trial. In this experiment the trials are the guesses. Since the deck is reschuffled each time it seems reasonable to assume the guesses are independent. It also seems reasonable to assume that the women's ability to guess the number remains the same on each trial. To test the hypothesis that the women is guessing at random the appropriate null hypothesis would be $H_0 : \theta = \frac{1}{5} = 0.2$.

(b) For $n = 20$ and $H_0 : \theta = 0.2$, we have $Y \backsim \text{Binomial}(20, 0.2)$ and $E(Y) = 20(0.2) = 4$. We use the test statistic or discrepancy measure $D = |Y - E(Y)| = |Y - 4|$. The observed value of D is $d = |8 - 4| = 4$. Then

$$\begin{aligned} p - value &= P(D \geq 4; H_0) = P(|Y - 4| \geq 4; H_0) \\ &= P(Y = 0) + P(Y \geq 8) \\ &= \binom{20}{0}(0.2)^0(0.8)^{20} + \sum_{y=8}^{20}\binom{20}{y}(0.2)^y(0.8)^{20-y} \\ &= 1 - \sum_{y=1}^{7}\binom{20}{y}(0.2)^y(0.8)^{20-y} = 0.04367 \quad \text{using } R. \end{aligned}$$

There is evidence based on the data against $H_0 : \theta = 0.2$. These data suggest that the woman might have some special guessing ability.

(c) For $n = 100$ and $H_0 : \theta = 0.2$, we have $Y \backsim \text{Binomial}(100, 0.2)$, $E(Y) = 100(0.2) = 20$ and $Var(Y) = 100(0.2)(0.8) = 16$. We use the test statistic or discrepancy measure $D = |Y - E(Y)| = |Y - 20|$. The observed value of D is $d = |32 - 20| = 12$. Then

$$\begin{aligned} p - value &= P(D \geq 12; H_0) = P(|Y - 20| \geq 12) \\ &= P(Y \leq 8) + P(Y \geq 32) \\ &= \sum_{y=0}^{8}\binom{100}{y}(0.2)^y(0.8)^{100-y} + \sum_{y=32}^{100}\binom{100}{y}(0.2)^y(0.8)^{100-y} \\ &= 1 - \sum_{y=9}^{31}\binom{100}{y}(0.2)^y(0.8)^{100-y} = 0.004 \quad \text{using } R. \end{aligned}$$

or

$$\begin{aligned} p - value &= P(D \geq 12; H_0) = P(|Y - 20| \geq 12) \\ &\approx P\left(|Z| \geq \frac{12}{\sqrt{16}}\right) \quad \text{where } Z \backsim N(0, 1) \\ &= 2[1 - P(Z \leq 3)] = 2(1 - 0.99865) = 0.0027. \end{aligned}$$

There is strong evidence based on the data against $H_0 : \theta = 0.2$. These data suggest that the woman has some special guessing ability. Note that we would not conclude that it has been proven that she does have special guessing ability!

5.2 Assuming $H_0 : \theta = 10$ is true $Y \backsim \text{Poisson}(10)$. Therefore

$$\begin{aligned} P(D &\geq 15; H_0) = P(Y - 10 \geq 15) = P(Y \geq 25) \\ &= 1 - \sum_{y=0}^{24} \frac{10^y e^{-10}}{y!} = 0.000047 \quad \text{using } R \end{aligned}$$

or

$$\begin{aligned} P(D &\geq 15; H_0) = P(Y \geq 25) \approx P\left(Z \geq \frac{25-10}{\sqrt{10}}\right) \quad \text{where } Z \backsim N(0,1) \\ &= P(Z \geq 4.74) \approx 0 \end{aligned}$$

There is very strong evidence based on the data against $H_0 : \theta = 10$.

5.3 (a) A qqplot of the data is given in Figure 12.21. Since the points in the qqplot lie reasonably along a straight line it seems reasonable to assume a Normal model for these data.

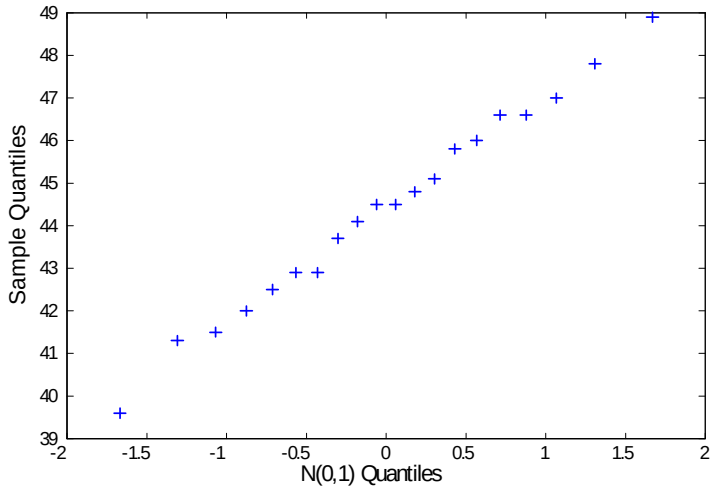


Figure 12.21: Qqplot for Dioxin data

(b) A study population is a bit difficult to define in this problem. One possible choice is to define the study population to be all measurements that could be taken on a given day by this instrument on a standard solution of 45 parts per billion dioxin. The parameter μ corresponds to the mean measurement made by this instrument on the standard solution. The parameter σ corresponds to the standard deviation of the measurements made by this instrument on the standard solution.

(c) For these data

$$\bar{y} = \frac{888.1}{20} = 44.405 \text{ and } s = \left[\frac{39545.03 - 20\left(44.405\right)^2}{19}\right]^{1/2} = 2.3946$$

To test $H_0 : \mu = 45$ we use the test statistic

$$D = \frac{\left|\bar{Y} - 45\right|}{S/\sqrt{20}} \text{ where } T = \frac{\bar{Y} - 45}{S/\sqrt{20}} \backsim t\left(19\right).$$

The observed value of D is

$$d = \frac{|44.405 - 45|}{2.3946/\sqrt{20}} = 1.11$$

and

$$\begin{aligned} p - value &= P\left(D \geq d; H_0\right) \\ &= P\left(|T| \geq 1.11\right) \quad \text{where } T \backsim t\left(19\right) \\ &= 2\left[1 - P\left(T \leq 1.11\right)\right] \\ &= 0.2803 \quad \text{(calculated using } R\text{)}. \end{aligned}$$

Alternatively using the t-tables in the Course Notes we have $P\left(T \leq 0.8610\right) = 0.8$ and $P\left(T \leq 1.3277\right) = 0.9$ so

$$\begin{aligned} 2\left(1 - 0.9\right) \leq p - value \leq 2\left(1 - 0.8\right) \\ \text{or } \ 0.2 \leq p - value \leq 0.4. \end{aligned}$$

In either case since the $p - value$ is larger than 0.1 and we would conclude that, based on the observed data, there is no evidence against the hypothesis $H_0 : \mu = 45$. (Note: This does not imply the hypothesis is true!).

A $100p\%$ confidence interval for μ based on the pivotal quantity

$$T = \frac{\bar{Y} - 45}{S/\sqrt{20}} \backsim t\left(19\right)$$

is given by

$$\left[\bar{y} - as/\sqrt{20}, \bar{y} + as/\sqrt{20}\right]$$

where $P\left(T \leq a\right) = (1+p)/2$. From t- tables we have $P\left(T \leq 2.093\right) = \left(1 + 0.95\right)/2 = 0.975$. Therefore the 95% confidence interval for μ is

$$\left[\bar{y} - 2.093s/\sqrt{20}, \bar{y} + 2.093s/\sqrt{20}\right] = [43.28, 45.53]$$

Based on these data it would appear that the new instrument is working as it should be since there was not evidence against $H_0 : \mu = 45$. We might notice

that the value $\mu = 45$ is not in the center of the 95% confidence interval but closer to the upper endpoint suggesting that the instrument might be under reading the true value of 45. It would be wise to continue testing the instrument on a regular basis on a known sample to ensure that the instrument is continuing to work well.

(d) To test $H_0 : \sigma^2 = \sigma_0^2$ we use the test statistic

$$U = \frac{(n-1)\,S^2}{\sigma_0^2} \backsim \chi^2\,(n-1)\,.$$

For $n = 20$ and $H_0 : \sigma^2 = 4$, or equivalently $H_0 : \sigma = 2$, we have

$$U = \frac{19S^2}{(4)} \backsim \chi^2\,(19)\,.$$

The observed value of U is

$$u = \frac{19\,(5.7342)}{4} = 27.24$$

and

$$\begin{aligned} p - value &= 2P\,(U \geq 27.24) \quad \text{where } U \backsim \chi^2\,(19) \\ &= 0.20 \quad \text{(calculated using } R\text{)}. \end{aligned}$$

Alternatively using the Chi-squared tables in the Course Notes we have $P\,(U \geq 27.204) = 1 - 0.9 = 0.1$ so $p - value \approx 2\,(0.1) = 0.2$. In either case, since the $p - value$ is larger than 0.1, we would conclude that there is no evidence against the hypothesis $H_0 : \sigma^2 = 4$ based on the observed data.

A $100p\%$ confidence interval for σ based on the pivotal quantity

$$U = \frac{(n-1)\,S^2}{\sigma^2} \backsim \chi^2\,(n-1)$$

is given by

$$\left[\sqrt{\frac{(n-1)\,s^2}{b}}, \sqrt{\frac{(n-1)\,s^2}{a}}\right]$$

where $P\,(U \leq a) = (1-p)\,/2 = P\,(U \geq b)$. For $n = 20$ and $p = 0.95$ we have $P\,(U \leq 8.907) = 0.025 = P\,(U \geq 32.852)$ and the confidence interval for σ is

$$\left[\sqrt{\frac{19\,(5.7342)}{32.852}}, \sqrt{\frac{9\,(5.7342)}{8.907}}\right] = [1.82, 3.50]\,.$$

Based on these data there is no evidence to contradict the manufacturer's claim that the variability in measurements is less than two parts per billion. Note however that the confidence for σ does contain values of σ larger than 2 so again it would be wise to continue testing the instrument on a regular basis on a known sample to ensure that the instrument is continuing to work well.

(e) For the new data the observed value of D is

$$d = \frac{|44.1 - 45|}{2.1/\sqrt{25}} = 2.1429$$

and

$$\begin{aligned} p - value &= P(D \geq d; H_0) \\ &= P(|T| \geq 2.1429) \quad \text{where } T \backsim t(24) \\ &= 2[1 - P(T \leq 2.1429)]. \end{aligned}$$

From t-tables $P(T \leq 2.0639) = 0.975$ and $P(T \leq 2.4922) = 0.99$ so

$$\begin{aligned} 2(1 - 0.99) \leq p - value \leq 2(1 - 0.975) \\ \text{or} \quad 0.02 \leq p - value \leq 0.05 \end{aligned}$$

and therefore there is evidence against the hypothesis $H_0 : \mu = 45$.
Since $P(T \leq 2.0639) = (1 + 0.95)/2 = 0.975$ a 95% confidence interval for μ is

$$\left[44.1 - 2.0639(2.1)/\sqrt{25}, \bar{y} + 2.0639(2.1)/\sqrt{25}\right] = [43.23, 44.97]$$

which only contains values of μ less than 45. Based on these data it would appear that the new instrument is giving measurements on average which are below the true value of 45 parts per billion and therefore the new instrument needs to be adjusted.

For these new data a statistically significant result of under measuring has been determined. The question of whether this result is of practical significance can only be answered by the people who use these results to make a decision. With many labs results decisions are made based on whether the observed measurement is within a certain interval which is considered to "safe"or not. Dioxins are poisonous to humans. Unfortunately dioxins are present in the food we eat. The 95% confidence interval suggests that the new instrument is giving results which are under reporting by $1 - 2$ parts per billion on average. What we need now is an expert on dioxin who can tell us how much a difference $1 - 2$ parts per billion makes in the context of how these results are used in the hospital lab.

(f) Here is the R code plus the output:

```
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 mean=45

One Sample t-test

data: y
```

```
t = -1.1112, df = 19, p-value = 0.2803
alternative hypothesis: true mean is not equal to 45
95 percent confidence interval:
43.28429 45.52571
sample estimates:
mean of x
44.405

> # and gives 1 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
[1] 3.31634 12.23256
> c(sqrt(s2*df/b),sqrt(s2*df/a)) # confidence interval for sigma
[1] 1.821082 3.497508
> sigma0sq<-2^2 # test hypotheis sigma=2 or sigmasq=4
> chitest<-s2*df/sigma0sq
> q<-pchisq(chitest,df)
> min(2*q,2*(1-q))
[1] 0.1984887
```

5.4 To test $H_0 : \mu = 45$ when $\sigma^2 = 4$ is known we use the discrepancy measure

$$D = \frac{\left|\bar{Y} - 45\right|}{2/\sqrt{20}} \text{ where } Z = \frac{\bar{Y} - 45}{2/\sqrt{20}} \backsim N(0,1).$$

The observed value of D is

$$d = \frac{|44.405 - 45|}{2/\sqrt{20}} = 1.33$$

and

$$\begin{aligned} p - value &= P(D \geq d; H_0) \\ &= P(|Z| \geq 1.33) \quad Z \backsim N(0,1) \\ &= 2\left[1 - P(Z \leq 1.33)\right] = 2(1 - 0.90824) \\ &= 0.18352. \end{aligned}$$

Based on these data there is no evidence to contradict the manufacturer's claim that $H_0 : \mu = 45$.

A 95% confidence interval for μ is given by

$$\begin{aligned} & \left[44.405 - 1.96\,(2)/\sqrt{20}, 44.405 + 1.96\,(2)/\sqrt{20}\right] \\ = \quad & [43.52, 45.29] \end{aligned}$$

5.5 (a) To test the hypothesis $H_0 : \mu = 105$ we use the discrepancy measure or test statistic

$$D = \frac{\left|\bar{Y} - 105\right|}{S/\sqrt{12}}$$

where

$$S = \left[\frac{1}{11}\sum_{i=1}^{12}\left(Y_i - \bar{Y}\right)^2\right]^{1/2}$$

and the t-statistic

$$T = \frac{\bar{Y} - 105}{S/\sqrt{12}} \backsim t\,(11)$$

assuming the hypothesis $H_0 : \mu = 105$ is true.
For these data $\bar{y} = 104.13$, $s^2 = 88.3115$ and $s = 9.3974$. The observed value of the discrepancy measure D is

$$d = \frac{|\bar{y} - 105|}{s/\sqrt{12}} = \frac{|104.13 - 105|}{9.3974/\sqrt{12}} = 0.3194$$

and

$$\begin{aligned} p - value &= P\,(D \geq d; H_0) \\ &= P\,(|T| \geq 0.3194) \quad \text{where } T \backsim t\,(11) \\ &= 2\,[1 - P\,(T \leq 0.3194)] = 2\,(0.3777) \\ &= 0.7554 \quad \text{(calculated using } R\text{)}. \end{aligned}$$

Alternatively using the t-tables in the Course Notes we have $P\,(T \leq 0.260) = 0.6$ and $P\,(T \leq 0.54) = 0.7$ so

$$\begin{aligned} & 2\,(1 - 0.7) \leq p - value \leq 2\,(1 - 0.6) \\ & \text{or} \quad 0.6 \leq p - value \leq 0.8. \end{aligned}$$

In either case since the $p - value$ is much larger than 0.1 and we would conclude that, based on the observed data, there is no evidence against the hypothesis $H_0 : \mu = 105$. (Note: This does not imply the hypothesis is true!)
From t- tables we have $P\,(T \leq 2.201) = (1 + 0.95)/2 = 0.975$ where $T \backsim t\,(11)$.
A 95% confidence interval for μ is

$$\left[\bar{y} - 2.201s/\sqrt{12}, \bar{y} + 2.201s/\sqrt{12}\right] = [98.16, 110.10]\,.$$

(b) From Chi-squared tables $P(W \leq 3.816) = 0.025$ and $P(W \leq 21.920) = 0.975$. A 95% confidence interval for σ is

$$\left[\sqrt{\frac{11(88.3115)}{21.920}}, \sqrt{\frac{11(88.3115)}{3.816}}\right] = [6.66, 15.96].$$

(c) Since there was no evidence against $H_0 : \mu = 105$ and since the value $\mu = 105$ is near the center of the 95% confidence interval for μ, the data support the conclusion that the detector is accurate, that is, that the detector is not giving biased readings. The confidence interval for σ, however, indicates that the precision of the detectors might be of concern. The 95% confidence interval for σ suggests that the standard deviation could be as large as 16 parts per billion. As a statistician you would need to rely on the expertise of the researchers for a decision about whether the size of the σ is scientifically significant and whether the precision of the detectors is too low. You would also point out to the researchers that this evidence is based on a fairly small sample of only 12 detectors.

5.6 To test $H_0 : \sigma^2 = \sigma_0^2$ when μ is known we use the test statistic

$$U = \frac{\sum_{i=1}^{12}(Y_i - \mu)^2}{\sigma_0^2} \backsim \chi^2(n).$$

For $n = 12$, $\mu = 105$ and $H_0 : \sigma^2 = 100$, we have

$$U = \frac{\sum_{i=1}^{12}(Y_i - 105)^2}{100} \backsim \chi^2(12).$$

Since

$$\begin{aligned}\sum_{i=1}^{12}(y_i - 105)^2 &= \sum_{i=1}^{12} y_i^2 - 2(105)\sum_{i=1}^{12} y_i + 12(105)^2 \\ &= 131096.44 - 210(1249.6) + 12(105)^2 = 980.44\end{aligned}$$

The observed value of U is

$$u = \frac{980.44}{100} = 9.8044$$

and

$$\begin{aligned}p - value &= 2P(U \leq 9.8044) \quad \text{where } U \backsim \chi^2(12) \\ &= 0.73 \quad \text{(calculated using } R\text{)}.\end{aligned}$$

Alternatively using the Chi-squared tables in the Course Notes we have $P(U \leq 9.034) = 0.3$ so $p - value > 2(0.3) = 0.6$. In either case since the $p - value$ is larger than 0.1 and we would conclude that, based on the observed data, there is no evidence against the hypothesis $H_0 : \sigma^2 = 100$.

5.7 (a) The respondents to the survey are students who heard about the online referendum and then decided to vote. These students may not be representative of all students at the University of Waterloo. For example, it is possible that the students who took the time to vote are also the students who most want a fall study break. Students who don't care about a fall study break probably did not bother to vote. This is an example of sample error. Any online survey such as this online referendum has the disadvantage that the sample of people who choose to vote are not necessarily a representative sample of the study population of interest. The advantage of online surveys is that they are inexpensive and easy to conduct. To obtain a representative sample you would need to select a random sample of all students at the University of Waterloo. Unfortunately taking such a sample would be much more time consuming and costly then conducting an online referendum.

(b) A suitable target population would be the $30,990$ eligible voters. This would also be the study population. Note that all undergraduates were able to vote but it is not clear how the list of undergraduates is determined.

(c) The attribute of interest is the proportion of the $30,990$ eligible voters (the study population) who would respond yes to the question. The parameter θ in the Binomial model corresponds to this attribute. A Binomial model assumes independent trials (students) which might not be a valid assumption. For example, if groups of students, say within a specific faculty, all got together and voted, their responses may not be independent events.

(d) The maximum likelihood estimate of θ based on the observed data is

$$\hat{\theta} = \frac{4440}{6000} = 0.74.$$

Since this estimate is not based on a random sample it is not possible to say how accurate this estimate is.

(e) An approximate 95% confidence interval for θ is given by

$$0.74 \pm 1.96\sqrt{\frac{0.74\,(0.26)}{6000}} = 0.74 \pm 0.01 = [0.73, 0.75]$$

(f) Since $\theta = 0.7$ is not a value contained in the approximate 95% confidence interval $[0.73, 0.75]$ for θ, therefore the approximate $p - value$ for testing $H_0 : \theta = 0.7$ is less than 0.05. (Note that since $\theta = 0.7$ is far outside the interval, the $p - value$ would be much smaller than 0.05.)

5.8 (a) If $H_0 : \theta = 3$ is true then since Y_i has a Poisson distribution with mean 3, $i = 1, 2, \ldots, 25$ independently, then $\sum_{i=1}^{25} Y_i$ has a Poisson distribution with mean

$3 \times 25 = 75$. The discrepancy measure

$$D = \left|\sum_{i=1}^{25} Y_i - 75\right| = \left|\sum_{i=1}^{25} Y_i - E\left(\sum_{i=1}^{25} Y_i\right)\right|$$

is reasonable since it is measuring the agreement between the data and $H_0 : \theta = 3$ by using the distance between the observed value of $\sum_{i=1}^{25} Y_i$ and its expected value $E\left(\sum_{i=1}^{25} Y_i\right) = 75$.

For the given data, $\sum_{i=1}^{25} y_i = 51$. The observed value of the discrepancy measure is

$$d = \left|\sum_{i=1}^{25} y_i - 75\right| = |51 - 75| = 24$$

and

$$\begin{aligned} p - value &= P\left(D \geq d; H_0\right) \\ &= P\left(\left|\sum_{i=1}^{25} Y_i - 75\right| \geq 24; H_0\right) \\ &= \sum_{x=0}^{51} \frac{75^x e^{-75}}{x!} + \sum_{x=99}^{\infty} \frac{75^x e^{-75}}{x!} \\ &= 1 - \sum_{x=52}^{98} \frac{75^x e^{-75}}{x!} \\ &= 0.006716 \quad \text{(calculated using } R\text{)}. \end{aligned}$$

Since $0.001 < 0.006716 < 0.01$ we would conclude that, based on the data, there is strong evidence against the hypothesis $H_0 : \theta = 3$.

(b) If Y_i has a Poisson distribution with mean θ and variance θ, $i = 1, 2, \ldots, n$ independently then by the Central Limit Theorem

$$\frac{\bar{Y} - E\left(\bar{Y}\right)}{\sqrt{Var\left(\bar{Y}\right)}} = \frac{\bar{Y} - \theta}{\sqrt{\theta/n}}$$

has approximately a $N\left(0, 1\right)$ distribution.

(c) If $H_0 : \theta = 3$ is true then $E\left(\bar{Y}\right) = 3$. The discrepancy measure $D = \left|\bar{Y} - 3\right|$ is reasonable for testing $H_0 : \theta = 3$ since it is measuring the agreement between the data and $H_0 : \theta = 3$ by using the distance between the observed value of $\bar{Y}$ and its expected value $E\left(\bar{Y}\right) = 3$.

The observed value of the discrepancy measure is

$$d = |\bar{y} - 3| = \left|\frac{51}{25} - 3\right| = |2.04 - 3| = 0.96$$

and also

$$\frac{|\bar{y}-3|}{\sqrt{3/25}} = \frac{0.96}{\sqrt{3/25}} = 2.77.$$

Therefore

$$\begin{aligned} p-value &= P\left(D \geq d; H_0\right) \\ &= P\left(\left|\bar{Y}-3\right| \geq 0.96; H_0\right) \\ &\approx P\left(|Z| \geq \frac{0.96}{\sqrt{3/25}}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1-P\left(Z \leq 2.77\right)\right] = 0.005584 \end{aligned}$$

The approximate p-value of 0.005584 is close to the $p-value$ calculated in (a) which is the exact $p-value$. Since we are only interested in whether the $p-value$ is bigger than 0.1 or between 0.1 and 0.05 etc. we are not as worried about how good the approximation is. In this example the conclusion about H_0 is the same for the approximate $p-value$ as it is for the exact $p-value$.

