UNIVERSITY OF WATERLOO

WATERLOO MATHEMATICS

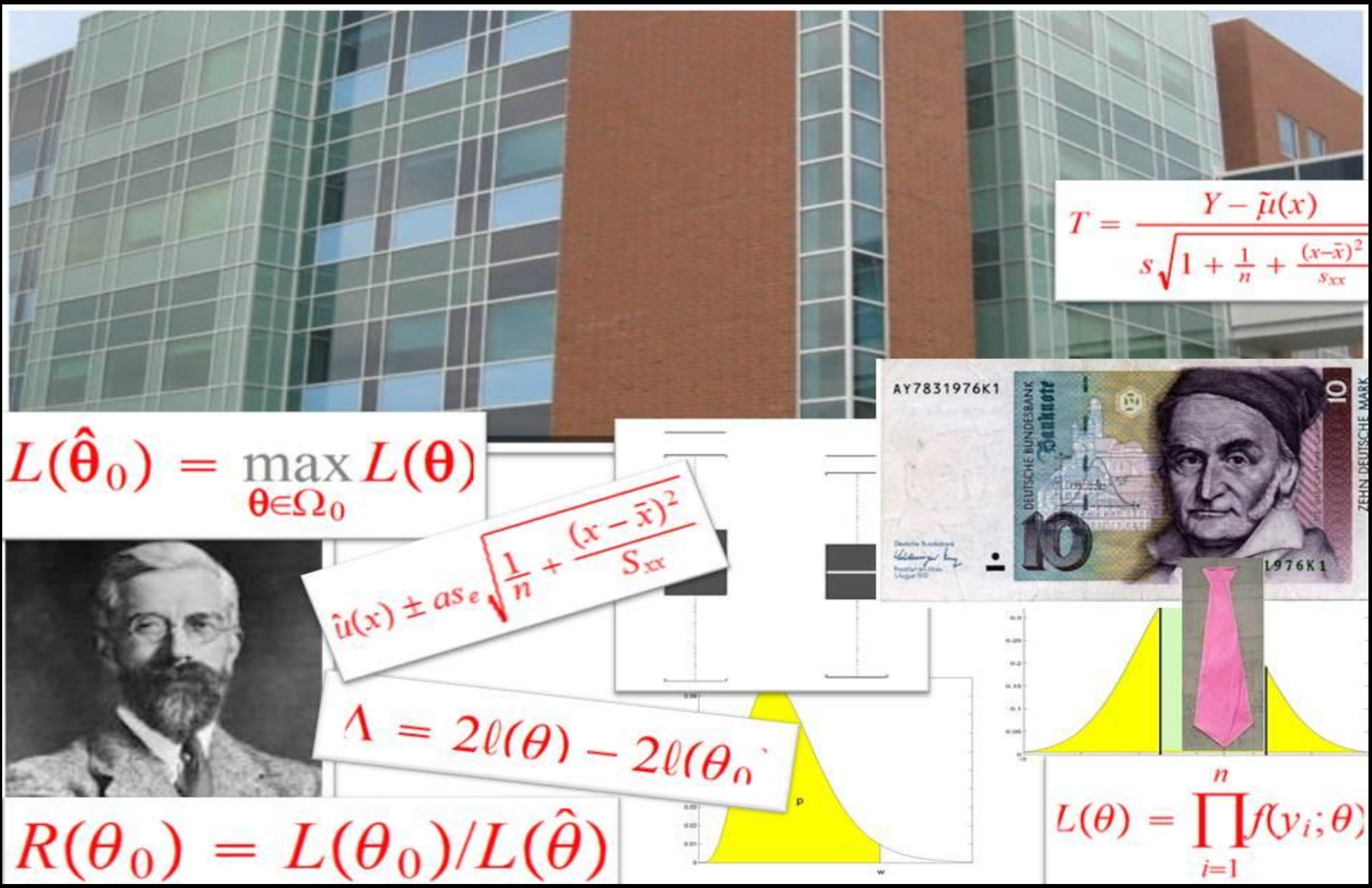
Contact SAS

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

STATISTICS 231 COURSE NOTES

WINTER 2013 EDITION

STATISTICS AND ACTUARIAL SCIENCE


$T = \frac{Y - \tilde{\mu}(x)}{s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$
$L(\hat{\boldsymbol{\theta}}_0) = \max_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta})$
$\hat{u}(x) \pm a s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$
AY7831976K1
DEUTSCHE BUNDESBANK
Banknote
10
ZEHN DEUTSCHE MARK
1976K1
$\Lambda = 2\ell(\theta) - 2\ell(\theta_0$
$R(\theta_0) = L(\theta_0)/L(\hat{\theta})$
$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$

Contents

Preface

These notes are a work-in-progress with contributions from those students taking the courses and the instructors teaching them. The original version of the notes was prepared by Jerry Lawless with additions and other editorial changes made by Jock MacKay, Don McLeish, Cyntha Struthers and others. Richard Cook furnished 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 dlmcleis@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 will reference these videos in the text.

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.

Empirical studies deal with *populations* and *processes*; both of which are collections of individual *units*. In order to increase our knowledge about a process, we examine a *sample* of units generated by the process. To study a population of units we examine a sample of units carefully selected from that population. Two challenges arise since we only see a sample from the process or population and not all of the units are the same. For example, scientists 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 variations in currency or stock values, variation in sales for a company over time, or variation 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 (i.e. informative, cost-effective) ways to collect and analyze data about such processes.

We can have various objectives when we collect and analyze data on 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 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. 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, Hypergeometric, 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 a 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. We finish this introduction with a recent quote from Hal Varien, Google's chief economist.

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

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

1.2 Collecting Data

The objects of study in this course are referred to as populations or processes. A *population* is a collection of *units*. For example, a population of interest may be all persons under the age of 18 in Canada as of September 1, 2012 or all car insurance policies issued by a company over a one year period. A *process* is a mechanism by which units are produced. For example, hits on a website constitute a process (the units are the distinct hits). Another process is the sequence of claims generated by car insurance policy holders (the units are the individual claims). 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. For example, variates can be *measured quantities* such as weight and blood pressure, *discrete quantities* such as the presence or absence of a disease or the number of damaged pixels in a monitor, *categorical quantities* such as colour or marital status, or more *complex quantities* 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. 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 of the claim or the response time to a hit in the processes mentioned above. 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 in the population or a sample of units taken from the process.

In planning to collect data about some process or population, 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. There are three broad approaches:

(i) **Sample Surveys** The object of many studies is to learn about a finite population (e.g. all persons over 19 in Ontario as of September 12 in a given year or all cars produced by the car manufacturer General Motors in the past calendar 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. 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 process or population 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** An experiment is a study in which the experimenter (i.e. the person conducting the study) intervenes and changes or sets the values of one or more variates on 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.

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

Example 1.2.1 A sample survey about smoking

Suppose we wish to study the smoking behaviour of Ontario residents aged 14-20 years. (Think about reasons why such studies are considered important.) Of course, the population of Ontario residents aged 14-20 years and their smoking habits both change over time, so we will content ourselves with a snapshot of the population at some point in time (e.g. the second week of September in a given year). Since we cannot afford to contact all persons in the population, we decide to select a sample of persons from the population of interest.

(Think about how we might do this - it is quite difficult!) We decide to measure the following variates on each person in the sample: age, sex, place of residence, occupation, current smoking status, length of time smoked, etc.

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

Example 1.2.2 A study of a manufacturing process

When a manufacturer produces