5.9 The observed value of the likelihood ratio test statistic for testing $H_0 : \theta = 3$ is

$$\begin{aligned} \lambda(3) &= -2\log R(3) = -2\log\left[\left(\frac{3}{2.04}\right)^{51} e^{25(2.04-3)}\right] \\ &= -2\log(0.01315) = 8.6624 \end{aligned}$$

and

$$\begin{aligned} p-value &= P\left(\Lambda(3) \geq 8.6624; H_0\right) \\ &\approx P\left(W \geq 8.6624\right) \quad \text{where } W \sim \chi^2(1) \\ &= P\left(|Z| \geq \sqrt{8.6624}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1-P\left(Z \leq 2.94\right)\right] = 0.00328 \end{aligned}$$

The $p-value$ is close to the $p-values$ calculated in (a) and (b).

5.10 Since

$$R(\theta) = \left[\frac{3.6}{\theta} e^{(1-3.6/\theta)}\right]^{20} \quad \theta > 0.$$

then

$$R(5) = \left[\frac{3.6}{5} e^{(1-3.6/5)}\right]^{20} = 0.3791$$

and

$$\lambda(5) = -2\log R(5) = -2\log(0.3791) = 1.9402.$$

Therefore

$$\begin{aligned} p-value &\approx P(W \geq 1.9402) \quad \text{where} \quad W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{1.9402}\right)\right] \quad \text{where} \quad Z \backsim N(0,1) \\ &= 2[1 - P(Z \leq 1.39)] = 2(1 - 0.91774) = 0.16452 \end{aligned}$$

and since $p - value > 0.1$ there is no evidence, based on the data, to contradict $H_0 : \theta = 5$. The approximate 95% confidence interval for θ is $[2.40, 5.76]$ which contains the value $\theta = 5$. This also implies that the $p - value > 0.05$ and so the approximate confidence interval is consistent with the test of hypothesis.

5.11 Since

$$r(\theta) = 15\log[2.3(\theta+1)] - 34.5(\theta+1) + 15 \quad \text{for } \theta > -1$$

then

$$r(-0.1) = 15\log[2.3(-0.1+1)] - 34.5(-0.1+1) + 15 = -5.1368$$

and

$$\lambda(5) = -2r(-0.1) = -2(-5.1368) = 10.2735$$

Therefore

$$\begin{aligned} p-value &\approx P(W \geq 10.2735) \quad \text{where} \quad W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{10.2735}\right)\right] \quad \text{where} \quad Z \backsim N(0,1) \\ &= 2[1 - P(Z \leq 3.21)] = 2(1 - 0.99934) = 0.00132 \end{aligned}$$

and since $0.001 < p - value < 0.01$ there is strong evidence, based on the data, to contradict $H_0 : \theta = -0.1$. The approximate 95% confidence interval for θ is $[-0.75, -0.31]$ which does not contain the value $\theta = -0.1$. This also implies that the $p - value < 0.05$ and so the approximate confidence interval is consistent with the test of hypothesis.

5.12 Since

$$R(\theta) = \frac{\theta^{16}(1-\theta)^{66}}{(8/41)^{16}(33/41)^{66}} \quad \text{for } 0 < \theta \leq \frac{1}{2}.$$

then

$$R(0.18) = \frac{(0.18)^{16}(1-0.18)^{66}}{(8/41)^{16}(33/41)^{66}} = 0.9397$$

and

$$\lambda(0.18) = -2\log R(0.18) = -2\log(0.9397) = 0.1244$$

Therefore

$$\begin{aligned} p-value &\approx P(W \geq 0.1244) \quad \text{where} \quad W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.1244}\right)\right] \quad \text{where} \quad Z \backsim N(0,1) \\ &= 2[1 - P(Z \leq 0.35)] = 2(1 - 0.63683) = 0.72634 \end{aligned}$$

and since $p - value > 0.1$ there is no evidence, based on the data, to contradict $H_0 : \theta = 0.18$. The approximate 95% confidence interval for θ is $[0.12, 0.29]$ which contains the value $\theta = 0.18$. This also implies that the $p - value > 0.05$ and so the approximate confidence interval is consistent with the test of hypothesis.

5.13 (a) The maximum likelihood estimate of θ is $\hat{\theta} = 18698.6/20 = 934.93$. The agreement between the plot of the empirical cumulative distribution function and the cumulative distribution function of an Exponential(934.93) random variable given in Figure 12.22 indicates that the Exponential is reasonable.

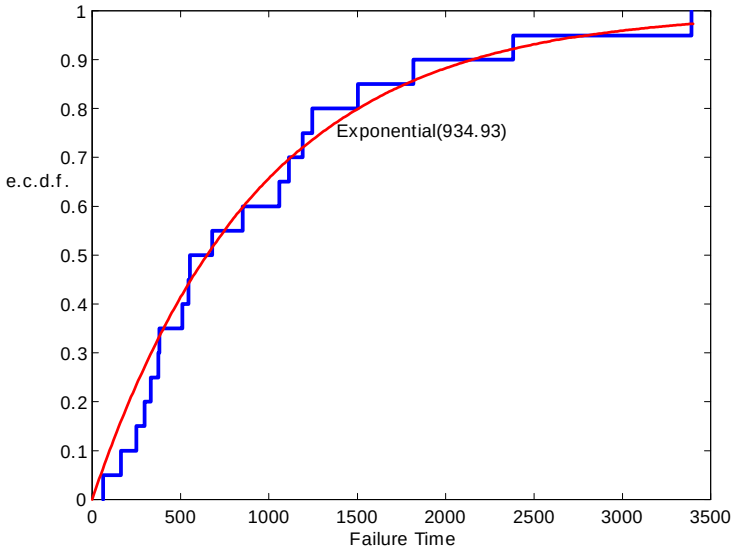


Figure 12.22: Empirical c.d.f. and Exponential(934.93) c.d.f. for failure times of power systems

(b) The observed value of the likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ for Exponential data is (see Example 5.3.2) is

$$\lambda(\theta_0) = -2\log R(\theta_0) = -2\log\left[\left(\frac{\hat{\theta}}{\theta_0}\right)^n e^{n\left(1-\hat{\theta}/\theta_0\right)}\right]$$

where $\hat{\theta} = \bar{y}$. For $n = 20$, $\hat{\theta} = \bar{y} = 934.93$ and $\theta_0 = 1000$ we have

$$\lambda(1000) = -2\log\left[\left(\frac{934.93}{1000}\right)^{20} e^{20(1-934.93/1000)}\right] = 2\log(0.95669) = 0.0885$$

with

$$\begin{aligned} p-value &\approx P\left(W \geq 0.0885\right) \quad \text{where } W \backsim \chi^2\left(1\right) \\ &= 2\left[1-P\left(Z \leq \sqrt{0.0885}\right)\right] \quad \text{where } \ Z \backsim N\left(0,1\right) \\ &= 2\left[1-P\left(Z \leq 0.30\right)\right] = 2\left(1-0.61791\right) = 0.76418. \end{aligned}$$

There is no evidence against the hypothesis $H_0 : \theta = 1000$ based on the observed data.

5.14 A test statistic that could be used will be to test the mean of the generated sample. The mean should be closed to 0.5 if the random number generator is working well.

5.15 (a) For each given region the assumptions of independence, individuality and homogeneity would need to hold for the number of events per person per year.

(b) Assume the observations $y_1, y_2, \ldots, y_K$ from the different regions are independent. Since $Y_j \backsim \text{Poisson}(P_j\theta_j t)$ then the likelihood function for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is

$$L\left(\boldsymbol{\theta}\right) = \prod_{j=1}^{K} \frac{\left(P_j\theta_j t\right)^{y_j} e^{-P_j\theta_j t}}{y_j!}$$

or more simply

$$L\left(\boldsymbol{\theta}\right) = \prod_{j=1}^{K} \theta_j^{y_j} e^{-P_j\theta_j t}$$

and the log likelihood function is

$$l\left(\boldsymbol{\theta}\right) = \sum_{j=1}^{K} \left[y_j \log \theta_j - P_j\theta_j t\right].$$

Since

$$\frac{\partial l}{\partial \theta_j} = \frac{y_j}{\theta_j} - P_j t = \frac{y_j - \left(P_j t\right)\theta_j}{\theta_j} = 0$$

for $\theta_j = y_j/\left(P_j t\right)$, the maximum likelihood estimate of θ_j is $\hat{\theta}_j = y_j/\left(P_j t\right)$, $j = 1, 2, \ldots, K$. So

$$\begin{aligned} l(\hat{\boldsymbol{\theta}}) &= \sum_{j=1}^{K} \left[y_j \log \hat{\theta}_j - P_j\hat{\theta}_j t\right] = \sum_{j=1}^{K} \left[y_j \log\left(\frac{y_j}{P_j t}\right) - y_j\right] \\ &= \sum_{j=1}^{K} y_j \left[\log\left(\frac{y_j}{P_j t}\right) - 1\right]. \end{aligned}$$

The likelihood function assuming $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ is given by

$$L\left(\theta\right) = \prod_{j=1}^{K} \theta^{y_j} e^{-P_j\theta t}$$

with log likelihood function

$$l(\theta) = \left(\sum_{j=1}^{K} y_j\right) \log\theta - \theta t \sum_{j=1}^{K} P_j.$$

Since

$$l'(\theta) = \frac{1}{\theta}\sum_{j=1}^{K} y_j - \sum_{j=1}^{K} P_j t = \frac{1}{\theta}\left[\sum_{j=1}^{K} y_j - \theta t \sum_{j=1}^{K} P_j\right] = 0$$

if $\theta = \sum_{j=1}^{K} y_j / \sum_{j=1}^{K} P_j t$, the maximum likelihood estimate of θ assuming $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ is $\hat{\theta}_0 = \sum_{j=1}^{K} y_j / t \sum_{j=1}^{K} P_j$. So

$$\begin{aligned} l(\hat{\theta}_0) &= \left(\sum_{j=1}^{K} y_j\right) \log\hat{\theta}_0 - \hat{\theta}_0 \sum_{j=1}^{K} P_j t \\ &= \left(\sum_{j=1}^{K} y_j\right) \log\left(\frac{\sum_{j=1}^{K} y_j}{t\sum_{j=1}^{K} P_j}\right) - \left(\frac{\sum_{j=1}^{K} y_j}{t\sum_{j=1}^{K} P_j}\right) \sum_{j=1}^{K} P_j t \\ &= \left(\sum_{j=1}^{K} y_j\right)\left[\log\left(\sum_{j=1}^{K} y_j / t\sum_{j=1}^{K} P_j\right) - 1\right]. \end{aligned}$$

The likelihood ratio test statistic for testing $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ is

$$\begin{aligned} \Lambda &= 2l(\tilde{\boldsymbol{\theta}}) - 2l(\tilde{\theta}_0) \\ &= 2\sum_{j=1}^{K} Y_j\left[\log\left(\frac{Y_j}{P_j t}\right) - 1\right] - 2\left(\sum_{j=1}^{K} Y_j\right)\left[\log\left(\sum_{j=1}^{K} Y_j / t\sum_{j=1}^{K} P_j\right) - 1\right]. \end{aligned}$$

The observed value of Λ is

$$\begin{aligned} \lambda &= 2l(\hat{\boldsymbol{\theta}}) - 2l(\hat{\theta}_0) \\ &= 2\sum_{j=1}^{K} y_j\left[\log\left(\frac{y_j}{P_j t}\right) - 1\right] - 2\left(\sum_{j=1}^{K} y_j\right)\left[\log\left(\sum_{j=1}^{K} y_j / t\sum_{j=1}^{K} P_j\right) - 1\right]. \end{aligned}$$

The $p - value$ is

$$P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda) \quad \text{where} \quad W \backsim \chi^2(K-1).$$

(c) For the given data

$$\hat{\boldsymbol{\theta}} = \left(\frac{27}{5(2025)}, \frac{18}{5(1116)}, \frac{41}{5(3210)}, \frac{29}{5(1687)}, \frac{31}{5(2840)}\right) \quad \text{and} \quad \hat{\theta}_0 = \frac{146}{5(10878)}$$

$\lambda = 3.73$ and $p - value \approx P(W \geq 3.73) = 0.44$ where $W \backsim \chi^2(4)$. There is no evidence based on the data against $H_0 : \theta_1 = \theta_2 = \cdots = \theta_5$, that is, that the rates are equal.

5.16 (a) $\tilde{\mu} = \overline{Y}$, $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})$, $\hat{\mu}_0 = \mu_0$, $\tilde{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)$ and $\Lambda(\mu_0) = n\log\left(\tilde{\sigma}_0^2/\tilde{\sigma}^2\right)$.

$$\frac{\tilde{\sigma}_0^2}{\tilde{\sigma}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}) + n(\overline{Y} - \mu_0)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})}$$

so that

$$\Lambda(\mu_0) = n\log\left[1 + \frac{n(\overline{Y} - \mu_0)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})}\right] = n\log\left(1 + \frac{T^2}{n-1}\right)$$

Chapter 6

6.1. (a) The maximum likelihood estimates of α and β are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2325.20}{2802.00} = 0.83, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 133.56 - (0.83)(43.20) = 97.71$$

and an unbiased estimate of σ^2 is

$$s_e^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy}) = \frac{1}{23}\left[3284.16 - (0.83)(2325.20)\right] = 58.8968.$$

(b) The scatterplot with fitted line and the residual plots shown in Figure 12.23 show no unusual patterns. The model fits the data well.

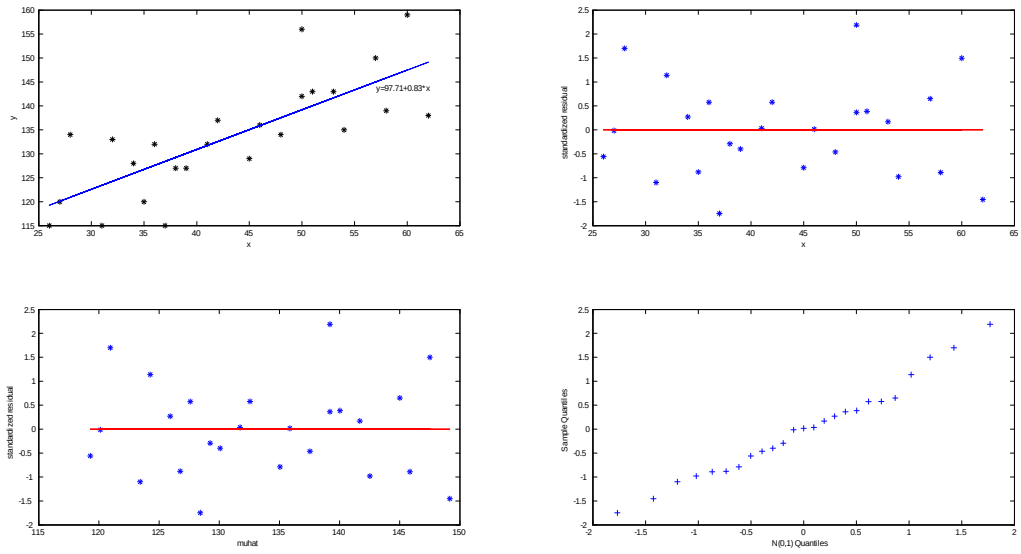


Figure 12.23: Scatterplot and residual plots for nurses data

(c) Since $P(T \leq 2.0687) = 0.975$ where $T \backsim t(23)$ and

$$s_e = \left[\frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})\right]^{1/2} = 7.6744$$

therefore a 95% confidence interval for β is

$$\hat{\beta} \pm 2.0687\,(7.6744)\,/\sqrt{2802.0} = 0.8298 \pm 0.2999 = [0.53, 1.13].$$

Meaning: Suppose we repeat the experiment (select 25 female nurses working at the large hospital at random and record their age and systolic blood pressure) a large number of times and each time we construct a 95% confidence interval for

β for the observed data. Then, approximately 95% of the constructed intervals would contain the true, but unknown value of β. We say that we are 95% confident that our interval contains the true value of β.

(d) Since $P(T \leq 1.7139) = 0.95$ where $T \backsim t\,(23)$, a 90% confidence interval for the mean systolic blood pressure of nurses aged $x = 35$ is

$$\begin{aligned} & \hat{\alpha} + \hat{\beta}(35) \pm 1.7139\,(7.6744)\left[\frac{1}{25} + \frac{(35 - 43.20)^2}{2802.00}\right]^{1/2} \\ = \quad & 126.7553 \pm 3.3274 = [123.43, 130.08]\,. \end{aligned}$$

(e) Since $P(T \leq 2.8073) = 0.995$ where $T \backsim t\,(23)$, a 99% prediction interval for the systolic blood pressure of a nurse aged $x = 50$ is

$$\begin{aligned} & \hat{\alpha} + \hat{\beta}(50) \pm 2.8073\,(7.6744)\left[1 + \frac{1}{25} + \frac{(50 - 43.20)^2}{2802.00}\right]^{1/2} \\ = \quad & 139.2029 \pm 23.0108 = [116.19, 162.21]\,. \end{aligned}$$

6.2 (a) The maximum likelihood estimate of α and β are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{22769.645}{6283.422} = 3.6238, \;\; \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 187.975 - (3.6238)\,(43.03) = 32.0444$$

The fitted line is $y = 32.04 + 3.62x$.

(b) The scatterplot with fitted line and the residual plots shown in Figure 12.24 show no unusual patterns. There is one residual value which is larger than 3 for $x = 50.3$.

(c) Since

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \;\text{ and }\; \text{sample correlation } = r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

therefore

$$r = \hat{\beta}\left(\frac{S_{xx}}{S_{yy}}\right)^{1/2} \quad \text{or} \quad \hat{\beta} = r\left(\frac{S_{yy}}{S_{xx}}\right)^{1/2}.$$

(d) Since $P(T \leq 2.1009) = 0.975$ where $T \backsim t\,(18)$ and

$$s_e = \left[\frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})\right]^{1/2} = 100.6524$$

therefore a 95% confidence interval for β is

$$\hat{\beta} \pm 2.1009\,(100.6524)\,/\sqrt{6283.422} = 3.6238 \pm 2.6677 = [0.9561, 6.2915].$$

The study population is all the actors listed at boxofficemojo.com/people/. The parameter β represents the mean increase in the amounted by a movie for a unit

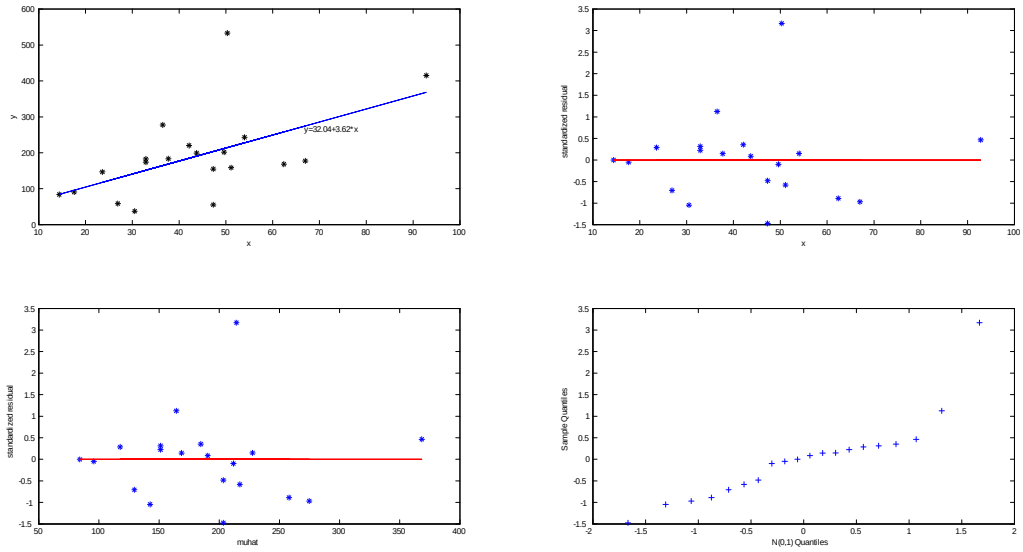


Figure 12.24: Scatterplot and residual plots for actor data

change in the value of an actor. However, since the 20 data points were obtained by taking the first 20 actors in the list, the sample is not a random sample. If actors with last names starting with letters at the beginning of the alphabet are more successful then other actors then the estimate of β might be biased.

(e) The hypothesis of no relationship is equivalent to $H_0 : \beta = 0$. Since

$$p - value = 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right)\right] = 2\left[1 - P\left(T \leq 2.85\right)\right] = 0.011$$

(using R), there is evidence based on the data against $H_0 : \beta = 0$. Note that this is consistent with the fact that the 95% confidence interval for β does not contain the value $\beta = 0$.

(f) Since $P(T \leq 2.1009) = 0.975$ where $T \backsim t\,(18)$, a 95% confidence interval for the mean amount grossed by movies for actors whose value is $x = 50$ is

$$\begin{aligned} & 32.0444 + (3.6238)\,(50) \pm 2.1009\,(100.6524)\left[\frac{1}{20} + \frac{(50 - 43.03)^2}{6283.422}\right]^{1/2} \\ = \quad & 213.2326 \pm 50.8090 = [162.4236, 264.0417]\,. \end{aligned}$$

A 95% confidence interval for the mean amount grossed by movies for actors

whose value is $x = 100$ is

$$\begin{aligned} & 32.0444 + (3.6238)(100) \pm 2.1009(100.6524)\left[\frac{1}{20} + \frac{(100-43.03)^2}{6283.422}\right]^{1/2} \\ = \; & 394.4209 \pm 159.1644 = [235.2565, 553.5853]. \end{aligned}$$

The largest observed x value is $x = 92.8$. By constructing a confidence interval for the mean amount grossed by movies for actors whose value is $x = 100$, we are assuming that the linear relationship hold beyond the observed data.

6.3 (a) Recall this was a regression of the form $E(Y_i) = \alpha + \beta x_{1i}$ where $x_{1i} = x_i^2$, and x_i = bolt diameter. Now $n = 30$, $\hat{\alpha} = 1.6668$, $\hat{\beta} = 2.8378$, $s_e = 0.05154$, $S_{xx} = 0.2244$, $\bar{x}_1 = 0.11$. A point estimate of the mean breaking strength at $x_1 = (0.35)^2 = 0.1225$ is

$$\hat{\mu}(0.1225) = \hat{\alpha} + \hat{\beta}(0.1225) = 1.667 + 2.838(0.1225) = 2.01447$$

A confidence interval for $\mu(0.1225)$ is

$$\hat{\mu}(0.1225) \pm a s_e \sqrt{\frac{1}{n} + \frac{(0.1225 - \bar{x}_1)^2}{S_{xx}}}$$

From t-tables, $P(T \leq 2.0484) = 0.975$ where $T \backsim t(28)$. The 95% confidence interval is

$$\begin{aligned} & 2.01447 \pm 2.0484(0.05154)\sqrt{\frac{1}{30} + \frac{(0.1225 - 0.11)^2}{0.2244}} \\ = \; & 2.01447 \pm 0.01932 = [1.9952, 2.0338] \end{aligned}$$

(b) A 95% prediction interval for the strength at $x_1 = (0.35)^2 = 0.1225$ is

$$\begin{aligned} & \hat{\mu}(0.1225) \pm a s_e \sqrt{1 + \frac{1}{n} + \frac{(0.1225 - \bar{x}_1)^2}{S_{xx}}} \\ = \; & 2.01447 \pm 2.0484(0.05154)\sqrt{1 + \frac{1}{30} + \frac{(0.1225 - \bar{x}_1)^2}{0.2244}} \\ = \; & 2.01447 \pm 0.10732 = [1.9072, 2.1218] \end{aligned}$$

This interval is wider since it is an interval estimate for a single observation (a random variable) at $x_1 = 0.35$ rather than an interval estimate for a mean (a constant).

(c) Since Y represents the mean strength of the bolt of diameter $x = 0.35$, then based on the assumed model $Y \backsim G(\alpha + \beta(0.1225), \sigma)$. Since α, β and σ are unknown we estimate them using $\hat{\alpha} = 1.6668$, $\hat{\beta} = 2.8378$, and $s_e = 0.05154$ and

use $Y \backsim G\left(2.01447, 0.05154\right)$. Since $V \backsim G\left(1.60, 0.10\right)$ independently of $Y \backsim G\left(2.01447, 0.05154\right)$ then $V-Y \backsim G\left(1.60-2.01447, \sqrt{(0.1)^2+(0.05154)^2}\right)$ or $V-Y \backsim G\left(-0.41447, 0.1125\right)$. Therefore an estimate of $P\left(V>Y\right)$ is

$$\begin{aligned}\hat{P}\left(V>Y\right) &= \hat{P}\left(V-Y>0\right)=P\left(Z>\frac{0-(-0.41447)}{0.1125}\right) \quad \text{where } Z \backsim G\left(0,1\right)\\ &= 1-P\left(Z\leq 3.68\right)\approx 0.\end{aligned}$$

6.4 (a)

$$\hat{\beta}=\frac{S_{xy}}{S_{xx}}=\frac{2818.556835}{2818.946855}=0.9999$$

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}=23.5505-23.7065\times 0.9999=-0.1527$$

The scatterplot with fitted line and the residual plots shown in Figure 12.25 show no unusual patterns. The model fits the data well.

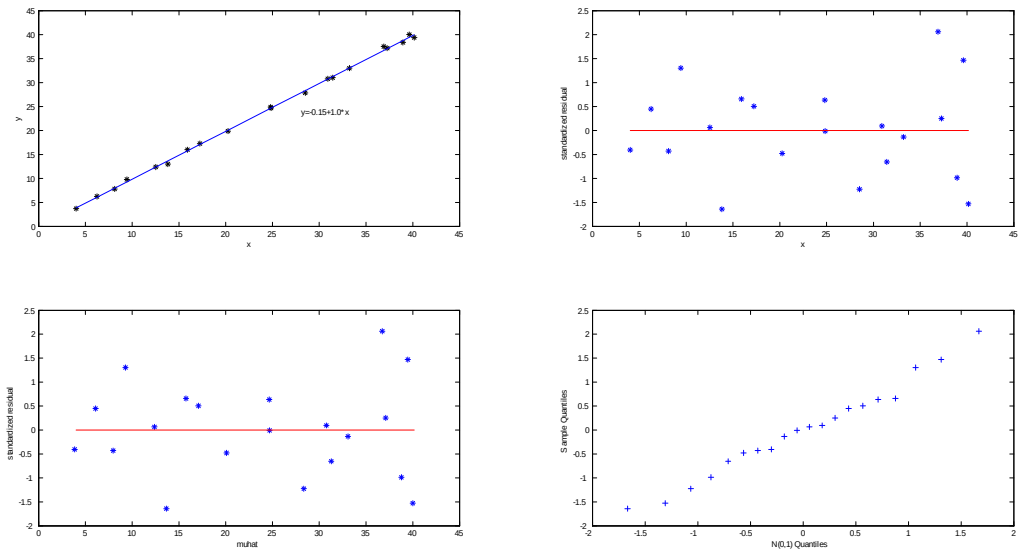


Figure 12.25: Scatterplot and residual plots for cheap versus expensive procedures

(b) Since $P(T\leq 2.1009)=0.975$ where $T \backsim t\left(18\right)$ and

$$\begin{aligned}s_e &= \left(\frac{S_{yy}-\hat{\beta}S_{xy}}{n-2}\right)^{1/2}\\ &= \left[\frac{2820.862295-\left(0.9998616\right)\left(2818.556835\right)}{18}\right]^{1/2}=0.3870\end{aligned}$$

a 95% confidence interval for β is

$$0.9999 \pm 2.1009\,(0.3870)\,/\sqrt{2818.946855} = [0.9845,\ 1.0152]\,.$$

Since the value $\beta = 1$ is inside the 95% confidence interval for β we know the $p - value$ for testing $H_0 : \beta = 1$ is greater than 0.05. Alternatively

$$p - value = 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 1\right|}{s_e/\sqrt{S_{xx}}}\right)\right] = 2\left[1 - P\left(T \leq 0.019\right)\right] = 0.99$$

using R and there is no evidence based on the data against $H_0 : \beta = 1$.
A 95% confidence interval for α is

$$-0.1527 \pm 2.1009(0.3870)\sqrt{\frac{1}{20} + \frac{(0 - 23.7065)^2}{2818.946855}} = [-0.5587,\ 0.2533]$$

Since $\alpha = 0$ is inside the 95% confidence interval for α we know the $p - value$ for testing $H_0 : \alpha = 0$ is greater than 0.05. Alternatively

$$p{-}value = 2\left[1 - P\left(T \leq \frac{|\hat{\alpha} - 0|}{s_e\sqrt{\frac{1}{n} + \frac{(0-\bar{x})^2}{S_{xx}}}}\right)\right] = 2\left[1 - P\left(T \leq 0.7903\right)\right] = 0.4396$$

using R and there is no evidence based on the data against $H_0 : \alpha = 0$.
The question of interest is how well the cheaper way of determining concentrations compares with the more expensive way. To put this question in terms of the model we first note that the assumed model is

$$Y_i \sim G(\alpha + \beta x_i, \sigma) \quad i = 1, 2, \ldots, n \ \text{ independently.}$$

If the cheaper way worked perfectly then the measurements using the cheaper way would be identical to the more expensive way plus some variability. That is, the model would be

$$Y_i \sim G(x_i, \sigma) \quad i = 1, 2, \ldots, n \ \text{ independently.}$$

This means we are interested in whether the model with $\beta = 1$ and $\alpha = 0$ fits the data well. This is the reason why we test the hypotheses $H_0 : \beta = 1$ and $H_0 : \alpha = 0$.

(c) The scatterplot plus the fitted line indicates good agreement between the cheaper way of determining concentrations and the more expensive way. The points lie quite close to the fitted line. The data suggest that the cheaper way of determining concentrations is quite accurate since the cheaper way does not appear to consistently give values which are systematically above (or below) the concentration determined by the more expensive way.

(d) Since the fitted model is

$$y = -0.1527 + 0.9999x,$$

the point estimate of the y-intercept is $\hat{\alpha} = -0.1527$ which is slightly negative which suggests the cheaper way is giving values lower than the true concentration as determined by the more expensive way. However, the confidence interval for α was $[-0.5587, 0.2533]$ which certainly includes the value $\alpha = 0$ as well as values of α above and below zero. The data do not suggest the cheaper way is giving lower values. If the confidence interval only contained negative values then this would suggest that the cheaper way is giving lower values.

6.5 (a) The likelihood function is for β is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right] \quad \text{for } \beta \in \Re$$

or more simply.

$$L(\beta) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right] \quad \text{for } \beta \in \Re$$

The log likelihood function is

$$l(\beta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2 \quad \text{for } \beta \in \Re.$$

Maximizing $l(\beta)$ is equivalent to minimizing $g(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2$ which is the criterion for determining the least squares estimate of β.
Solving

$$l'(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0$$

we obtain both the maximum likelihood estimate and the least squares estimate of β given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(b) Note that

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n}\left(\frac{x_i}{\sum_{i=1}^{n} x_i^2}\right) Y_i = \sum_{i=1}^{n} a_i Y_i \quad \text{where} \quad a_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

so $\tilde{\beta}$ is a linear combination of independent Normal random variables and therefore has a Normal distribution. Since

$$E(\tilde{\beta}) = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2} \right) E(Y_i) = \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} (x_i)(\beta x_i) = \frac{\beta}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i^2 = \beta.$$

and

$$Var(\tilde{\beta}) = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)^2 Var(Y_i) = \frac{1}{\left[\sum_{i=1}^{n} x_i^2 \right]^2} \sum_{i=1}^{n} x_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

therefore

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \backsim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \right).$$

(c)

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2 &= \sum_{i=1}^{n} (y_i^2 - 2 x_i y_i \hat{\beta} + x_i^2 \hat{\beta}^2) \\
&= \sum_{i=1}^{n} y_i^2 - 2 \underbrace{\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}_{\hat{\beta}} \sum_{i=1}^{n} x_i y_i + \underbrace{\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right]^2}_{\hat{\beta}^2} \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} y_i^2 - 2 \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2} + \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2} \\
&= \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

as required.

(d) Find a in the t-table such that $P(-a \leq T \leq a) = 0.95$ where $T \backsim t(n-1)$. Then since

$$\begin{aligned}
0.95 &= P\left(-a \leq \frac{\tilde{\beta} - \beta}{S_e / \sqrt{\sum_{i=1}^{n} x_i^2}} \leq a \right) \\
&= P\left(\tilde{\beta} - a S_e / \sqrt{\sum_{i=1}^{n} x_i^2} \leq \beta \leq \tilde{\beta} + a S_e / \sqrt{\sum_{i=1}^{n} x_i^2} \right)
\end{aligned}$$

a 95% confidence interval for β is given by $\left(\hat{\beta} - a \frac{s_e}{\sqrt{\sum_{i=1}^{n} x_i^2}}, \hat{\beta} + a \frac{s_e}{\sqrt{\sum_{i=1}^{n} x_i^2}} \right)$.

(e) Define the discrepancy measure

$$D = \frac{|\tilde{\beta} - \beta|}{S_e / \sqrt{\sum_{i=1}^{n} x_i^2}}.$$

Under the null hypothesis $H_0 : \beta = \beta_0$, the $p - value$ is given by

$$P\left(|T| > \frac{|\hat{\beta} - \beta_0|}{s_e / \sqrt{\sum_{i=1}^{n} x_i^2}} \right) = 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - \beta_0|}{s_e / \sqrt{\sum_{i=1}^{n} x_i^2}} \right) \right]$$

where $T \sim t(n-1)$.

6.6 (a)

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{13984.5554}{14058.9097} = 0.9947$$

and the fitted model is $y = 0.9947x$.

(b) A scatterplot of the data as well as the fitted line are given in top left panel of Figure 12.26. The straight line fits the data very well. The observed points all lie very close to the fitted line.

(c) Since $P(T \leq 2.0930) = 0.975$ where $T \backsim t(19)$ and

$$\begin{aligned} s_e &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i \right)^2}{\sum_{i=1}^{n} x_i^2} \right) \\ &= \left[\frac{1}{19}\left(13913.3833 - \frac{13984.5554^2}{14058.9097} \right) \right]^{1/2} = 0.3831 \end{aligned}$$

a 95% confidence interval for β is given by

$$\hat{\beta} \pm a s_e / \sqrt{\sum_{i=1}^{n} x_i^2} = 0.9947 \pm 2.0930(0.3831)/\sqrt{14058.9097} = [0.9879,\ 1.0015].$$

For testing $H_0 : \beta = 1$ we have

$$p - value = 2\left[1 - P\left(T \leq \frac{|0.9947 - 1|}{0.3831/\sqrt{14058.9097}} \right) \right] = 2[1 - P(T \leq 1.64)] = 0.12$$

where $T \sim t(19)$, and there is no evidence based on the data against $H_0 : \beta = 1$.

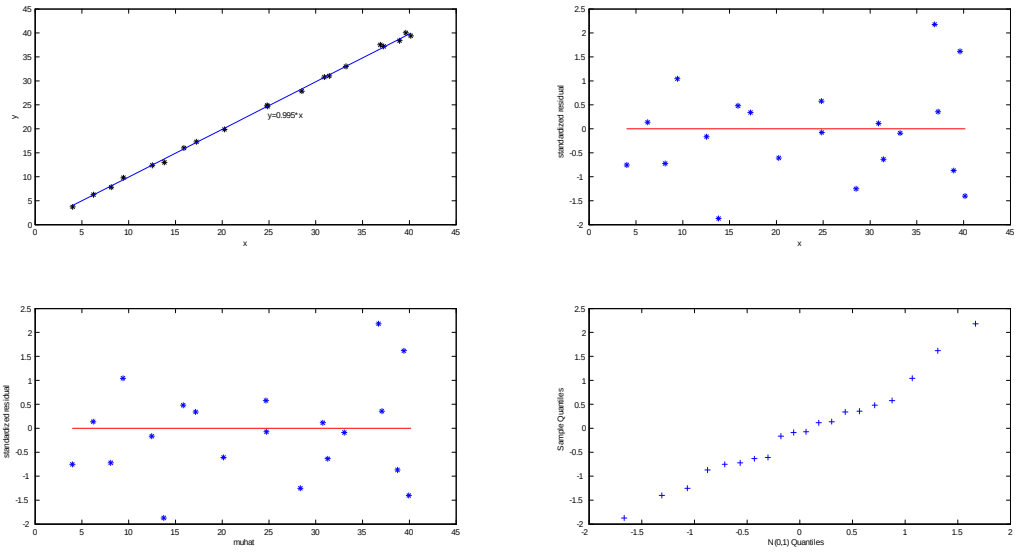


Figure 12.26: Scatterplot and residual plots for model through the origin

(d) The scatterplot with fitted line and the residual plots shown in Figure 12.26 show no unusual patterns. The model fits the data well.

(e) Based on this analysis we would conclude that the simple model $Y \sim G(\beta x_i, \sigma)$ is an adequate model for these data as compared to the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$.

6.7 (a) The maximum likelihood estimates of α and β are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{6175}{2155.2} = 2.8652 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 30.3869.$$

and the fitted line is $y = 30.3869 + 2.8652x$.

(b) The scatterplot with fitted line and the residual plots shown in Figure 12.27. There are a few large negative residuals but overall the model seems reasonable.

(c) An estimate of σ is

$$s_e = \left[\frac{S_{yy} - \hat{\beta}S_{xy}}{n-2}\right]^{1/2} = \left[\frac{24801.1521 - (2.8652)\,(6175.1522)}{44}\right]^{1/2} = 12.7096$$

The hypothesis of no relationship is equivalent to $H_0 : \beta = 0$. Since

$$p - value = 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}}\right)\right] = 2\left[1 - P\left(T \leq 10.47\right)\right] \approx 0$$

there is very strong evidence based on the data against $H_0 : \beta = 0$.

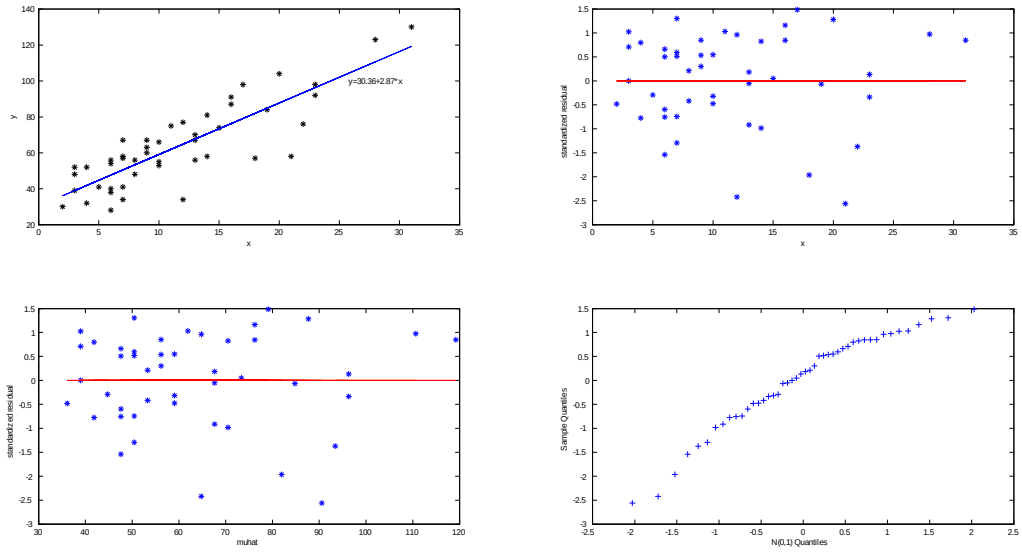


Figure 12.27: Scatterplot and residual plots for death rate due to cirrhosis of the liver versus wine consumption

(d) Since $P(T \leq 2.0154) = 0.975$ where $T \backsim t\,(44)$ a 95% confidence interval for β is

$$2.8652 \pm 2.0154\,(12.7096)\,/\sqrt{2155.1522} = [2.3135, 3.4171]\,.$$

6.8 (a) The command summary(RegModel) gives the output:

```
>Call:
>lm(formula = BodyDensity ~Skinfold, data = Dataset)
>Residuals:
>Min             1Q         Median      3Q       Max
>-0.0251400 -0.0040412 -0.0001752 0.0041324 0.0192336
>Coefficients:
>             Estimate  Std.  Error  t value   Pr(>|t|)
>(Intercept)  1.161139  0.005429    213.90   <2e-16 ***
>Skinfold    -0.062066  0.003353    -18.51   <2e-16 ***
>Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
>Residual standard error:  0.007877 on 90 degrees of freedom
>Multiple R-squared:  0.7919,    Adjusted R-squared:  0.7896
>F-statistic:  342.6 on 1 and 90 DF, p-value:  < 2.2e-16
```

The fitted line is $y = 1.161139 - 0.062066x$ where $y =$ BodyDensity and $x =$ Skinfold.

(b) For the hypothesis of no relationship the value of the test statistic is 213.90 and the $p-value$ is 2×10^{-16}. Since $p-value \approx 0$ there is very strong evidence against the hypothesis of no relationship.

(c) An estimate of σ is $se = 0.007877322$.

(d) The plots are given in Figure 12.28. The scatterplot and residual plots indicate that the model fits the data well.

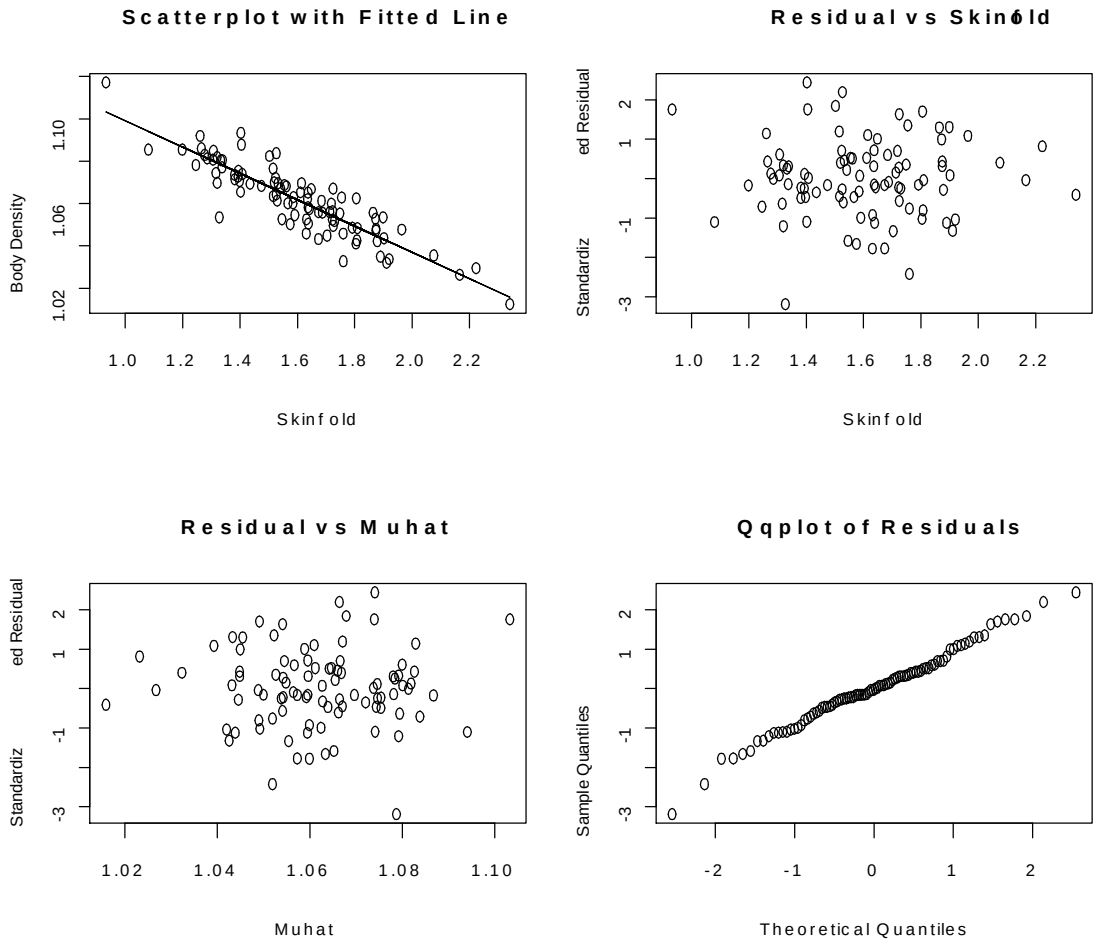


Figure 12.28: Fitted Line and Residual Plots for Skinfold Data

(e) The 95% confidence interval for β is $[-0.06872823, -0.05540425]$.

(f) A 95% prediction interval for the body density of a male with skinfold measurement of $x = 1.8$ is $[1.033629, 1.065211]$.

(g) Skinfold measurements seem to provide a reasonable approximation to body density measurements. However we notice that the range of body density mea-

surements is $[1.0126, 1.1171]$ with a width of 0.1045 and that the 95% prediction interval for the body density of a male with skinfold measurement of $x = 1.8$ has width 0.031582 which is approximately one third the width of the range of measurements. There is a fair bit of uncertainty in approximating the body density using the skinfold measurement. The decision to use the approximation or not would depend on issues such as what the body density measurement is to be used for and how accurate it needs to be and how much more cost and effort is required to measure body density directly. Note that an accurate body density measurement is usually done by weighing a person under water.

6.9 (a)

$$\begin{gathered}\bar{x} = 191.7871 \qquad \bar{y} = 20.0276 \\ S_{xx} = 2291.3148 \quad S_{yy} = 447.8497 \quad S_{xy} = 1008.8246\end{gathered}$$

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{1008.8246}{2291.3148} = 0.44028 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 20.0276 - (0.44028)(191.7871) = -64.4128\end{aligned}$$

The fitted line is $y = -64.4128 + 0.44028x$. The scatterplot and residual plot are given in the top two panels of Figure 12.29. Both graphs show a distinctive pattern. In the scatterplot as x increases the points lie above the line, then below then above. Correspondingly in the residual plot as x increases the residuals are positive then negative then positive. In the residual plot the points do not lie in a horizontal about the line $\hat{r}_i = 0$ which suggests that the linear model is not adequate.

(b)

$$\begin{gathered}\bar{x} = 191.7871 \qquad \bar{y} = 2.9804 \\ S_{xx} = 2291.3148 \quad S_{yy} = 1.00001 \quad S_{xy} = 47.81920\end{gathered}$$

$$\begin{aligned}\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{47.81920}{2291.3148} = 0.02087 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 2.9804 - (0.02087)(191.7871) = -1.02214\end{aligned}$$

The fitted line is $z = -1.02214 + 0.02087x$. The scatterplot and residual plots are given in the bottom two panels of Figure 12.29. In both of these plots we do not observe any unusual patterns. There is no evidence to contradict the linear model for log(pressure) versus temperature. However this does not "prove" that the theory's model is correct - only that there is no evidence to disprove it.

(c) Since $P(T \leq 2.0452) = 0.975$ where $T \backsim t(29)$, and

$$\begin{aligned}s_e &= \left(\frac{S_{yy} - \hat{\beta}S_{xy}}{n-2}\right)^{1/2} \\ &= \left[\frac{1.00001 - (0.02087)(47.81920)}{29}\right]^{1/2} = 0.00838894\end{aligned}$$

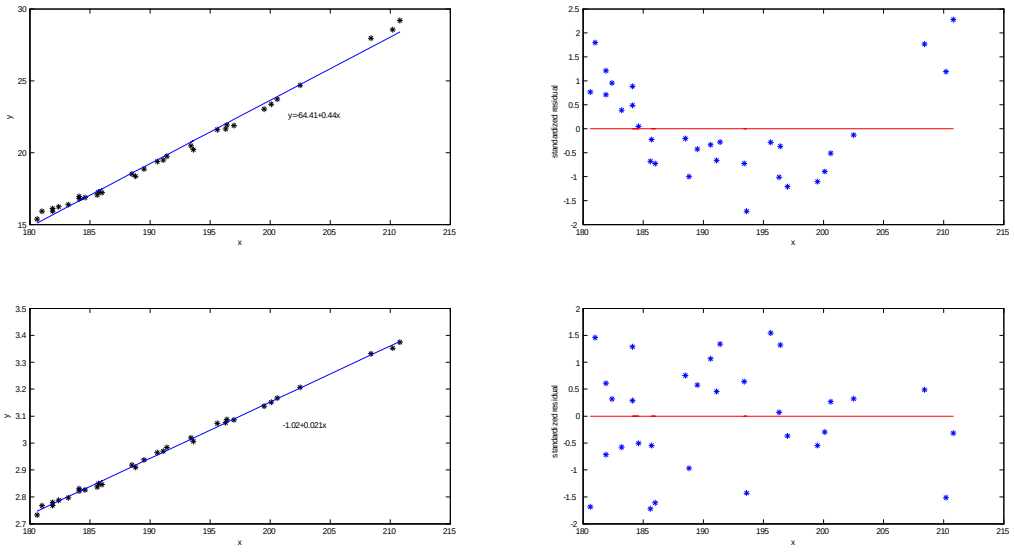


Figure 12.29: Fitted lines and residual plots for atmospheric pressure data

a 95% confidence interval for the mean log atmospheric pressure at a temperature of $x = 195$ is

$$-1.02214 + (0.02087)(195) \pm 2.0452\,(0.008389)\left[\frac{1}{31} + \frac{(100 - 191.7871)^2}{2291.3148}\right]^{1/2}$$

$$= \quad 3.04747 \pm 0.00329 = [3.04418, 3.05076]\,.$$

which implies a 95% confidence interval for the mean atmospheric pressure at a temperature of $x = 195$ is

$$[\exp(3.04418), \exp(3.05076)] = [20.9927, 21.1313]\,.$$

6.10 (a) We assume that the study population is the set of all Grade 3 students who are being taught the same curriculum. (For example in Ontario all Grade 3 students must be taught the same Grade 3 curriculum set out by the Ontario Government.) The parameter μ_1 represents the mean score on the DRP test if all Grade 3 students in the study population took part in the new directed readings activities for an 8-week period.
The parameter μ_2 represents the mean score on the DRP test for all Grade 3 students in the study population without the directed readings activities.
The parameter σ represents the standard deviation of the DRP scores for all Grade 3 students in the study population which is assumed to be the same whether the students take part in the new directed readings activities or not.

(b) The qqplot of the responses for the treatment group and the qqplot of the responses for the control group are given in Figures 12.30 and 12.31. Looking at these plots we see that the points lie reasonably along a straight line in both plots and so we would conclude that the normality assumptions seem reasonable.

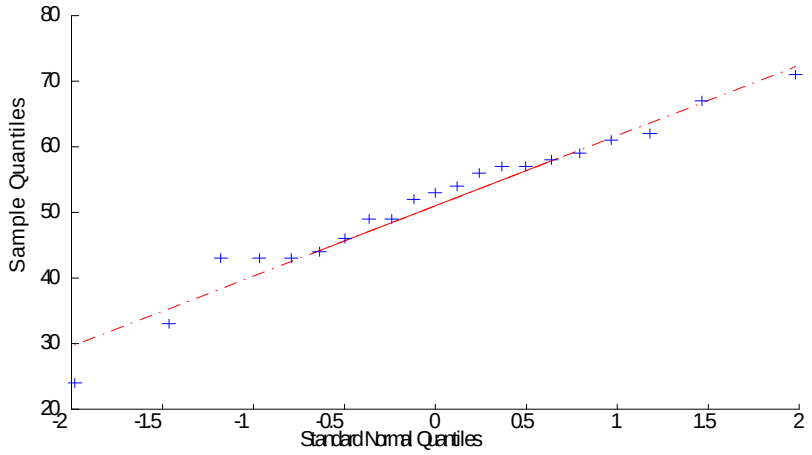


Figure 12.30: Normal Qqplot of the Responses for the Treatment Group

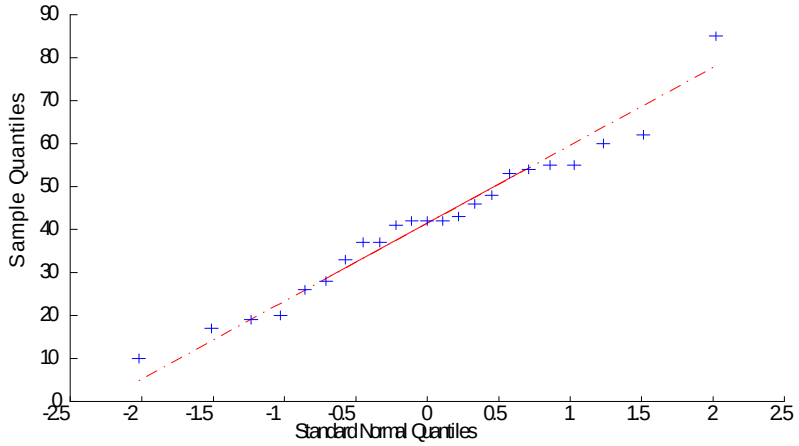


Figure 12.31: Normal Qqplot for the Responses in the Control Group

(c) For the given data

$$s_p = \left[\frac{1}{21+23-2}\left(2423.2381 + 6469.7391\right)\right]^{1/2} = 14.5512$$

Also $P\left(T \leq 2.018\right) = 0.975$ where $T \sim t\left(42\right)$. A 95% confidence interval for the

difference in the means, $\mu_1 - \mu_2$ is

$$\begin{aligned} & 51.4762 - 41.5217 \pm (2.018)(14.5512)\sqrt{\frac{1}{21}+\frac{1}{23}} \\ = \; & 9.9545 \pm 8.8628 = [1.0916,\ 18.8173] \end{aligned}$$

(d) To test the hypothesis of no difference between the means, that is, to test the hypothesis $H_0 : \mu_1 = \mu_2$ we use the discrepancy measure

$$D = \frac{\left|\bar{Y}_1 - \bar{Y}_2 - 0\right|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t\left(n_1 + n_2 - 2\right)$$

assuming $H_0 : \mu_1 = \mu_2$ is true. The observed value of D for these data is

$$d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{|51.4762 - 41.5217 - 0|}{14.5512\sqrt{\frac{1}{21}+\frac{1}{23}}} = 2.2666$$

and

$$\begin{aligned} p - value &= 2\left[1 - P\left(T \leq 2.2666\right)\right] \quad \text{where } T \sim t\left(42\right) \\ &= 0.02863. \end{aligned}$$

Since the p-value is less than 0.05 there is evidence against the hypothesis $H_0 : \mu_1 = \mu_2$ based on the data.
Although the data suggest there is a difference between the treatment group and the control group we **cannot conclude that the difference is due to the the new directed readings activities**. The difference could simply be due to the differences in the two Grade 3 classes. Since randomization was **not** used to determine which student received the treatment and which student was in the control group, the difference in the DRP scores could have existed before the treatment was applied.

(e) Here is the output from running t t.test in R
> # 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=TRUE,conf.level=0.95)

Two Sample t-test

data: DRP by Group
t = -2.2666, df = 42, p-value = 0.02863
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-18.817650 -1.091253
sample estimates:
mean in group Control mean in group Treatment
41.52174 51.47619

6.11 (a) The pooled estimate of variance is

$$s_p = \sqrt{\frac{209.02961 + 116.7974}{18}} = 4.25.$$

From t tables, $P(T < 1.734) = 0.95$ where $T \backsim t\,(18)$. The 90% confidence interval is

$$10.693 - 6.750 \pm 1.734\,(4.25)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = [0.647, 7.239]$$

(b) We test the hypothesis $H_0 : \mu_1 = \mu_2$ or equivalently $H_0 : \mu_1 - \mu_2 = 0$ using the pivotal

$$D = \frac{|\bar{Y}_1 - \bar{Y}_2|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The observed value of this statistic is

$$d = \frac{|10.693 - 6.750|}{4.25\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.074$$

with

$$p - value = 2\,[1 - P(T \leq 2.074)] = 0.05 \quad \text{where} \quad T \backsim t(18)$$

so there is weak evidence against H_0 based on the data.

(c) We repeat the above using as data $Z_{ij} = \log(Y_{ij})$. This time the sample means are 2.248, 1.7950 and the sample variances are 0.320, 0.240 respectively,. The pooled estimate of variance is $s_p = \sqrt{\frac{0.320+0.240}{2}} = 0.529$. The observed value of the discrepancy measure is

$$d = \frac{|2.248 - 1.795|\,0}{0.529\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.9148$$

with

$$p - value = 2\,[1 - P(T \leq 1.91)] \approx 0.07 \quad \text{where} \quad T \backsim t(18)$$

so there is even less evidence against H_0 based on the data.

(d) One could check the Normality assumption with qqplots for each of the variables Y_{ij} and $Z_{ij} = \log(Y_{ij})$ although with such a small sample size these will be difficult to interpret.

6.12 (a) The pooled estimate of the common standard deviation σ is

$$s_p = \sqrt{\frac{3050 + 2937}{58}} = 10.1599.$$

Using R, $P(T \leq 2.0017) = 0.975$ where $T \backsim t\,(58)$. The 95% confidence interval for $\mu_1 - \mu_2$ is

$$120 - 114 \pm 2.0017\,(10.1599)\sqrt{\frac{1}{30} + \frac{1}{30}} = 6 \pm 5.2511 = [0.7489, 11.2511]\,.$$

(b) Since

$$d = \frac{|120 - 114 - 0|}{10.1599\sqrt{\frac{1}{30} + \frac{1}{30}}} = 2.2872$$

with

$$p - value = 2\,[1 - P(T \leq 2.2872)] = 0.026 \quad \text{where} \quad T \backsim t(58)$$

there is evidence against the hypothesis of no difference based on the data. This is consistent with the fact that the 95% confidence interval for $\mu_1 - \mu_2$ did not contain the value $\mu_1 - \mu_2 = 0$.

6.13 Let μ_1 be the mean log failure time for welded girders and μ_2 be the mean score for log failure time for repaired welded girders. The pooled estimate of the common standard deviation σ is

$$s_p = \sqrt{\frac{13\,(0.0914) + 9\,(0.0422)}{22}} = 0.26697$$

From t tables, $P(T < 2.0739) = 0.975$ where $T \backsim t\,(22)$. The 95% confidence interval for $\mu_1 - \mu_2$ is

$$14.564 - 14.291 \pm 2.0739\,(0.26697)\sqrt{\frac{1}{14} + \frac{1}{10}} = 0.273 \pm 0.22924 = [0.04376, 0.50224]\,.$$

Since

$$d = \frac{|14.564 - 14.291 - 0|}{0.26697\sqrt{\frac{1}{14} + \frac{1}{10}}} = 2.4698$$

with

$$p - value = 2\,[1 - P(T \leq 2.4698)] = 0.02175 \quad \text{where} \quad T \backsim t(22)$$

there is evidence against the hypothesis of no difference based on the data. This is consistent with the fact that the 95% confidence interval for $\mu_1 - \mu_2$ did not contain the value $\mu_1 - \mu_2 = 0$.

6.14 (a) For the female coyotes we have

$$\bar{y}_f = 89.24, \quad s_f^2 = 42.87887, \quad n_f = 40.$$

For the male coyotes we have

$$\bar{y}_m = 92.06, \quad s_m^2 = 44.83586, \quad n_m = 43.$$

Since $n_f = 40$ and $n_m = 43$ are reasonably large we have that $\bar{Y}_f$ has approximately a $N(89.24, 42.87887/40)$ distribution and $\bar{Y}_m$ has approximately a $N(92.06, 44.83586/43)$ distribution. An approximate 95% confidence interval for $\mu_f - \mu_m$ is given by

$$89.24 - 92.06 \pm 1.96\sqrt{\frac{42.87887}{40} + \frac{44.83586}{43}} = [-5.67, 0.03].$$

The value $\mu_f - \mu_m = 0$ is just inside the right hand endpoint and the $p-value$ for testing $H_0 : \mu_f - \mu_m = 0$ would be close to 0.05. There is weak evidence based on the data of a difference between mean length for male and female coyotes. Since the interval contains mostly negative values the data suggest the mean length for males is slightly larger than for females.

(b) Using $Y_1 \backsim N(89.24, 42.87887)$, $Y_2 \backsim N(92.06, 44.83586)$ and $Y_1 - Y_2 \backsim N(89.24 - 92.06, 42.87887 + 44.83586)$ or $Y_1 - Y_2 \backsim N(-2.82, 87.71473)$ we estimate $P(Y_1 > Y_2) = P(Y_1 - Y_2 > 0)$ as

$$P\left(Z > \frac{0 - (-2.82)}{\sqrt{87.71473}}\right) = 1 - P(Z \leq 0.30) = 1 - 0.61791 = 0.38209.$$

(c) Since $P(T \leq 2.0227) = 0.975$ where $T \backsim t(39)$ a 95% confidence interval the mean length of female coyotes is

$$89.24 \pm 2.0227\sqrt{42.87887/40} = 89.24 \pm 2.0942 = [87.1457, 91.3342].$$

Since $P(T \leq 2.0181) = 0.975$ where $T \backsim t(42)$ a 95% confidence interval the mean length of male coyotes is

$$92.06 \pm 2.0181\sqrt{44.83586/43} = 92.06 \pm 2.06073 = [89.9993, 94.1207].$$

6.15 We assume that the observations for the "Alcohol" group are a random sample from a $G(\mu_1, \sigma)$ distribution and that the observations for the "Non-Alcohol" group are a random sample from a $G(\mu_2, \sigma)$ distribution. To see if there is any difference between the two groups we construct a 95% confidence interval for the mean difference in reaction times $\mu_1 - \mu_2$.

The pooled estimate of the common standard deviation is

$$s_p = \sqrt{\frac{0.608 + 0.35569}{22}} = 0.2093.$$

Since $P(T \leq 2.0739) = 0.975$ where $T \backsim t\,(22)$, a 95% confidence interval for $\mu_1 - \mu_2$ is

$$1.370 - 1.599 \pm 2.0739\,(0.2093)\sqrt{\frac{1}{12} + \frac{1}{12}} = [-0.4064, -0.0520].$$

This interval does not contain $\mu_1 - \mu_2 = 0$ and only contains negative values. The data suggest that $\mu_1 < \mu_2$, that is, the mean reaction time for the "Alcohol" group is less than the mean reaction time for the "Non-Alcohol" group. We are not told the units of these reaction times so it is unclear whether this difference is of practical significance.

6.16 (a) We assume that the observed differences are a random sample from a $G\,(\mu, \sigma)$ distribution. An estimate of σ is

$$s = \sqrt{\frac{17.135}{7}} = 1.5646$$

Since $P(T \leq 2.3646) = 0.975$ where $T \backsim t\,(7)$, a 95% confidence interval for μ is

$$1.075 \pm 2.3646\,(1.5646)\,/\sqrt{8} = 1.075 \pm 1.3080 = [-0.2330, 2.3830]\,.$$

(b) If the natural pairing is ignored an estimate of the common standard deviation is

$$s_p = \sqrt{\frac{535.16875 + 644.83875}{14}} = 9.18075$$

Since $P(T \leq 2.1148) = 0.975$ where $T \backsim t\,(14)$, a 95% confidence interval for $\mu_1 - \mu_2$ is

$$23.6125 - 22.5375 \pm 2.1148\,(9.18075)\sqrt{\frac{1}{8} + \frac{1}{8}} = [-8.7704, 10.9204].$$

We notice that although both intervals in (a) and (b) are centered at the value 1.075, the interval in (b) is very much wider.

(c) A matched pairs study allows for a more precise comparison since differences between the 8 pairs have been eliminated. That is by analyzing the differences we do not need to worry that there may have been large differences in the 8 cars which were used in the study with respect to other explanatory variates which might affect gas mileage (the response variate) such as size of engine, make of car, etc.

6.17 (a) We assume that the study population is the set of all factories of similar size. The parameter μ represents the mean difference in the number of staff hours per month lost due to accidents before and after the introduction of an industrial safety program in the study population.

(b) For these data

$$s = \left[\frac{1}{7}\left(1148.79875\right)\right]^{1/2} = 12.8107.$$

From t tables $P\left(T \leq 2.3646\right) = 0.975$ where $T \sim t\left(7\right)$. A 95% confidence interval for μ is

$$-15.3375 \pm 2.3646\left(12.8107\right)/\sqrt{8} = -15.3375 \pm 10.6891 = [-26.0266, -4.6484]$$

(c) Since

$$d = \frac{|\bar{y} - 0|}{s/\sqrt{n}} = \frac{|-15.3375 - 0|}{12.8107/\sqrt{8}} = 3.39$$

with

$$\begin{aligned} p - value &= 2\left[1 - P\left(T \leq 3.39\right)\right] \quad \text{where } T \sim t\left(7\right) \\ &= 0.012. \end{aligned}$$

Since the p-value is between 0.01 and 0.05 there is reasonable evidence against the hypothesis $H_0 : \mu = 0$ based on the data.
Since this experimental study was conducted as a matched pairs study, an analysis of the differences, $y_i = y_{1ii} - y_{2i}$, allows for a more precise comparison since differences between the 8 pairs have been eliminated. That is by analyzing the differences we do not need to worry that there may have been large differences in the safety records between factories due to other variates such as differences in the management at the different factories, differences in the type of work being conducted at the factories etc. Note however that a drawback to the study was that we were not told how the 8 factories were selected. To do the analysis above we have assumed that the 8 factories are a random sample from the study population of all similar size factories but we do not know if this is the case.

6.19 (a) Since two algorithms are each run on the same 20 sets of numbers we analyse the differences $y_i = y_{Ai} - y_{Bi}$, $i = 1, 2, \ldots, 20$. Since $P(T < 2.8609) = (1 + 0.99)/2 = 0.995$ where $T \backsim t\left(19\right)$, we obtain the confidence interval

$$0.409 \pm 2.8609\left(0.487322\right)/\sqrt{20} = [0.097, 0.721]$$

These values are all positive indicating strong evidence based on the data against $H_0 : \mu_A - \mu_B = 0$ $(p - value < 0.01)$, that is, the data suggest that algorithm B is faster.

(b) To check the Normality assumption we plot a qqplot of the differences. See Figure 12.32. The data lie reasonably along a straight line and therefore a Normal model is reasonable.

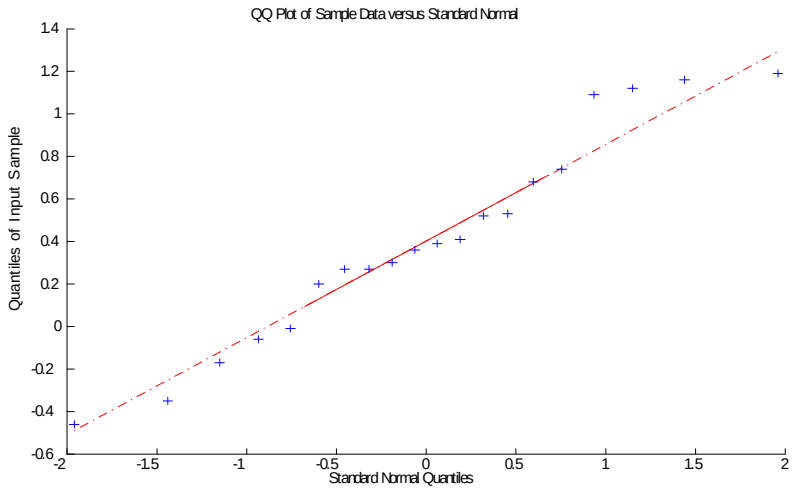


Figure 12.32: Qqplot for sorting algorithm data.

(c) We can estimate the probability by using the fact that $Y_A - Y_B \sim G(\mu, \sigma)$. We estimate the parameters using $\hat{\mu} = 0.40$ and $s = 0.487322$. Since

$$\begin{aligned} P(Y_A &> Y_B) = P(Y_A - Y_B > 0) = P\left(Z > \frac{0 - 0.409}{0.487322}\right) \\ &= P\left(Z > -0.84\right) = P\left(Z < 0.84\right) = 0.80 \quad \text{where } Z \backsim N\left(0,1\right) \end{aligned}$$

an estimate of the probability that algorithm B sorts a randomly selected list faster than A is 0.80.

(d) An estimate of p is $\hat{p} = 15/20 = 0.75$ and an approximate 95% is given by

$$\begin{aligned} \hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} &= 0.75 \pm 1.96\sqrt{\frac{0.75\left(0.25\right)}{20}} \\ &\text{or} \quad [0.56, 0.94] \end{aligned}$$

(e)

$$s_p = \sqrt{\frac{1.4697 + 0.9945}{2}} = 1.1100$$

Using R we have $P(T < 2.7116) = (1 + 0.99)/2 = 0.995$ where $T \backsim t(38)$. The interval, assuming common variance, is

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{20} + \frac{1}{20}} = 0.409 \pm 2.7116(1.1100)\sqrt{\frac{1}{20} + \frac{1}{20}}$$

or

$$[-0.543, 1.361]$$

This second interval $[-0.543, 1.361]$ is much wider than the first interval $[0.097, 0.721]$ biased on the paired experiment and unlike the first interval, it contains the value

zero. Unlike the paired design, independent samples of the same size (20 different problems run with each algorithm) is too small to demonstrate the superiority of algorithm B. The independent samples is a less efficient way to analyse the difference. This is why in computer simulations, it is essential to be able to run different simulations using the same random number seed.

(f) Here is the *R* output for doing the t tests and confidence intervals for the paired analysis and the unpaired analysis:

```
> t.test(Time~Algorithm,data=sortingdata,paired=TRUE,conf.level=0.99)

Paired t-test

data: Time by Algorithm
t = 3.7534, df = 19, p-value = 0.001346
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
0.09724793 0.72075207
sample estimates:
mean of the differences
0.409

> t.test(Time~Algorithm,data=sortingdata,paired=FALSE,var.equal=TRUE,conf.level=0.99)

Two Sample t-test

data: Time by Algorithm
t = 1.1652, df = 38, p-value = 0.2512
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
-0.5427918 1.3607918
sample estimates:
mean in group A mean in group B
4.7375 4.3285
```

Chapter 7

7.1 Here is the R output

```
> 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
> lambda
[1] 74.10284
> df<-7 # degrees for freedom for this example equal 7
> 1-pchisq(lambda,df) # p-value for LR test
[1] 2.181588e-13
> d<-sum((y-e)^2/e) # observed value of Pearson goodness of fit statistic
> d
[1] 72.86367
> 1-pchisq(d,df) # p-value for Pearson goodness of fit test
[1] 3.890221e-13
```

In both cases there is very strong evidence against the hypothesis that the distribution of colours is uniform.

7.2 The maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{1}{200}\sum_{i=1}^{5} i f_i = \frac{230}{200} = 1.15.$$

The expected frequencies assuming a Poisson(1.15) distribution are given in the table below in brackets

Number of Interruptions	0	1	2	3	4	≥ 5	Total
f_i	64	71	42	18	4	1	200
e_i	63.33	72.83	41.88	16.05	4.61	1.30	

where

$$e_i = 200 \times \frac{(1.15)^i\, e^{-1.15}}{i!} \quad \text{for} \quad i = 0, 1, \ldots, 4$$

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

Number of Interruptions	0	1	2	3	≥ 4	Total
f_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

$$2\left[64\log\left(\frac{64}{63.33}\right) + 71\log\left(\frac{71}{72.83}\right) + \cdots + 5\log\left(\frac{5}{5.91}\right)\right] = 0.43$$

and $p - value \approx P(W > 0.43) = 0.93$ where $W \backsim \chi^2(3)$. Based on the data there is no evidence against the hypothesis that the Poisson model fits the data.
Here is R code to do this analysis:

```
# estimate of theta = mean interruption time
th<-(0*64+1*71+2*42+3*18+4*4+5*1)/200
# observed frequencies for collapsed table
f=c(64,71,42,18,5)
# expected frequencies based on Poisson model
e<-200*c(dpois(0,th),dpois(1,th),dpois(2,th),dpois(3,th),1-ppois(3,th))
lambda<-2*sum(f*log(f/e))   # observed value of LR statistic
pvalue<-1-pchisq(lambda,3)   # p-value for LR test
c(lambda,pvalue)
```

7.3 The total number of defectives among the $250 \times 12 = 3000$ items inspected is

$$80 \times 1 + 31 \times 2 + 19 \times 3 + 11 \times 4 + 5 \times 5 + 1 \times 6 = 274$$

and the maximum likelihood estimate of θ = the proportion of defectives is

$$\hat{\theta} = \frac{274}{3000} = 0.09133.$$

We want to test the hypothesis that the number of defectives in a box is Binomial$(12, \theta)$. Under this hypothesis and using $\hat{\theta} = 0.091333$ we obtain the expected numbers in each category

Number defective	0	1	2	3	4	5	≥ 6	Total
e_i	79.21	95.54	52.82	17.70	4	0.64	0.08	250

where

$$e_i = 250 \times \binom{12}{i} \hat{\theta}^i (1 - \hat{\theta})^{12-i} \quad \text{for} \quad i = 0, 1, \ldots, 5$$

and the last category is obtained by subtraction. Since the expected numbers in the last three categories are all less than 5 we pool these categories to improve the Chi-squared approximation and obtain

Number defective	0	1	2	3	≥ 4	Total
f_i (e_i)	103(79.21)	80(95.54)	31(52.82)	19(17.7)	17(4.72)	250

The observed value of the likelihood ratio statistic is

$$\begin{aligned} & 2\left[103 \log\left(\frac{103}{79.21}\right) + 80 \log\left(\frac{80}{95.54}\right) + \cdots + 17 \log\left(\frac{17}{4.72}\right)\right] \\ = \; & 38.8552. \end{aligned}$$

Under the null hypothesis we had to estimate the parameter θ. The degrees of freedom are is $4-1=3$. The $p-value$ is $P(W>38.8552)\approx 0$ where $W \backsim \chi^2(3)$, so based on the data there is very strong evidence that the Binomial model does not fit. The likely reason is that the defects tend to occur in batches when packed (so that there more cartons with no defects than one would expect).

7.4 (a) Here is the R code for this problem:

```
> data<-c(70,75,63,59,81,92,75,100,63,58)
> L<-dmultinom(data,prob=data)
> L1<-dmultinom(data,prob=rep(1,10)) #This is L(θ̂)
> lambda<-2*(log(L)-log(L1))
> pvalue<-1-pchisq(lambda,9)
```

$\lambda = 23.605,\ \ p-value = 0.005$

Since the p-value is so small, there is strong evidence based on the data against the hypothesis that the machine is operating in a truly "random" fashion.

(b) $p-value = 1-(0.995)^6 = 0.03$

7.5 (a) For $n=2$, the likelihood function is

$$L_2(\theta_2) = \left[\binom{2}{0}(1-\theta_2)^2\right]^{23}\left[\binom{2}{1}\theta_2(1-\theta_2)\right]^{44}\left[\binom{2}{2}\theta_2^2\right]^{13} \quad 0<\theta_2<1$$

or more simply

$$L_2(\theta_2) = (1-\theta_2)^{2(23)}\theta_2^{44}(1-\theta_2)^{44}\theta_2^{2(13)} = \theta_2^{70}(1-\theta_2)^{90} \quad 0<\theta_2<1$$

which is maximized for

$$\hat{\theta}_2 = \frac{70}{160} = 0.4375.$$

For $n=3$

$$\begin{aligned} L_3(\theta_3) &= (1-\theta_3)^{3(10)}\theta_3^{25}(1-\theta_3)^{2(25)}\theta_3^{2(48)}(1-\theta_3)^{1(48)}\theta_3^{3(13)} \\ &= \theta_3^{160}(1-\theta_3)^{128} \quad 0<\theta_3<1 \end{aligned}$$

which is maximized for

$$\hat{\theta}_3 = \frac{160}{288} = 0.5556.$$

For $n=4$

$$\begin{aligned} L_4(\theta_4) &= (1-\theta_4)^{4(5)}\theta_4^{30}(1-\theta_4)^{3(30)}\theta_4^{2(34)}(1-\theta_3)^{2(34)} \\ &\quad \times\theta_4^{3(22)}(1-\theta_4)^{1(22)}\theta_4^{4(5)} \\ &= \theta_4^{184}(1-\theta_4)^{200} \quad 0<\theta_4<1 \end{aligned}$$

which is maximized for

$$\hat{\theta}_4 = \frac{184}{384} = 0.4792.$$

The expected frequencies assuming the Binomial model, are calculated using

$$e_{nj} = y_{n+}\binom{n}{j}\hat{\theta}_n^j\left(1-\hat{\theta}_n\right)^{n-j} \quad j = 0, 1, \ldots, n; \quad n = 2, 3, 4$$

and are given below:

		Number of females $= j$					Total number of litters
	e_{nj}	0	1	2	3	4	y_{n+}
Litter	2	25.3125	39.375	15.3125			80
Size $= n$	3	8.4280	31.6049	39.5062	16.4609		96
	4	7.0643	25.9964	35.8751	22.0034	5.0608	96

For $n = 2$ the observed value of the likelihood ratio statistic is

$$2\left[23\log\left(\frac{23}{25.3125}\right) + 44\log\left(\frac{44}{39.375}\right) + 13\log\left(\frac{13}{15.3125}\right)\right] = 1.11.$$

The degrees of freedom are $3 - 1 - 1 = 1$ since θ_2 was estimated.

$$\begin{aligned} p-value &\approx P(W \geq 1.11) \text{ where } W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{1.11}\right)\right] \text{ where } Z \backsim N(0,1) \\ &= 2\left[1 - P\left(Z \leq 1.05\right)\right] \\ &= 0.29220 \end{aligned}$$

$$p-value = P(W \geq 1.11)$$

and there is no evidence based on the data against the Binomial model. Similarly for $n = 3$, we obtain $\lambda = 4.22$ and $P(W \geq 4.22) = 0.12$ where $W \backsim \chi^2(2)$ and there is no evidence based on the data against the Binomial model. For $n = 4$, $\lambda = 1.36$ and $P(W \geq 1.36) = 0.71$ where $W \backsim \chi^2(3)$ and there is also no evidence based on the data against the Binomial model.

(b) The likelihood function for $\theta_1, \theta_2, \theta_3, \theta_4$ is

$$\begin{aligned} L(\theta_1, \theta_2, \theta_3, \theta_4) &= \theta_1^{12}(1-\theta_1)^8\,\theta_2^{70}(1-\theta_2)^{90}\,\theta_3^{160}(1-\theta_3)^{128}\,\theta_4^{184}(1-\theta_4)^{200} \\ 0 &< \theta_n < 1; \quad n = 1, 2, 3, 4. \end{aligned}$$

Under the hypothesis $H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta$ the likelihood function is

$$\begin{aligned} L(\theta) &= \theta^{12}(1-\theta)^8\,\theta^{70}(1-\theta)^{90}\,\theta^{160}(1-\theta)^{128}\,\theta^{184}(1-\theta)^{200} \\ &= \theta^{12+70+160+184}(1-\theta)^{8+90+128+200} \\ &= \theta^{426}(1-\theta)^{426} \quad 0 < \theta < 1 \end{aligned}$$

which is maximized for $\hat{\theta} = \frac{426}{852} = 0.5$. The expected frequencies, assuming H_0 are calculated using

$$e_{nj} = y_{n+}\binom{n}{j}(0.5)^n \quad j = 0, 1, \ldots, n; \quad n = 2, 3, 4$$

and are given below:

		Number of females $= j$					Total number
	e_{nj}	0	1	2	3	4	of litters $= y_{n+}$
Litter	1	10	10				20
Size $= n$	2	20	40	20			80
	3	12	36	36	12		96
	4	6	24	36	24	6	96

The observed value of the likelihood ratio statistic is

$$2\left[8\log\left(\frac{8}{10}\right) + 12\log\left(\frac{12}{10}\right) + \cdots + 22\log\left(\frac{22}{24}\right) + 5\log\left(\frac{5}{6}\right)\right] = 14.27.$$

The degrees of freedom $= (1 + 2 + 3 + 4) - 1 = 9$ and
$p - value \approx P(W \geq 14.27) = 0.11$ where $W \backsim \chi^2(9)$. There is no evidence based on the data against the hypothesis $\theta_1 = \theta_2 = \theta_3 = \theta_4$.

7.6 This process can be thought of as an experiment in which we observe $y_i =$ the number of non-zero digits (Failures) until the first zero (Success) for $i = 1, 2, \ldots, 50$ and $P(\text{Success}) = 0.1$. Therefore the Geometric(0.1) distribution is an appropriate model for these data. To test the fit of the model we summarize the data in a frequency table:

# between 2 zeros	0	1	2	3	4	5	6	7	8	10	12
# of occurrences	6	4	9	3	5	2	2	3	2	2	1
# between 2 zeros	13	14	15	16	18	19	20	21	22	26	
# of occurrences	1	1	1	1	1	1	1	1	2	1	

The expected frequencies are

$$e_j = 50 \times (0.1)(1 - 0.1)^j, \quad j = 0, 1, \ldots.$$

To obtain expected frequencies of at least five we join adjacent categories to obtain:

Observation between two 0's	0	1	$2-3$	$4-5$	$6-7$	$8-10$	≥ 11	Total
Observed Frequency.: f_j	6	4	12	7	5	4	12	50
Expected Frequency.: e_i	5	4.5	7.695	6.233	5.049	5.833	15.691	50

The observed value of the likelihood ratio statistic is $\lambda = 3.984$. The degrees of freedom for the Chi-squared approximation are $7 - 1 = 6$ and

$$p - value \approx P\left(W \geq 3.984\right) \approx 0.68 \text{ where } W \backsim \chi^2\left(6\right).$$

There is no evidence based on the data against the hypothesis that the Geometric(0.1) distribution is a good model for these data.

7.7 (a) The expected frequencies are:

e_{ij}	Rust-Proofed	Not Rust Proofed	Total
Rust present	$\frac{42\times 50}{100} = 21$	21	42
Rust absent	29	29	58
Total	50	50	100

The observed value of the likelihood ratio statistic is likelihood ratio statistic is

$$\lambda = 2\left[14\log\left(\frac{14}{21}\right) + 28\log\left(\frac{28}{21}\right) + 36\log\left(\frac{36}{29}\right) + 22\log\left(\frac{22}{29}\right)\right] = 8.1701$$

with

$$\begin{aligned} p - value &\approx P\left(W \geq 8.1701\right) \text{ where } W \backsim \chi^2\left(1\right) \\ &= 2\left[1 - P\left(Z \leq \sqrt{8.1701}\right)\right] \text{ where } Z \backsim N\left(0,1\right) \\ &= 2\left[1 - P\left(Z \leq 2.86\right)\right] = 0.0042587 \end{aligned}$$

so there is strong evidence against the hypothesis that the probability of rust occurring is the same for rust-proofed and non-rust-proofed cars based on the observed data.

7.8 (a) The expected frequencies are:

$y_{ij}\left(e_{ij}\right)$	Both	Mother	Father	Neither	Total
Above Average	$\frac{30\times 50}{100} = 15$	$\frac{16\times 50}{100} = 8$	$\frac{18\times 50}{100} = 9$	18	50
Below Average	15	8	9	18	50
Total	30	16	18	36	100

The observed value of the likelihood ratio statistic is $\lambda = 10.8$. The degrees of freedom for the Chi-squared approximation are $\left(4 - 1\right)\left(2 - 1\right) = 3$ and

$$p - value \approx P(W \geq 10.8) = 0.013 \text{ where } W \backsim \chi^2\left(3\right).$$

Therefore there is evidence based on the data against the hypothesis that birth weight is independent of parental smoking habits.

(b) The expected frequencies depending on whether the mother is a smoker or non-smoker are:

Mother smokes

$y_{ij}\,(e_{ij})$	Father smokes	Father non-smoker	Total
Above average	$\frac{30\times15}{46} = 9.78$	5.22	15
Below average	20.22	10.78	31
Total	30	16	46

Mother non-smoker

$y_{ij}\,(e_{ij})$	Father smokes	Father non-smoker	Total
Above average	$\frac{18\times35}{54} = 11.67$	23.33	35
Below average	6.33	12.67	19
Total	18	36	54

For the Mother smokes table, the observed value of the likelihood ratio statistic is $\lambda = 0.2644$. The degrees of freedom for the Chi-squared approximation are $(2-1)(2-1) = 1$ and

$$\begin{aligned} p-value &\approx P(W \geq 0.2644) \quad \text{where } W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.2644}\right)\right] \quad \text{where } Z \backsim N(0,1) \\ &= 2\left[1 - P(Z \leq 0.51)\right] = 0.60710 \end{aligned}$$

For the Mother non-smoker table, the observed value of the likelihood ratio statistic is $\lambda = 0.04078$. The degrees of freedom for the Chi-squared approximation are $(2-1)(2-1) = 1$ and

$$\begin{aligned} p-value &\approx P(W \geq 0.04078) \quad \text{where } W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.04078}\right)\right] \quad \text{where } Z \backsim N(0,1) \\ &= 2\left[1 - P(Z \leq 0.20)\right] = 0.83997 \end{aligned}$$

In both cases there is no evidence based on the data against the hypothesis that, given the smoking habits of the mother, birth weight is independent of the smoking habits of the father.

7.9 The expected frequencies are:

	Normal	Enlarged	Much enlarged	Total
Carrier present	$\frac{516\times72}{1398} = 26.57$	$\frac{589\times72}{1398} = 30.33$	15.09	72
Carrier absent	489.43	558.67	277.91	1326
Total	516	589	293	1398

The observed value of the likelihood ratio statistic is 7.3209 with

$$\begin{aligned} p-value &\approx P(W \geq 7.3209) = 0.026 \text{ where } W \backsim \chi^2(2) = \text{Exponential}(2) \\ &= e^{-7.3209/2} = 0.02572 \end{aligned}$$

so there is evidence based on the data against the hypothesis that the two classifications are independent.

7.10 The observed frequencies are:

y_{ij}	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

The expected frequencies are:

e_{ij}	Tall wife	Medium wife	Short wife	Total
Tall husband	$\frac{65\times50}{210} = 15.476$	$\frac{65\times104}{210} = 32.191$	17.333	65
Medium husband	$\frac{99\times50}{210} = 23.571$	$\frac{99\times50}{210} = 49.029$	26.400	99
Short husband	10.952	22.781	12.267	46
Total	50	104	56	210

The observed value of the likelihood ratio statistic is

$$\begin{aligned} \lambda &= 2[18\log\left(\frac{18}{15.476}\right) + 28\log\left(\frac{28}{32.191}\right) + 19\log\left(\frac{19}{17.333}\right) \\ &\quad +20\log\left(\frac{20}{23.571}\right) + 51\log\left(\frac{51}{49.029}\right) + 28\log\left(\frac{28}{26.400}\right) \\ &\quad +12\log\left(\frac{12}{10.952}\right) + 25\log\left(\frac{25}{22.781}\right) + 9\log\left(\frac{9}{12.267}\right)] \\ &= 3.1272 \end{aligned}$$

The degrees of freedom for the Chi-squared approximation are $(3-1)(3-1) = 4$ and

$$p-value \approx P(W \geq 3.1272) = 0.5368 \text{ where } W \backsim \chi^2(4).$$

There is no evidence based on the data against the hypothesis that the heights of husbands and wives are independent.

7.11 (a) The expected frequencies are:

e_{ij}	3 boys	2 boys	2 girls	3 girls	Total
Mother under 30	$\frac{29\times11}{64} = 4.9844$	$\frac{29\times18}{64} = 8.1563$	$\frac{29\times22}{64} = 9.96883$	5.8906	29
Mother over 30	6.0156	9.8438	12.0313	7.1094	35
Total	11	18	22	13	64

The observed value of the likelihood ratio statistic is $\lambda = 0.5587$. The degrees of freedom for the Chi-squared approximation are $(4-1)(2-1) = 3$ and

$$p - value \approx P(W \geq 0.5587) = 0.9058 \text{ where } W \backsim \chi^2(3).$$

There is no evidence based on the data to contradict the hypothesis of no association between the sex distribution and age of the mother.

(b) The expected frequencies are:

y = no. of boys	3	2	1	0	Total
Observed Frequency	11	18	22	13	64
Expected Frequency	$64(0.5)^3 = 8$	$64\binom{3}{2}(0.5)^2 = 24$	$64\binom{3}{1}(0.5)^2 = 24$	8	64

The observed value of the likelihood ratio statistic is $\lambda = 5.4441$. The degrees of freedom for the Chi-squared approximation are $4 - 1 = 3$ and

$$p - value \approx P(W \geq 5.4441) = 0.1420 \text{ where } W \backsim \chi^2(3).$$

There is no evidence based on the data against the Binomial$(3, 0.5)$ model.

7.12 The data in a two way table are:

f_{ij} $[e_{ij}]$	Cold	No Cold	Total
Vitamin C	20 [25]	80 [75]	100
Placebo	30 [25]	70 [75]	100
Total	50	150	200

If the probability of catching the cold is the same for each group, then an estimate of this probability is $50/200 = 0.25$. The expected frequencies and observed frequencies are shown in the table. The original model consists of two independent Binomial models each with their own unknown parameter. Under the null hypothesis that the probability of catching a cold is the same for both groups the model is two independent Binomial models with only one unknown parameter. Therefore the degrees of freedom for the Chi-squared approximation are $2-1 = 1$. The observed value of the likelihood ratio statistic is

$$2\left[20\log\left(\frac{20}{25}\right) + 80\log\left(\frac{80}{75}\right) + 30\log\left(\frac{30}{25}\right) + 70\log\left(\frac{70}{75}\right)\right] = 2.6807$$

and

$$\begin{aligned} p - value &\approx P(W \geq 2.68) \text{ where } W \backsim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{2.68}\right)\right] \text{ where } Z \backsim N(0,1) \\ &= 2\left[1 - P(Z \leq 1.64)\right] \\ &= 0.10157 \end{aligned}$$

Based on the observed data there is no evidence against the hypothesis that the probability of catching a cold during the study period was the same for each group.

Chapter 8

8.1 (a) The observed and expected frequencies (in square brackets) assuming independence are given in the table where $e_{11} = (3301 \times 28358)/50267 = 1862.25$, $e_{12} = (3301 \times 15328)/50267 = 1006.57$ and all other expected frequencies can be determined by subtraction.

No. of cigarettes	0	$1-20$	>20	Total
Weight ≤ 2.5	1322 [1862.25]	1186 [1006.57]	793 [432.17]	3301
Weight > 2.5	27036 [26495.75]	14142 [14321.42]	5788 [6148.83]	46966
Total	28358	15328	6581	50267

The observed value of the likelihood ratio statistic is 480.644 and $p-value \approx P(W \geq 480.644) \approx 0$ where $W \backsim \chi^2(2) = \text{Exponential}(2)$. Based on the data there is very strong evidence against the hypothesis that birth weight is independent of the mother's smoking habits. The data suggest that lower birth weights are associated with mothers who smoke more.

(b) Since this is an observational study, evidence of an association does not imply a causal relationship. In particular the researchers cannot conclude that if the mothers stopped smoking then birth weights would increase.

The researchers would need to conduct an experimental study in which they controlled how much the mothers smoked in order to conclude that the evidence of a relationship between mother's smoking habits and birth weights implies a causal relationship. Of course a study in which the researchers "controlled" the smoking habits of the mothers would be very difficult to conduct.

(c) An association between the smoking habits of fathers and birth weights is to be expected since there is probably an association between the smoking habits of the fathers and the smoking habits of the mothers. That is, the association between the smoking habits of fathers and birth weights is a result of the association between the smoking habits of the fathers and the smoking habits of the mothers together with the association between the smoking habits of the mothers and birth weights.

8.2 (a) The observed and expected frequencies (in square brackets) assuming independence are given in the table.

	Mark ≤ 80	Mark > 80	Total
Standard Lecture	60 $\left[\frac{75\times106}{150} = 53\right]$	15 [22]	75
CAI	46 [53]	15 [22]	75
Total	106	44	150

The observed value of the likelihood ratio statistic is 6.3874 and

$$\begin{aligned} p-value &\approx P\left(W \geq 6.3874\right) \quad \text{where } W \backsim \chi^2\left(1\right) \\ &= 2\left[1-P\left(Z \leq \sqrt{6.3874}\right)\right] \quad \text{where } Z \backsim N\left(0,1\right) \\ &= 2\left[1-P\left(Z \leq 2.53\right)\right] = 0.0114 \end{aligned}$$

Based on the data there is evidence against the hypothesis of independence, that is, against the hypothesis that marks are independent of whether the student received the standard lecture or some CAI.

(b) In order to conclude that CAI increases the chances of achieving a mark over 80%, randomization of the students to either a standard lecture or to CAI would need to have been done.

8.3 (a) The observed and expected frequencies (in square brackets) assuming independence are given in the table.

	Admitted	Not Admitted	Total
Standard Lecture	3738 $\left[\frac{8442\times 5232}{12763} = 3460.67\right]$	4704 [4981.33]	8442
CAI	1494 [1771.33]	2827 [2549.67]	4321
Total	5232	7531	12763

The observed value of the likelihood ratio statistic is $\lambda = 112.398$ and

$$\begin{aligned} p-value &\approx P\left(W \geq 112.398\right) \quad \text{where } W \backsim \chi^2\left(1\right) \\ &= 2\left[1-P\left(Z \leq \sqrt{112.398}\right)\right] \quad \text{where } Z \backsim N\left(0,1\right) \\ &= 2\left[1-P\left(Z \leq 10.60\right)\right] \approx 0 \end{aligned}$$

Based on the data there is very strong evidence against the hypothesis of independence, that is, against the hypothesis that whether the student is admitted or not is independent of their sex.

(b) Only Program A shows any evidence of non-independence, and that is in the direction of a lower admission rate for males.

(c) This is an example of Simpson's Paradox. The association is observed in the collapsed table since in the table broken down by program we observe that over 50% of the men applied to programs A and B which had higher admission rates while over 50% of the women applied to programs C - F which had much lower admission rates.

8.4 (a) The observed value of the test statistic

$$d = \frac{|11.7-12.0|}{\sqrt{\frac{(2.1)^2}{100}+\frac{(2.4)^2}{100}}} = 0.9407$$

and

$$\begin{aligned} p-value &\approx P\left(|Z| \geq 0.94\right) \quad \text{where } Z \backsim N\left(0,1\right) \\ &= 2\left[1-P\left(Z \leq 0.94\right)\right] \\ &= 2\left[1-P\left(Z \leq 10.60\right)\right] = 0.34722 \end{aligned}$$

Since the $p-value > 0.1$, there is not evidence based on the data against the hypothesis of no difference between the mean amount of rust for rust-proofed cars as compared to non-rust-proofed cars.

(b) Since the cars were not randomly assigned to rust-proofing or not, a variate that the manufacturer is not aware of which is not rust-proofing could have had an effect on the results. For example, maybe the cars that were rust-proofed were owned by drivers who lived in areas where salt is used frequently in winter and therefore they had decided to use rust-proofing to reduce the effects of the salt. The drivers who did not chose rust-proofing might live in areas where driving conditions do not affect the rusting of cars. It would have been better to use randomization to decide which of the cars received rust-proofing and which did not. In this way the variates that affect the rusting of cars that the manufacturer is not aware of are balanced in the two groups.

8.5 This study is an observational study based only on data from the United States. A causal relationship cannot be concluded only on the basis of these data. To establish a causal relationship a strong association would need to be observed in numerous studies in many countries. Other possible sources of confounding variates would need to be examined in these studies to determine if they could explain the association. A pathway by which drinking wine causes cirrhosis of the liver would need to be established.

8.6 This is an experimental study since Hooker observed the boiling point of water at many different elevation levels. (We don't know how he chose these levels.) We are assuming that his method for boiling water and for measuring water temperature and atmospheric pressure were controlled as much as possible at the different elevations to avoid other variates affecting the relationship. Recall that Hooker was interested in using the boiling point of water as the explanatory variate and atmospheric pressure as the response variate since measuring the boiling point would give travelers a quick way to estimate elevation, using the known relationship between elevation and barometric pressure, and the model relating pressure to boiling point. The causal relationship actually works in the reverse direction, that is, it is atmospheric pressure which is causing the change in the boiling point of water. This conclusion however requires an argument based on physics. Pressure on the surface of water tends to keep the water molecules contained. As pressure increases, water molecules need additional heat to gain the speed necessary for escape. Lowering the pressure lowers the boiling point because the molecules need less speed to escape.

8.7 It is important that the subject does not whether they are receiving the treatment since if they do know they might think the treatment is working just because they know that are receiving a treatment (the placebo effect). It is important that the physician not know whether the subject is receiving the treatment or not since knowing might affect their decision about whether the treatment is working or not.

APPENDIX B: SAMPLE TESTS

Sample Midterm Test 1

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?

   (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. Fill in the blanks below. You may use a numerical value or one of the following words or phrases: *sample skewness, sample kurtosis, sample variance, sample mean, relative frequencies, frequencies, histogram, boxplot.*

(a) A large positive value of the _____ indicates that the distribution is not symmetric and the right tail is larger than the left.

(b) The sum of the _____ equals 1.

(c) For a random sample from an Exponential(θ) distribution, the value of θ can be estimated using the _____.

(d) Suppose $y_{(1)}, y_{(2)}, \ldots, y_{(99)}, y_{(100)}$ are the ordered values of a data set with $y_{(1)} = \min(y_1, y_2, \ldots, y_{100})$ and $y_{(n)} = \max(y_1, \ldots, y_{100})$. Suppose $IQR = 3.85$ is the interquartile range of the data set. Then the IQR of the data set $y_{(1)}, ..., y_{(99)}$, $y_{(100)} + 5$ (that is, 5 is added only to the largest value) is equal to _____.

(e) Suppose $s^2 = 2.6$ is the sample variance of the data set $y_1, y_2, ..., y_{100}$. Then the sample variance of the data set $y_1 + 2, y_2 + 2, \ldots, y_{100} + 2$ (that is, 2 is added to every value) is _____.

[4] (f) The data $y_1, y_2, \ldots, y_{100}$ is recorded in kilometers (km) and the sample mean and sample skewness are recorded. If we decide instead to record the data in meters instead of kilometers, (1 meter is 0.001 km) then the sample mean is changed by a factor of _____ and the sample skewness is changed by a factor of _____.

3. Researchers are interested in the relationship between a certain gene and the risk of contracting diabetes. A gene is said to be expressed if its coded information is converted into certain proteins. A team of researchers investigates whether there is a relationship between a certain gene being expressed, and whether or not a person contracts diabetes in their lifetime. The team takes a random sample of 100 people who are aged 55 or above. For each person selected they determine (i) age, (ii) whether or not the gene is expressed, (iii) the person's insulin level, and (iv) if the person has diabetes.

a. This study is an example of **(check only those that apply)**

i. an experimental study because we need to experiment with the genes. ☐

ii. an observational study because we are recording observations for each sampled unit. ☐

iii. a probability model because probability is required to predict whether a person will contract diabetes. ☐

iv. a causative study because the diabetes causes the gene. ☐

v. a response study because the patient responds to the clinician. ☐

b. The "age" of the subject is an example of **(check only those that apply)**

i. an explanatory variate because it explains how long the subject is in the study. ☐

ii. an explanatory variate because it may help to explain whether a given person will contract diabetes. ☐

iii. a non-Normal variate because subjects may lie about their age. ☐

iv. a response variate because it responds to many different circumstances. ☐

c. The Plan step in PPDAC for this experiment includes **(check only those that apply)**

i. the question of whether or not diabetes was related to the expression of the gene. ☐

ii. the sampling protocol or the procedure used to select the sample. ☐

iii. the specification of the sample size. ☐

iv. the questions the researchers wished to investigate. ☐

v. a determination of the units that are available to be included in the study. ☐

d. In the Problem step of PPDAC, we **(check only those that apply)**

i. solve the problem for the maximum likelihood estimate. ☐

ii. list all problems that might be encountered in our analysis. ☐

iii. decide what questions we wish to address with this study. ☐

iv. decide what group of individuals we wish to apply the conclusions. ☐

v. define the variates that may be needed. ☐

4. In an experimental study conducted by Baumann and Jones of methods of teaching reading comprehension, the values of $n = 66$ test scores were recorded. Graphical summaries of the data are given in Figures 11.2-11.3. The summary statistics for these data are:

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Sample s.d.
30	40	45	44.015	49	57	6.644

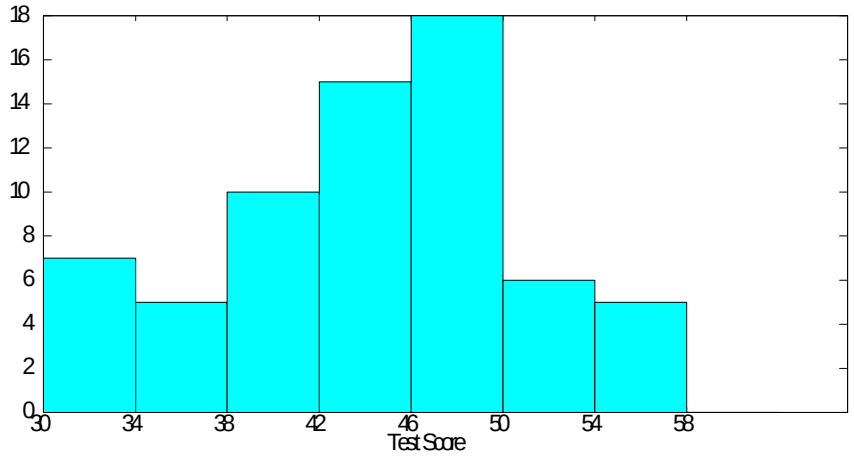
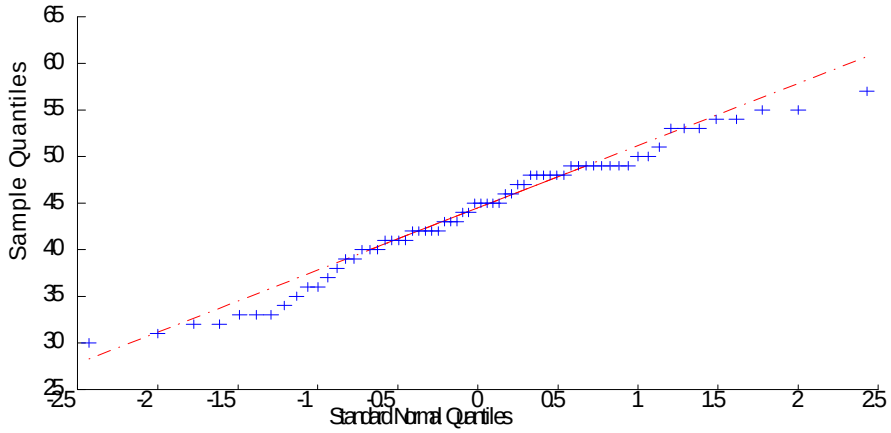


Figure 11.2: Frequency Histogram of Test Scores



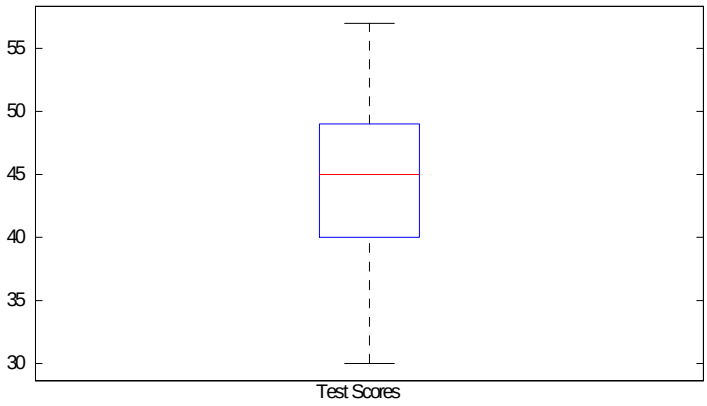


Figure 11.3: Boxplot of Test Scores

Based on these plots and statistics circle True or False for the following statements.

(a) The interquartile range is 9. True False

(b) The distribution has very large tails, too large to be consistent with the Normal distribution. True False

(c) The sample skewness is positive. True False

(d) About half of the test scores fall outside the interval $(40, 49)$. True False

(e) The shape of the Normal qqplot would change if 5 marks were added to each test score. True False

5. a. Suppose $y_1, y_2, \ldots, y_{25}$ are the observed values in a random sample from the Poisson(θ) distribution. Find the maximum likelihood estimate of θ. Show all your steps.

   b. Suppose $y_1, y_2, \ldots, y_{10}$ are the observed values in a random sample from the probability density function

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

   where $0 < \theta < \infty$. Find the maximum likelihood estimate of θ. Show all your steps.

Sample Midterm Test 1 Solutions

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?

A unit is a university student

(b) Define the target population.

The *set of all university students* in Ontario

(c) Define the study population.

The *set of university students living in the Kitchener Waterloo region in September 2012.*

(d) What are two variates in this problem?

sex (male/female), and agree/disagree with the statement

(e) What is the sampling protocol?

take a random sample of 250 students attending a specific Laurier-Waterloo football game

(f) A possible source of study error is:

There may be a difference between KW university students and the population of Ontario university students, for e.g., university students in Toronto and Thunder Bay may have different financial worries then KW university students.

(g) A possible source of sample error is:

Since more males tend to go to football games there may be a difference between the proportion of males in the sample and the proportion of males in the study population.

(h) Describe an attribute of interest for the target population and provide an estimate based on the given data.

An attribute of interest is the proportion of the target population that "agrees" with the statement. The estimate is 110/250 or 44%

2. Fill in the blanks below. You may use a numerical value or one of the following words or phrases: *sample skewness, sample kurtosis, sample variance, sample mean, relative frequencies, frequencies, histogram, boxplot.*

   a. A large positive value of the sample skewness indicates that the distribution is not symmetric and the right tail is larger than the left.

   b. The sum of the relative frequencies equals 1.

   c. For a random sample from an Exponential(θ) distribution, the value of θ can be estimated using the sample mean .

   d. Suppose $y_{(1)}, y_{(2)}, \ldots, y_{(99)}, y_{(100)}$ are the ordered values of a data set with $y_{(1)} = \min(y_1, y_2, \ldots, y_{100})$ and $y_{(100)} = \max(y_1, ..., y_{100})$. Suppose $IQR = 3.85$ is the interquartile range of the data set. Then the IQR of the data set $y_{(1)}, \ldots, y_{(99)}$, $y_{(100)} + 5$ (that is, 5 is added only to the largest value) is equal to 3.85 .

   e. Suppose $s^2 = 2.6$ is the sample variance of the data set $y_1, y_2, \ldots, y_{100}$. Then the sample variance of the data set $y_1 + 2, y_2 + 2, \ldots, y_{100} + 2$ (that is, 2 is added to every value) is 2.6 .

   f. he data $y_1, y_2, \ldots, y_{100}$ is recorded in kilometers (km) and the sample mean and sample skewness is recorded. If we decide instead to record the data in meters instead of kilometers, (1 meter is 0.001 km) then the sample mean is changed by a factor of 1000 and the sample skewness is changed by a factor of one (or the same) .

3. Researchers are interested in the relationship between a certain gene and the risk of contracting diabetes. A gene is said to be expressed if its coded information is converted into certain proteins. A team of researchers investigates whether there is a relationship between a certain gene being expressed, and whether or not a person contracts diabetes in their lifetime. The team takes a random sample of 100 people who are aged 55 or above. For each person selected they determine (i) age, (ii) whether or not the gene is expressed, (iii) the person's insulin level, and (iv) if the person has diabetes.

a. This study is an example of **(check only those that apply)**

   i. an experimental study because we need to experiment with the genes. ☐

   ii. an observational study because we are recording observations for each sampled unit. ☑

   iii. a probability model because probability is required to predict whether a person will contract diabetes. ☐

   iv. a causative study because the diabetes causes the gene. ☐

   v. a response study because the patient responds to the clinician. ☐

b. The "age" of the subject is an example of **(check only those that apply)**

   i. an explanatory variate because it explains how long the subject is in the study. ☐

   ii. an explanatory variate because it may help to explain whether a given person will contract diabetes. ☑

   iii. a non-Normal variate because subjects may lie about their age. ☐

   iv. a response variate because it responds to many different circumstances. ☐

c. The Plan step in PPDAC for this experiment includes **(check only those that apply)**

   i. the question of whether or not diabetes was related to the expression of the gene. ☐

   ii. the sampling protocol or the procedure used to select the sample. ☑

   iii. the specification of the sample size. ☑

   iv. the questions the researchers wished to investigate. ☐

   v. a determination of the units that are available to be included in the study. ☑

d. In the Problem step of PPDAC, we **(check only those that apply)**

i. solve the problem for the maximum likelihood estimate. ☐

ii. list all problems that might be encountered in our analysis. ☐

iii. decide what questions we wish to address with this study. ☑

iv. decide what group of individuals we wish to apply the conclusions. ☑

v. define the variates that may be needed. ☑

4. In an experimental study conducted by Baumann and Jones of methods of teaching reading comprehension, the values of $n = 66$ test scores were recorded. Graphical summaries of the data are given in Figures 1-3. The summary statistics for these data are:

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Sample s.d.
30	40	45	44.02	49	57	6.65

Based on these plots and statistics circle True or False for the following statements.

(a) The interquartile range is 9. **True** False

(b) The distribution has very large tails, too large to be consistent with the Normal distribution. True **False**

(c) The sample skewness is positive. True **False**

(d) About half of the test scores fall outside the interval $(40, 49)$. **True** False

(e) The shape of the Normal qqplot would change if 5 marks were added to each test score. True **False**

5. a. Suppose $y_1, y_2, ..., y_{25}$ are the observed values in a random sample from the Poisson(θ) distribution. Find the maximum likelihood estimate of θ. Show all your steps.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} \\ &= \left(\prod_{i=1}^{n} \frac{1}{y_i!} \right) \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \quad \text{note that the term } \prod_{i=1}^{n} \frac{1}{y_i!} \text{ is optional.} \end{aligned}$$

The log likelihood is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} y_i \log(\theta) - n\theta \\ l'(\theta) &= \frac{1}{\theta}\sum_{i=1}^{n} y_i - n = 0 \text{ for } \theta = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \end{aligned}$$

The maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}$.

(a) Suppose $y_1, y_2, ..., y_{10}$ are the observed values in a random sample from the probability density function

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

where $0 < \theta < \infty$. Find the maximum likelihood estimate of θ. Show all your steps.

$$L(\theta) = \prod_{i=1}^{n} \frac{y_i}{\theta^2} e^{-y_i/\theta} = \left(\prod_{i=1}^{n} y_i\right) \frac{1}{\theta^{2n}} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i\right) \text{ for } \theta > 0$$

or more simply

$$L(\theta) = \frac{1}{\theta^{2n}} \exp\left(-\frac{n\bar{y}}{\theta}\right) \text{ for } \theta > 0.$$

The log likelihood is

$$\begin{aligned} l(\theta) &= -2n\ln(\theta) - \frac{1}{\theta} n\bar{y} \text{ for } \theta > 0 \\ l'(\theta) &= -2n\left(\frac{1}{\theta}\right) + \left(\frac{1}{\theta^2}\right) n\bar{y} = 0 \text{ or } \frac{n}{\theta^2}(-2\theta + \bar{y}) = 0 \end{aligned}$$

The maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}/2$.

Sample Midterm Test 2

1.

(a) Suppose $Y \backsim \text{Binomial}(n, \theta)$. An experiment is to be conducted in which data y are to be collected to estimate θ. To ensure that the width of the approximate 90% confidence interval for θ is no wider that $2\,(0.02)$, the sample size n should be at least _____.

(b) Between December 20, 2013 and February 7, 2014 the Kitchener City Council conducted an online survey which was posted on the City of Kitchener's website. The online survey was publicized in the local newspapers, radio stations and TV news. The propose of the survey was to determine whether or not the citizens of Kitchener supported a proposal to put life sized bronze statues of Canada's past prime ministers in Victoria Park, Kitchener as a way to celebrate Canada's 150th. The community group that had proposed the idea had already received 2 million dollars in pledges and was asking the city for a contribution of $300,000 over three years.

People who took part in the survey were asked "Do you support the statue proposal in concept, by which we mean do you like the idea even if you don't agree with all aspects of the proposal?" Of the 2441 who took the survey, 1920 answered no to this question.

(i) Explain clearly whether you think using the online survey was a good way for the City of Kitchener to determine whether or not the citizens of Kitchener support the Prime Ministers' Statues Project.

(ii) Assume the model $Y \backsim \text{Binomial}(n, \theta)$ where Y = number of people who responded no to the question "Do you support the statue proposal in concept, by which we mean do you like the idea even if you don't agree with all aspects of the proposal?" What does the parameter θ represent in this study?

(iii) A point estimate of θ based on the observed data is _____.

(iv) An approximate 95% confidence interval for θ based on the observed data is _____.

(v) By reference to the confidence interval, indicate what you know about the $p-value$ for a test of the hypothesis $H_0 : \theta = 0.8$?

(c) Suppose a Binomial experiment is conducted and the observed 95% confidence interval for θ is $[0.1, 0.2]$. This means (circle the letter for the correct answer):

A : The probability that θ is contained in the interval $[0.1, 0.2]$ equals 0.95.

B : If the Binomial experiment was repeated 100 times independently and a 95% confidence interval was constructed each time then approximately 95 of these intervals would contain the true value of θ.

2. At the R.A.T. laboratory a large number of genetically engineered rats are raised for conducting research. Twenty rats are selected at random and fed a special diet. The weight gains (in grams) from birth to age 3 months of the rats fed this diet are:

63.4	68.3	52.0	64.5	62.3	55.8	59.3	62.4	75.8	72.1
55.6	73.2	63.9	60.7	63.9	60.2	60.5	67.1	66.6	66.7

Let y_i = weight gain of the i'th rat, $i = 1, 2, \ldots, 20$. For these data

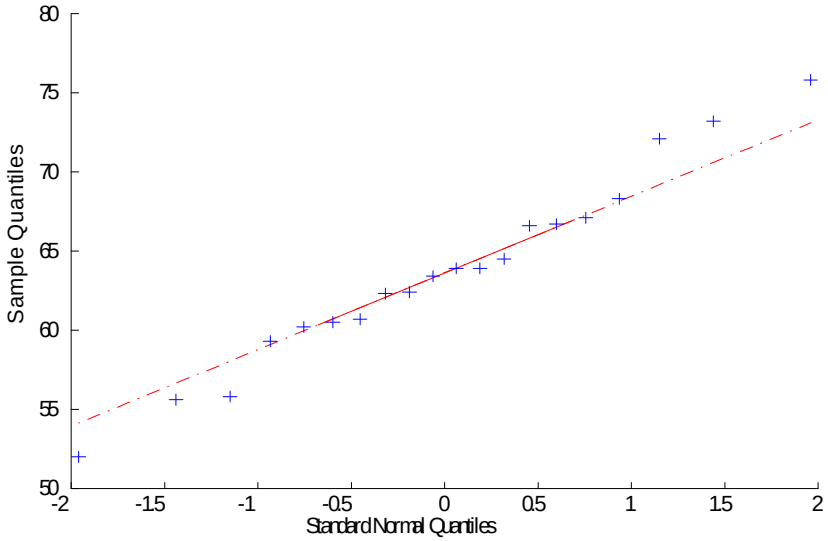
$$\sum_{i=1}^{20} y_i = 1273.8 \quad \text{and} \quad \sum_{i=1}^{20} (y_i - \bar{y})^2 = 665.718.$$

To analyze these data the model

$$Y_i \backsim N\left(\mu, \sigma^2\right) = G\left(\mu, \sigma\right), \quad i = 1, 2, \ldots, 20$$

is assumed where μ and σ are unknown parameters.

(a) Comment on how reasonable the Gaussian model is for these data based on the qqplot below:



(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 maximum likelihood estimate of μ is _____

The maximum likelihood estimate of σ is _____
(You do not need to derive these estimates.)

(d) Let

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{20}} \quad \text{where} \quad S^2 = \frac{1}{19}\sum_{i=1}^{20}\left(Y_i - \bar{Y}\right)^2.$$

The distribution of T is _____.

(e) The company, R.A.T. Chow, that produces the special diet claims that the mean weight gain for rats that are fed this diet is 67 grams.

The $p-value$ for testing the hypothesis $H_0 : \mu = 67$ is between _____ and _____.

What would you conclude about R.A.T. Chow's claim?

(f) Let $W = \frac{1}{\sigma^2}\sum_{i=1}^{20}\left(Y_i - \bar{Y}\right)^2$.

The distribution of W is _____.

Let a and b be such that $P(W \leq a) = 0.05 = P(W \geq b)$.

Then $a =$ _____ and $b =$ _____.

(g) A 90% confidence interval for σ for the given data is _____.

3. Let Y have an Exponential(θ) distribution with probability density function

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

(a) Show that $W = 2Y/\theta$ has probability density function given by

$$g(w) = \frac{1}{2}e^{-w/2}, \qquad \text{for} \quad w > 0$$

which is the probability density function of a random variable with a $\chi^2(2)$ distribution.

(b) Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Exponential(θ) distribution. Use your result from (a) and theorems that you have learned in class to prove that

$$U = \frac{2}{\theta}\sum_{i=1}^{n} Y_i \sim \chi^2(2n).$$

(c) Explain clearly how the pivotal quantity U can be used to obtain a two-sided $100p\%$ confidence interval for θ.

(d) Suppose $n = 25$ so that

$$U = \frac{2}{\theta} \sum_{i=1}^{25} Y_i \sim \chi^2 (50) .$$

Let a and b be such that $P(U \leq a) = 0.05 = P(U \geq b)$.

Then $a =$ _____ and $b =$ _____.

(e) Suppose $y_1, y_2, \ldots, y_{25}$ is an observed random sample from the Exponential(θ) distribution with $\sum_{i=1}^{25} y_i = 560$.

The maximum likelihood estimate for θ is _____.
(You do not need to derive this estimate.)

A 90% confidence interval for θ based on U is _____.

(f) Suppose an experiment is conducted and the hypothesis $H_0 : \theta = \theta_0$ is tested using a test statistic D with observed value d. If the $p - value = 0.01$ then this means (circle the letter for the correct answer):

A : the probability that $H_0 : \theta = \theta_0$ is correct equals 0.01.

B : the probability of observing a D value greater than or equal to d, assuming $H_0 : \theta = \theta_0$ is true, equals 0.01.

Sample Midterm Test 2 Solutions

1.
(a) [3] Suppose $Y \backsim$ Binomial(n, θ). An experiment is to be conducted in which data y are to be collected to estimate θ. To ensure that the width of the approximate 90% confidence interval for θ is no wider that $2\,(0.02)$, the sample size n should be at least __1692__.

Justification: An approximate 90% confidence interval for θ is given by $\hat{\theta} \pm 1.645\sqrt{\hat{\theta}(1-\hat{\theta})/n}$ since $P(Z \leq 1.645) = 0.95$ where $Z \backsim N(0,1)$ which has width $2\,(1.645)\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. Therefore we need n such that

$$(1.645)\sqrt{\hat{\theta}(1-\hat{\theta})/n} \leq 0.02$$

$$\text{or } n \geq \left(\frac{1.645}{0.02}\right)^2 \hat{\theta}(1-\hat{\theta})$$

Since we don't know $\hat{\theta}$ and the right side of the inequality takes on its largest value for $\hat{\theta} = 0.5$ we chose n such that

$$n \geq \left(\frac{1.645}{0.02}\right)^2 (0.5)^2 = 1691.3$$

Since n must be an integer we take $n = 1692$.

(b) Between December 20, 2013 and February 7, 2014 the Kitchener City Council conducted an online survey which was posted on the City of Kitchener's website. The online survey was publicized in the local newspapers, radio stations and TV news. The propose of the survey was to determine whether or not the citizens of Kitchener supported a proposal to put life sized bronze statues of Canada's past prime ministers in Victoria Park, Kitchener as a way to celebrate Canada's 150th. The community group that had proposed the idea had already received 2 million dollars in pledges and was asking the city for a contribution of $300,000 over three years.

People who took part in the survey were asked "Do you support the statue proposal in concept, by which we mean do you like the idea even if you don't agree with all aspects of the proposal?" Of the 2441 who took the survey, 1920 answered no to this question.

(i) Explain clearly whether you think using the online survey was a good way for the City of Kitchener to determine whether or not the citizens of Kitchener support the Prime Ministers' Statues Project.

This is not a good way for the City of Kitchener to determine whether or not the citizens of Kitchener support the Prime Ministers' Statues Project.

The respondents to the survey are people who heard about the survey through local media, had access to the internet and then took the time to complete the survey. These people are probably not representative of all citizens of Kitchener. This is an example of sample error.

To obtain a representative sample you would need to select a random sample of all citizens living in Kitchener.

(*ii*) Assume the model $Y \backsim \text{Binomial}(n, \theta)$ where $Y =$ number of people who responded no to the question "Do you support the statue proposal in concept, by which we mean do you like the idea even if you don't agree with all aspects of the proposal?" The parameter θ corresponds to what attribute of interest in the study population? Be sure to define the study population

The parameter θ corresponds to the proportion of people in the study population, which consists of all citizens of Kitchener, who would respond no to the question.

(*iii*) A point estimate of θ based on the observed data is $1920/2441 = 0.7866$.

(*iv*) An approximate 95% confidence interval for θ based on the observed data is $[0.7703, 0.8029]$.

$$\frac{1920}{2441} \pm 1.96\sqrt{\frac{1920}{2441}\left(1 - \frac{1920}{2441}\right)/2441} = 0.7866 \pm 0.0163 = [0.7703, 0.8029]$$

(*v*) By reference to the confidence interval, indicate what you know about the $p-value$ for a test of the hypothesis $H_0 : \theta = 0.8$?

Since $= 0.8$ is a value contained in the interval $[0.7703, 0.8029]$ therefore the $p - value$ for testing $H_0 : \theta = 0.8$ is greater than or equal to 0.05.

(Note that since $\theta = 0.8$ is very close to the upper endpoint of the interval that the $p - value$ would be very close to 0.05.)

(*c*) Suppose a Binomial experiment is conducted and the observed 95% confidence interval for θ is $[0.1, 0.2]$. This means (circle the letter for the correct answer):

A : The probability that θ is contained in the interval $[0.1, 0.2]$ equals 0.95.

$\boxed{B}$: If the Binomial experiment was repeated 100 times independently and a 95% confidence interval was constructed each time then approximately 95 of these intervals would contain the true value of θ.

2. At the R.A.T. laboratory a large number of genetically engineered rats are raised for conducting research. Twenty rats are selected at random and fed a special diet. The weight gains (in grams) from birth to age 3 months of the rats fed this diet are:

63.4	68.3	52.0	64.5	62.3	55.8	59.3	62.4	75.8	72.1
55.6	73.2	63.9	60.7	63.9	60.2	60.5	67.1	66.6	66.7

Let y_i = weight gain of the i'th rat, $i = 1, 2, \ldots, 20$. For these data

$$\sum_{i=1}^{20} y_i = 1273.8 \quad \text{and} \quad \sum_{i=1}^{20} (y_i - \bar{y})^2 = 665.718.$$

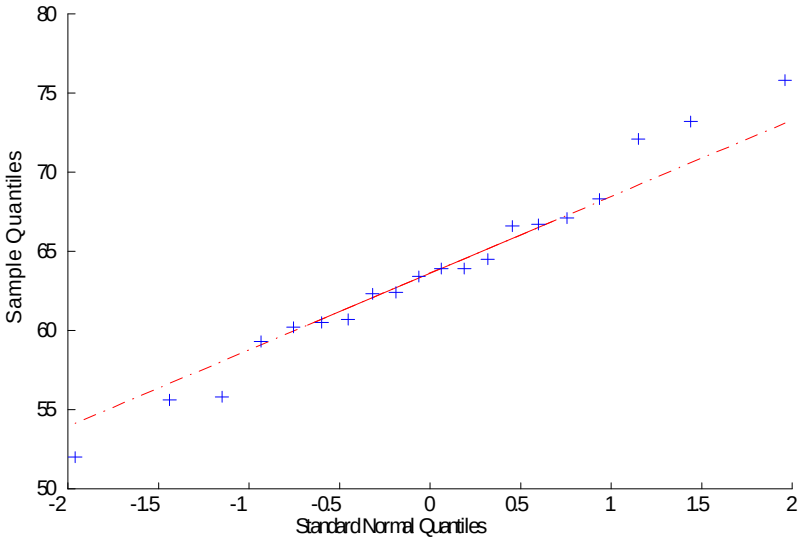
To analyze these data the model

$$Y_i \backsim N\left(\mu, \sigma^2\right) = G\left(\mu, \sigma\right), \quad i = 1, 2, \ldots, 20$$

is assumed where μ and σ are unknown parameters.

(a) Comment on how reasonable the Gaussian model is for these data based on the qqplot below:

Since the points in the qqplot lie reasonably along a straight line the Gaussian model seems 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?

A suitable study population consists of the genetically engineered rats which are raised for conducting research at the R.A.T. laboratory.

The parameter μ corresponds to the mean weight gain of the rats fed the special diet from birth to age 3 months in the study population.

The parameter σ corresponds to the standard deviation of the weight gains of the rats fed the special diet from birth to age 3 months in the study population.

(c) The maximum likelihood estimate of μ is $1273.8/20 = 63.69$

The maximum likelihood estimate of σ is $\left[\frac{1}{20}(665.718)\right]^{1/2} = (33.2859)^{1/2} = 5.7694$
(You do not need to derive these estimates.)

(d) [1] Let

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{20}} \quad \text{where} \quad S^2 = \frac{1}{19}\sum_{i=1}^{20}\left(Y_i - \bar{Y}\right)^2.$$

The distribution of T is $t(19)$.

(e) [6] The company, R.A.T. Chow, that produces the special diet claims that the mean weight gain for rats that are fed this diet is 67 grams.

The $p - value$ for testing the hypothesis $H_0 : \mu = 67$ is between 0.02 and 0.05.

$$s = \left[\frac{1}{19}(665.718)\right]^{1/2} = 5.9193 \qquad \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} = \frac{|63.69 - 67|}{5.9193/\sqrt{20}} = 2.5008$$

$$p - value = P\left(|T| \geq 2.5008\right) = 2\left[1 - P\left(T \leq 2.5008\right)\right]$$

Since $P\left(T \leq 2.5395\right) = 0.99$ and $P\left(T \leq 2.0930\right) = 0.975$, therefore

$$2\left(1 - 0.99\right) \leq p - value \leq 2\left(1.0975\right) \quad \text{or} \quad 0.02 \leq p - value \leq 0.05.$$

What would you conclude about R.A.T. Chow's claim?

Since the $p - value \leq 0.05$, therefore there is evidence against R.A.T. Chow's claim, $H_0 : \mu = 67$, based on the observed data.

(f) Let $W = \frac{1}{\sigma^2}\sum_{i=1}^{20}\left(Y_i - \bar{Y}\right)^2$.

The distribution of W is $\chi^2(19)$.

Let a and b be such that $P\left(W \leq a\right) = 0.05 = P\left(W \geq b\right)$.

Then $a =$ 10.117 and $b =$ 30.144.

(g) [2] A 90% confidence interval for σ for the given data is [4.6994, 8.1118].

$$\left[\left(\frac{665.718}{30.144}\right)^{1/2}, \left(\frac{665.718}{10.117}\right)^{1/2}\right] = [22.0846, 65.8019] = [4.6994, 8.1118]$$

3. Let Y have an Exponential(θ) distribution with probability density function

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

(a) Show that $W = 2Y/\theta$ has probability density function given by

$$g(w) = \frac{1}{2}e^{-w/2}, \qquad \text{for} \quad w > 0$$

which is the probability density function of a random variable with a $\chi^2(2)$ distribution.

For $w \geq 0$,

$$G(w) = P(W \leq w) = P\left(\frac{2Y}{\theta} \leq w\right) = P\left(Y \leq \frac{\theta w}{2}\right) = F\left(\frac{\theta w}{2}\right)$$

where

$$F(y) = P(Y \leq y)$$

Therefore

$$g(w) = G'(w) = f\left(\frac{\theta w}{2}\right) \cdot \left(\frac{\theta}{2}\right) = \frac{1}{\theta}\exp\left[-\left(\frac{\theta w}{2}\right)/\theta\right] \cdot \left(\frac{\theta}{2}\right) = \frac{1}{2}e^{-w/2}, \quad \text{for } w \geq 0$$

as required.

(b) Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Exponential(θ) distribution. Use your result from (a) and theorems that you have learned in class to prove that

$$U = \frac{2}{\theta}\sum_{i=1}^{n} Y_i \sim \chi^2(2n).$$

From (a), $\frac{2}{\theta}Y_i \backsim \chi^2(2)$ $i = 1, 2, \ldots, n$ independently.

Since the sum of independent Chi-squared random variables has a Chi-squared distribution with degrees of freedom equal to the sum of the degrees of freedom of the Chi-squared random variables in the sum, therefore

$$U = \frac{2}{\theta}\sum_{i=1}^{n} Y_i \sim \chi^2\left(\sum_{i=1}^{n} 2\right) \quad \text{or} \quad \chi^2(2n)$$

as required.

(c) Explain clearly how the pivotal quantity U can be used to obtain a two-sided $100p\%$ confidence interval for θ.

Using Chi-squared tables find a and b such that $P(U \leq a) = \frac{1-p}{2} = P(U \geq b)$ where $U \sim \chi^2(2n)$

Since

$$\begin{aligned} p &= P(a \leq U \leq b) \\ &= P\left(\frac{1}{b} \leq \frac{\theta}{2\sum_{i=1}^{n} Y_i} \leq \frac{1}{a}\right) \\ &= P\left(\frac{2\sum_{i=1}^{n} Y_i}{b} \leq \theta \leq \frac{2\sum_{i=1}^{n} Y_i}{a}\right) \end{aligned}$$

then

$$\left[\frac{2\sum_{i=1}^{n} y_i}{b}, \frac{2\sum_{i=1}^{n} y_i}{a}\right]$$

is a $100p\%$ confidence interval for θ.

(d) Suppose $n = 25$ so that

$$U = \frac{2}{\theta}\sum_{i=1}^{25} Y_i \sim \chi^2(50).$$

Let a and b be such that $P(U \leq a) = 0.05 = P(U \geq b)$.

Then $a =$ 34.764 and $b =$ 67.505.

(e) Suppose $y_1, y_2, \ldots, y_{25}$ is an observed random sample from the Exponential(θ) distribution with $\sum_{i=1}^{25} y_i = 560$.

The maximum likelihood estimate for θ is $560/25 = 22.4$.(You do not need to derive this estimate.)

A 90% confidence interval for θ based on U is $[16.5914, 32.2172]$.

$$\left[\frac{2\,(560)}{67.505}, \frac{2\,(560)}{34.764}\right] = [16.5914, 32.2172]$$

(f) Suppose an experiment is conducted and the hypothesis $H_0 : \theta = \theta_0$ is tested using a test statistic D with observed value d. If the $p - value = 0.01$ then this means (circle the letter for the correct answer):

A : the probability that $H_0 : \theta = \theta_0$ is correct equals 0.01.

$\boxed{B}$: the probability of observing a D value greater than or equal to d, assuming $H_0 : \theta = \theta_0$ is true, equals 0.01.

Sample Final Exam

1. A marketing research firm designed a study to examine the relationship between the amount of money spent in advertising a product on local television in one week and the sales of the product in the following week. The firm selected 4 levels of spending (in thousands of dollars) on advertising a product on local television in one week: 1.2, 2.4, 3.6, 4.8. Twenty communities in Ontario were selected. Each of the 4 levels of spending on advertising were applied in 5 different communities. The sales of the product (in thousands of dollars) in the following week measured in each of the 20 communities are given below:

Cost of Advertising (x)	Number of Communities	Total Sales (y)				
1.2	5	4.9	3.0	3.6	4.4	8.8
2.4	5	8.6	6.8	8.4	8.7	7.8
3.6	5	8.3	8.3	8.0	8.8	7.7
4.8	5	11.0	10.8	11.6	12.0	10.1

$$\begin{aligned} \bar{x} &= 3,\ \bar{y} = 7.93,\ S_{xx} = \sum_{i=1}^{20} (x_i - \bar{x})^2 = 36, \\ S_{yy} &= \sum_{i=1}^{20} (y_i - \bar{y})^2 = 125.282,\ S_{xy} = \sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 61.32 \end{aligned}$$

To analyse these data the regression model

$$Y_i = \alpha + \beta x_i + R_i, \quad \text{where} \quad R_i \backsim N\left(0, \sigma^2\right) = G(0, \sigma), \quad i = 1, 2, \ldots, 20 \ \text{ independently}$$

is assumed where α, β and σ are unknown parameters and the $x_i's$ are assumed to be known constants.

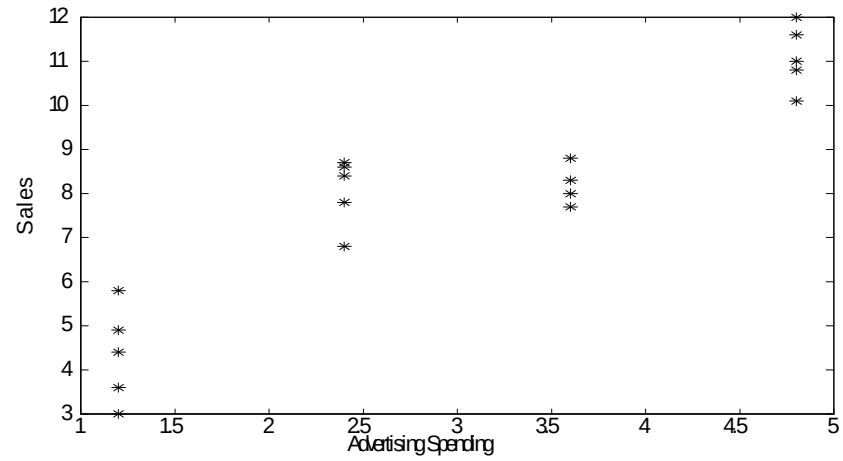
(a) Is this an experimental or observational study? Explain.

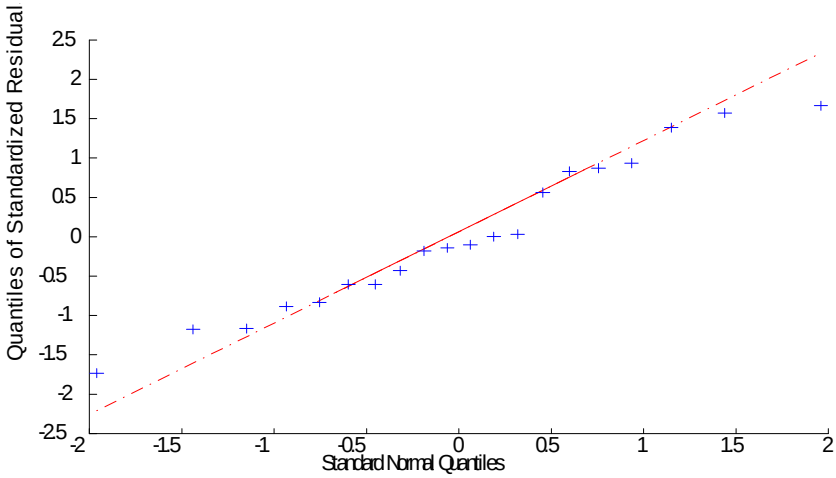
(b) Calculate the maximum likelihood estimates of α and β for these data and draw the fitted line on the scatterplot below. How well does this line fit the data? Do you notice anything unusual?

(c) A qqplot of the standardized residuals is given below. Explain clearly how this plot is obtained. What conclusions can be drawn from this plot about the validity of the assumed model for these data?

(d) Test the hypothesis that there is no relationship between the amount of money spent in advertising a product on local television in one week and the sales of the product in the following week. Show all your work.

(e) Would you conclude that an increase in the amount of money spent in advertising causes an increase in the sales of the product in the following week? Explain your answer.





(f) If the amount of dollars spent on advertising a product on local television in one week is 5 thousand dollars, find a 90% prediction interval for the sales of the product (in thousands of dollars) in the following week.

2. A wind farm is a group of wind turbines in the same location used for production of electric power. The number of wind farms is increasing as we try to move to more renewable forms of energy. Wind turbines are most efficient if the mean windspeed is 16 km/h or greater.

The windspeed Y at a specific location is modeled using the Rayleigh distribution which has probability density function

$$f(y;\theta) = \frac{2y}{\theta} e^{-y^2/\theta}, \qquad y \geq 0, \qquad \theta > 0$$

where θ is an unknown parameter which depends on the location.

(a) Let $y_1, y_2, \ldots, y_n$ be the windspeeds measured on n different days at a specific location. Assuming these observations represent n independent realizations of the random variable Y

which has the Rayleigh probability density function $f(y;\theta)$, find the Maximum Likelihood estimate of θ. Show all your work.

(b) To determine whether a location called Windy Hill is a good place for a wind farm, the windspeed was measured in km/h on 14 different days as given below:

14.7 30.0 13.3 41.9 25.6 39.6 34.5 9.9 13.6 24.2 5.1 41.4 20.5 22.2

$$\sum_{i=1}^{14} y_i = 336.5 \quad \text{and} \quad \sum_{i=1}^{14} y_i^2 = 9984.03$$

For these data calculate the Maximum Likelihood estimate of θ and give the Relative Likelihood function for θ.

(c) If the random variable Y has a Rayleigh distribution then $E(Y) = \sqrt{\theta\pi}/2$. Thus a mean of 20 km/h corresponds to $\theta = (40)^2/\pi \approx 509.3$. The owner of Windy Hill claims that the average windspeed at Windy Hill is 20 km/h. Test the hypothesis $H_0 : \theta = 509.3$ using the given data and the likelihood ratio test statistic. Show all your work.

(d) If Y_i has a Rayleigh distribution with parameter θ, $i = 1, 2, \ldots, n$ independently then

$$W = \frac{2}{\theta}\sum_{i=1}^{n} Y_i^2 \backsim \chi^2(2n).$$

If $n = 14$, find a and b such that $P(W \leq a) = 0.025 = P(W \geq b)$. Use the pivotal quantity W and the data from Windy Hill to construct an exact 95% confidence interval for θ. Show all your work.

(e) Would you recommend that a wind farm be situated at Windy Hill? Justify your answer.

3. Two drugs, both in identical tablet form, were each given to 10 volunteer subjects in a pilot drug trial. The order in which each volunteer received the drugs was randomized and the drugs were administered one day apart. For each drug the antibiotic blood serum level was measured one hour after medication. The data are given below:

Subject: i	1	2	3	4	5	6	7	8	9	10
Drug A: y_{1i}	1.08	1.19	1.22	0.60	0.55	0.53	0.56	0.93	1.43	0.67
Drug B: y_{2i}	1.48	0.62	0.65	0.32	1.48	0.79	0.43	1.69	0.73	0.71
Difference: $y_i = y_{1i} - y_{2i}$	−0.40	0.57	0.57	0.28	−0.93	−0.26	0.13	−0.76	0.70	−0.04

$$\sum_{i=1}^{10} y_i = -0.14 \quad \text{and} \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 2.90484.$$

To analyse these data the response model

$$Y_i \backsim G(\mu, \sigma), \quad i = 1, 2, \ldots, 10 \text{ independently}$$

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) Test the hypothesis of no difference in the mean response for the two drugs, that is, test $H_0 : \mu = 0$. Show all your work.

(c) Construct a 95% confidence interval for μ.

(d) This experiment is a matched pairs experiment. Explain why this type of design is better then a design in which 20 volunteers are randomly divided into two groups of 10 with one group receiving drug A and the other group receiving drug B.

(e) Explain the importance of randomizing the order of the drugs, the fact that the drugs where given in identical tablet form and the fact that the drugs were administered one day apart.

4. Exhaust emissions produced by motor vehicles is a major source of air pollution. One of the major pollutants in vehicle exhaust is carbon monoxide (CO). An environmental group interested in studying CO emissions for light-duty engines purchased 11 light-duty engines from Manufacturer A and 12 light-duty engines from Manufacturer B. The amount of CO emitted in grams per mile for each engine was measured. The data are given below:

Manufacturer A:

5.01 8.60 4.95 7.51 14.59 11.53 5.21 9.62 15.13 3.95 4.12

$$\sum_{j=1}^{11} y_{1j} = 90.22 \text{ and } \sum_{j=1}^{11} (y_{1j} - \bar{y}_1)^2 = 166.9860$$

Manufacturer B:

16.67 6.42 9.24 14.30 9.98 6.10 14.10 16.97 7.04 5.38 25.53 24.92

$$\sum_{j=1}^{12} y_{2j} = 136.65 \text{ and } \sum_{j=1}^{12} (y_{2j} - \bar{y}_2)^2 = 218.7656$$

To analyse these data assume the response model

$$Y_{1j} \backsim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 11 \text{ independently}$$

for Manufacturer A and independently

$$Y_{2j} \backsim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 12 \text{ independently}$$

for Manufacturer B where μ_1, μ_2 and σ are unknown parameters.

(a) Describe a suitable study population for this study. The parameters μ_1, μ_2 and σ correspond to what attributes of interest in the study population?

(b) Calculate a 99% confidence interval for the difference in the means: $\mu_1 - \mu_2$.

(c) Test the hypothesis $H_0 : \mu_1 = \mu_2$. Show all your work.

(d) What conclusions can the environmental group draw from this study? Justify your answer.

(e) A similar study was conducted but with 50 light-duty engines from manufacturer A and 50 light-duty engines from manufacturer B. The $p - value$ for testing
$H_0 : \mu_1 = \mu_2$ was equal to 0.018 and a 95% confidence for $\mu_1 - \mu_2$ was $[-3.0, -0.5]$. Explain the difference between a result which is statistically significant and a result which is of practical significance in the context of this larger study.

5. In a court case challenging an Oklahoma law that differentiated the ages at which young men and women could buy 3.2% beer, the Supreme Court examined evidence from a random roadside survey that measured information on age, gender, and drinking behaviour. The table below gives the results for the drivers under 20 years of age.

		Drank Alcohol in last 2 hours		Total
		Yes	No	
Gender	Male	77	404	481
of Driver	Female	16	122	138
Total		93	526	619

(a) Is this an experimental or observational study? Explain.

(b) Use the likelihood ratio statistic to test the hypothesis of no relationship (independence) between the two variates: gender and whether or not the driver drank alcohol in the last 2 hours. Show all your work.

(c) The Supreme Court decided to strike down the law that differentiated the ages at which young men and women could buy 3.2% beer based on the evidence presented. Do you agree with the Supreme Court's decision? Justify your answer.

6. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates university students' motivation, study habits, and attitudes toward university. At a small university college 19 students are selected at random and given the SSHA test. Their scores are:

10 10 11 12 13 13 13 14 14 14
14 15 15 15 16 16 17 18 20

Let y_i = score of the i'th student, $i = 1, 2, \ldots, 19$. For these data

$$\sum_{i=1}^{19} y_i = 270 \quad \text{and} \quad \sum_{i=1}^{19} y_i^2 = 3956.$$

For these data calculate the mean, median, mode, sample variance, range, and interquartile range.

7. A data set consisting of six columns was collected by interviewing 100 students on the University of Waterloo campus. The columns are:

Column 1: Sex of respondent (coded as 0 = Male and 1 = Female)

Column 2: Age of respondent

Column 3: Height of respondent

Column 4: Faculty of respondent

Column 5: Number of courses respondent has failed.

Column 6: Whether the respondent (i) strongly disagreed, (ii) disagreed, (iii) agreed or (iv) strongly agreed with the statement "The University of Waterloo is the best university in Ontario.

(a) For this data set give an example of each of the following types of data;

discrete _____

continuous _____

categorical _____

binary _____

ordinal _____

(b) Two ways to graphically represent categorical data are _____ and _____.

(c) A numerical summary of the relationship between two categorical data is _____.

(d) A graphical way to examine the relationship between heights and weights is a _____.

(e) If the sample correlation between heights and weights was 0.4 you would conclude _____.

(f) The data on height of respondent could be summarized graphically using:

_____

Sample Final Exam Solutions

1.

(*a*)

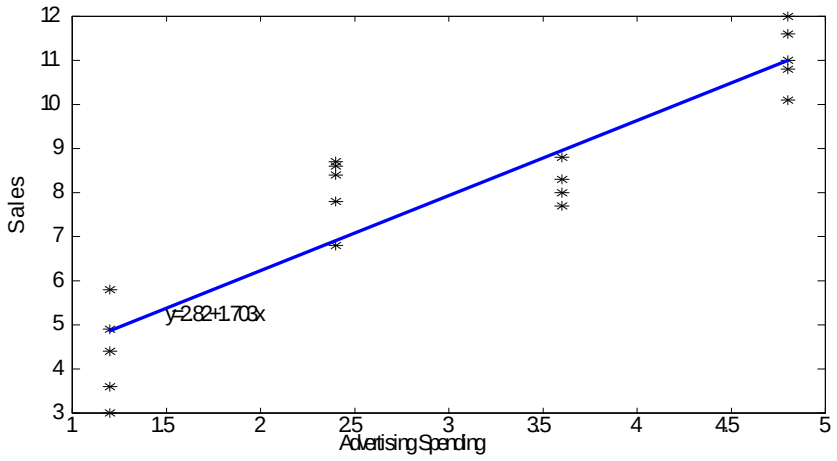
This is an experimental study since the research firm deliberately manipulated the levels of spending on advertising in each community.

(*b*)

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{61.32}{36} = 1.703, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 7.93 - \left(\frac{61.32}{36}\right)(3) = 2.82$$

The fitted line is $y = 2.82 + 1.703x$.

For $x = 1$, $y = 2.82 + 1.703(1) = 4.52$ and for $x = 5$, $y = 2.82 + 1.703(5) = 11.34$.



Looking at the scatterplot and the fitted line we notice that for $x = 2.4$, 4 of the 5 data points lie above the fitted line while for $x = 3.6$, all 5 of the data points lie below the fitted line. This suggests that the linear model might not be the best model for these data.

(*c*)

Calculate the estimated residuals $r_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$, $i = 1, 2, \ldots, 20$ and order the residuals from smallest to largest: $r_{(1)}, r_{(2)}, \ldots, r_{(n)}$.

Calculate q_i, $i = 1, 2, \ldots, 20$ where q_i satisfies $\mathrm{F}(q_i) = (i - 0.5)/20$ and F is the $\mathrm{N}(0, 1)$ cumulative distribution function. Plot $\left(r_{(i)}, q_i\right)$, $i = 1, 2, \ldots, 20$.

OR:

Calculate the estimated residuals $r_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$, $i = 1, 2, \ldots, 20$ and order the residuals from smallest to largest: $r_{(1)}, r_{(2)}, \ldots, r_{(n)}$.

Plot the ordered residuals against the theoretical quantiles of the Normal distribution.

Since there is no obvious pattern of departure from a straight line we would conclude that there is no evidence against the normality assumption $R_i \backsim N\left(0, \sigma^2\right)$, $i = 1, 2, \ldots, 20$.

(d) To test the hypothesis of no relationship we test $H_0 : \beta = 0$. We use the discrepancy measure

$$D = \frac{\left|\tilde{\beta} - 0\right|}{S/\sqrt{Sxx}}$$

where

$$T = \frac{\tilde{\beta} - 0}{S/\sqrt{Sxx}} \sim t\,(18) \quad \text{assuming } H_0 : \beta = 0 \text{ is true}$$

and

$$S^2 = \frac{1}{18}\sum_{i=1}^{20}\left(Y_i - \tilde{\alpha} - \tilde{\beta}x_i\right)^2.$$

Since $\hat{\beta} = 1.703$ and

$$s = \left[\frac{S_{yy} - \hat{\beta}S_{xy}}{18}\right]^{1/2} = \left[\frac{125.282 - (1.703)\,(61.32)}{18}\right]^{1/2} = (1.15742)^{1/2} = 1.0758$$

the observed value of D is

$$d = \frac{|1.703 - 0|}{1.0758/\sqrt{36}} = 9.50$$

and

$$\begin{aligned} p - value &= P\,(D \geq 9.50; H_0) \\ &= P\,(|T| \geq 9.50) \quad \text{where } T \sim t\,(18) \\ &\approx 0. \end{aligned}$$

Therefore there is very strong evidence based on the data against the hypothesis of no relationship between the amount of money spent in advertising a product on local television in one week and the sales of the product in the following week.

(e) Since this study was an experimental study, since there was strong evidence against $H_0 : \beta = 0$, and since the slope of the fitted line was $\hat{\beta} = 1.703 > 0$, the data suggest that an increase in the amount of money spent advertising causes an increase in the sales of the product in the following week. However we don't know if the 4 levels of spending on advertising were applied in the 5 different communities using randomization. If the levels of advertising were not randomly applied then the differences in the sales of the product could be due to differences between the communities. For example, if the highest (lowest) level was applied to the richest (poorest) communities you might expect to see the same pattern of response as was observed.

(f) From t tables $P(T \leq 1.73) = 0.95$ where $T \sim \mathrm{t}(18)$. A 90% prediction interval for the sales of the product (in thousands of dollars) in the following week if $x = 5$ is

$$\begin{aligned} & 2.82 + 1.703(5) \pm (1.73)(1.0758)\left[1 + \frac{1}{20} + \frac{(5-3)^2}{36}\right]^{1/2} \\ = \; & 11.3367 \pm 2.0055 \\ = \; & [9.33,\ 13.34] \end{aligned}$$

2.

(a) The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{2y_i}{\theta} e^{-y_i^2/\theta} = \left(\prod_{i=1}^{n} 2y_i\right) \theta^{-n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i^2\right), \qquad \theta > 0$$

and the log likelihood function is

$$l(\theta) = \log\left(\prod_{i=1}^{n} 2y_i\right) - n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} y_i^2, \qquad \theta > 0.$$

The derivative of the log likelihood function is

$$l'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} y_i^2 = \frac{1}{\theta^2}\left(-n\theta + \sum_{i=1}^{n} y_i^2\right), \qquad \theta > 0$$

and $l'(\theta) = 0$ if $\theta = \frac{1}{n}\sum_{i=1}^{n} y_i^2$. Therefore the Maximum Likelihood estimate of θ is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i^2.$$

(b) The Maximum Likelihood estimate for these data is

$$\hat{\theta} = \frac{9984.03}{14} = 713.145$$

and the Relative Likelihood function is

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{-14}\exp(-9984.03/\theta)}{\hat{\theta}^{-14}\exp\left(-9984.03/\hat{\theta}\right)} = \left(\frac{713.45}{\theta}\right)^{14}\exp(14 - 9984.03/\theta), \quad \theta > 0.$$

(c) The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0$ is

$$\Lambda = -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right]$$

which has approximately a $\chi^2(1)$ distribution if $H_0 : \theta = \theta_0$ is is true.

For these data the observed value of the likelihood ratio test statistic for $H_0 : \theta = 509.3$ is

$$\begin{aligned} d &= -2\log R(509.3) \\ &= -2\log\left[\left(\frac{713.45}{509.3}\right)^{14}\exp\left(14 - 9984.03/509.3\right)\right] \\ &= -2\log(0.4130) \\ &= 1.7688 \end{aligned}$$

and

$$\begin{aligned} p-value &= P(D \geq 1.7688; H_0) \\ &\approx P(W \geq 1.7688) \quad \text{where } W \sim \chi^2(1) \\ &= P\left(|Z| \geq \sqrt{1.7688}\right) \quad \text{where } Z \sim N(0,1) \\ &= 2\left[1 - P(Z \leq 1.33)\right] \\ &= 2(1 - 0.90824) = 2(0.09176) \\ &= 0.18352 \end{aligned}$$

Since the p-value $\approx 0.18352 > 0.1$, therefore there is no evidence based on the data against $H_0 : \theta = 509.3$.

(d) From χ^2 tables we have

$$P(W \leq 15.31) = 0.025 = P(W \geq 44.46) \quad \text{where } W \backsim \chi^2(28).$$

Since

$$\begin{aligned} 0.95 &= P\left(15.31 \leq \frac{2}{\theta}\sum_{i=1}^{n} Y_i^2 \leq 44.46\right) \\ &= P\left(\frac{2}{44.46}\sum_{i=1}^{n} Y_i^2 \leq \theta \leq \frac{2}{15.31}\sum_{i=1}^{n} Y_i^2\right) \end{aligned}$$

a 95% confidence interval for θ based on these data is given by

$$\begin{aligned} &\left[\frac{2(9984.03)}{44.46}, \frac{2(9984.03)}{15.31}\right] \\ = \ &[449.12,\ 1304.25]. \end{aligned}$$

(e) A 95% confidence interval for the mean windspeed $\sqrt{\theta\pi}/2$ based on these data is

$$\begin{aligned} &\left[\frac{\sqrt{449.12\pi}}{2}, \frac{\sqrt{1304.252\pi}}{2}\right] \\ = \ &[18.78,\ 32.01]. \end{aligned}$$

Since the values of this interval are all above 16, the data seem to suggest a mean windspeed greater than 16km/hr. However we don't know how the data were collected. It would be wise to determine how the data were collected before reaching a conclusion. Suppose that Windy Hill is only windy at one particular time of the year and that the data were collected only during the windy period. We would not want to make a decision only based on these data.

3.

(a) A suitable study population would consist of individuals who have volunteered to partake in clinical trials.

The parameter μ corresponds to the mean difference in antibiotic blood serum level between drugs A and B in the study population.

The parameter σ corresponds to the standard deviation of the differences in antibiotic blood serum level between drugs A and B in the study population.

(b) To test the hypothesis of no difference in the mean response for the two drugs, that is, $H_0 : \mu = 0$ we use the discrepancy measure

$$D = \frac{\left|\bar{Y} - 0\right|}{S/\sqrt{10}}$$

where

$$T = \frac{\bar{Y} - 0}{S/\sqrt{10}} \sim t\,(9) \quad \text{assuming } H_0 : \mu = 0 \text{ is true}$$

and

$$S^2 = \frac{1}{9}\sum_{i=1}^{10}\left(Y_i - \bar{Y}\right)^2.$$

Since $\bar{y} = -0.14/10 = -0.014$ and

$$s = \left[\frac{2.90484}{9}\right]^{1/2} = (0.32276)^{1/2} = 0.5681$$

the observed value of D is

$$d = \frac{|-0.014 - 0|}{0.5681/\sqrt{10}} = 0.078$$

and

$$\begin{aligned} p - value &= P\,(D \geq 0.078; H_0) \\ &= P\,(|T| \geq 0.078) \quad \text{where } T \sim t\,(9) \\ &= 2\,[1 - P\,(T \leq 0.078)]\,. \end{aligned}$$

From t tables $P\,(T \leq 0.261) = 0.6$ and $P\,(T \leq 0) = 0.5$ so

$$0.8 = 2\,(1 - 0.6) \leq p - value \leq 2\,(1 - 0.5) = 1.$$

Therefore there is no evidence based on the data against the hypothesis of no difference in the mean response for the two drugs, that is, $H_0 : \mu = 0$.

(c) From t tables $P(T \leq 2.26) = 0.975$ where $T \sim t(9)$. A 95% confidence interval for μ based on these data is

$$\begin{aligned} & \bar{y} \pm 2.26\,(s)/\sqrt{10} \\ = \; & -0.014 \pm 2.26\,(0.5681)/\sqrt{10} \\ = \; & [-0.4200,\ 0.3920]\,. \end{aligned}$$

(d) Since this experimental study was conducted as a matched pairs study, an analysis of the differences, $y_i = y_{1i} - y_{2i}$, allows for a more precise comparison since differences between the 10 pairs have been eliminated. That is, by analysing the differences we do not need to worry that there may have been large differences in the responses between subjects due to other variates such as age, general health, etc.

(e) It is important to randomize the order of the drugs in case the order in which the drugs are taken affects the outcome.

It is important to give the drugs in identical tablet form so the subject does not know which drug he or she is taking since knowing which drug is being taken could affect the outcome.

It is important that the drugs be administered one day apart to ensure that the effects of one drug are gone before the second drug is given.

4.

(a) The study population would consist of light-duty engines produced by Manufacturer A and Manufacturer B.

The parameter μ_1 corresponds to the mean amount of CO emitted by light-duty engines produced by Manufacturer A.

The parameter μ_2 corresponds to the mean amount of CO emitted by light-duty engines produced by Manufacturer B.

The parameter σ corresponds to the standard deviation of the CO emissions from light-duty engines produced by Manufacturers A and B. (Note that it has been assumed that this standard deviation is the same for both manufacturers.)

(b) From t tables $P(T \leq 2.83) = 0.995$ where $T \sim t(21)$. For these

$$s = \left[\frac{166.9860 + 218.7656}{21}\right]^{1/2} = (18.3691)^{1/2} = 4.2860$$

A 99% confidence interval for $\mu_1 - \mu_2$ based on these data is

$$\begin{aligned} & \bar{y}_1 - \bar{y}_1 \pm 2.83\,(s)\sqrt{\frac{1}{11}+\frac{1}{12}} \\ = \quad & \frac{90.22}{11} - \frac{136.65}{12} \pm 2.83\,(4.2860)\sqrt{\frac{1}{11}+\frac{1}{12}} \\ = \quad & -3.1857 \pm 5.0631 \\ = \quad & [-8.2487,\ 1.8774]\,. \end{aligned}$$

(c) To test the hypothesis of no difference in the mean response for the two drugs, that is, $H_0 : \mu_1 = \mu_2$ we use the discrepancy measure

$$D = \frac{\left|\bar{Y}_1 - \bar{Y}_2 - 0\right|}{S\sqrt{\frac{1}{11}+\frac{1}{12}}}$$

where

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{S\sqrt{\frac{1}{11}+\frac{1}{12}}} \sim t\,(21) \quad \text{assuming } H_0 : \mu_1 = \mu_2 \text{ is true}$$

and

$$S^2 = \frac{1}{21}\left[\sum_{i=1}^{11}\left(Y_{1i} - \bar{Y}_1\right)^2 + \sum_{i=1}^{12}\left(Y_{2i} - \bar{Y}_2\right)^2\right].$$

Since

$$\bar{y}_1 - \bar{y}_2 = \frac{90.22}{11} - \frac{136.65}{12} = -3.1857$$

and $s = 4.2860$ the observed value of D is

$$d = \frac{|-3.1857 - 0|}{4.2860\sqrt{\frac{1}{11}+\frac{1}{12}}} = 1.7806$$

and

$$\begin{aligned} p - value \quad &= \quad P\,(D \geq 1.7806; H_0) \\ &= \quad P\,(|T| \geq 1.7806) \quad \text{where} \quad T \sim t\,(21) \\ &= \quad 2\,[1 - P\,(T \leq 1.7806)]\,. \end{aligned}$$

From t tables $P\,(T \leq 1.73) = 0.95$ and $P\,(T \leq 2.09) = 0.975$ so

$$0.05 = 2\,(1 - 0.975) \leq p - value \leq 2\,(1 - 0.95) = 0.1$$

and therefore there is weak evidence based on the data against the hypothesis of no difference in the mean response for the two drugs, that is, $H_0 : \mu_1 = \mu_2$.

(d) Although there is weak evidence of a difference between the mean CO emissions for the two manufacturers it is difficult to draw much of a conclusion. The sample sizes $n_1 = 11$

and $n_2 = 12$ are small. We also don't know whether the engines were chosen at random from the two manufacturers on the day, week, or month. In other words we don't know if the samples are representative of all light-duty engines produced by these manufacturers.

(e) Since the $p - value$ for testing $H_0 : \mu_1 = \mu_2$ was equal to 0.018 then there is a statistically significant difference in the two means. The 95% confidence for $\mu_1 - \mu_2$ is $[-3.0, -0.5]$ where the units are grams per mile. The values in this interval seem small but we would need to consult with an environmental scientist to determine if a difference of between 1/2 a gram per mile and 3 grams per mile is of practical importance in terms of the damage to the earth's atmosphere.

5.

(a) This is an observational study because no explanatory variates were manipulated by the researcher.

(b) If the hypothesis of no relationship (independence) between the two variates gender and whether or not the driver drank alcohol in the last 2 hours is true then the expected frequency for the outcome male and drank alcohol in the last 2 hours for the given data is

$$e_{11} = \frac{93 \times 481}{619} = 72.27.$$

The other expected frequencies e_{12}, e_{21}, e_{22} can be obtained by subtraction from the appropriate row or column total. The expected frequencies are given in brackets in the table below.

		Drank Alcohol in last 2 hours		Total
		Yes	No	
Gender	Male	77 (72.27)	404 (408.73)	481
of Driver	Female	16 (20.73)	122 (117.27)	138
Total		93	526	619

The observed value of the likelihood ratio statistic is

$$\begin{aligned}
\lambda &= 2 \sum_{j=1}^{2} \sum_{j=1}^{2} y_{ij} \log \left(\frac{y_{ij}}{e_{ij}} \right) \\
&= 2 \left[77 \log \left(\frac{77}{72.27} \right) + 404 \log \left(\frac{404}{408.73} \right) + 16 \log \left(\frac{16}{20.73} \right) + 122 \log \left(\frac{122}{117.27} \right) \right] \\
&= 1.7208
\end{aligned}$$

Since the expected frequencies are all great than 5 then D has approximately a $\chi^2(1)$ distribution.

Thus

$$\begin{aligned} p-value &\approx P\left(W \geq 1.7208\right) \quad \text{where } W \backsim \chi^2\left(1\right) \\ &= P\left(|Z| \geq \sqrt{1.7208}\right) \quad \text{where } Z \sim N\left(0,1\right) \\ &= 2\left[1 - P\left(Z \leq 1.31\right)\right] \\ &= 2\left(1 - 0.90490\right) \\ &= 0.1902 \end{aligned}$$

Since the p-value $= 0.1902 > 0.1$ we would conclude that there is no evidence against the hypothesis of no relationship between between the two variates, gender and whether or not the driver drank alcohol in the last 2 hours.

(c) Although there is no evidence against the hypothesis of no relationship between the two variates: gender and whether or not the driver drank alcohol in the last 2 hours based on the data we cannot conclude there is no relationship since this is an observational study. Whether a causal relationship exists or not cannot be determined by an observational study only. A decision to strike down the law based on these data alone is unwise.

6.The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates university students' motivation, study habits, and attitudes toward university. At a small university college 19 students are selected at random and given the SSHA test. Their scores are:

10	10	11	12	13	13	13	14	14	14
14	15	15	15	16	16	17	18	20	

Let y_i = score of the i'th student, $i = 1, 2, \ldots, 19$. For these data

$$\sum_{i=1}^{19} y_i = 270 \quad \text{and} \quad \sum_{i=1}^{19} y_i^2 = 3956.$$

For these data calculate the mean, median, mode, sample variance, range, and interquartile range.

$$\begin{aligned} \text{mean} &= 14.21, \ \text{median} = 14, \ \text{mode} = 14, \ \text{sample variance} = 6.62, \\ \text{range} &= 20 - 10 = 10, \ \text{IQR} = 16 - 13 = 3 \end{aligned}$$

7. A data set consisting of six columns of data was collected by interviewing 100 students on the University of Waterloo campus. The columns are:

Column 1: Sex of respondent

Column 2: Age of respondent

Column 3: Height of respondent

Column 4: Faculty of respondent

Column 5: Number of courses respondent has failed.

Column 6: Whether the respondent (i) strongly disagreed, (ii) disagreed, (iii) agreed or (iv) strongly agreed with the statement "The University of Waterloo is the best university in Ontario.

(a) For this data set give an example of each of the following types of data;

discrete number of courses failed

continuous height or age

categorical faculty or sex

binary sex

ordinal degree of agreement with statement

(b) Two ways to graphically represent categorical data are pie charts and bar charts.

(c) A numerical summary of the relationship between two categorical data is relative risk.

(d) A graphical way to examine the relationship between heights and weights is a scatterplot .

(e) If the sample correlation between heights and weights was 0.4 you would conclude that there is a positive linear relationship between heights and weights.

(f) The data on height of respondent could be summarized graphically using: a relative frequency histogram, an empirical c.d.f., or a boxplot .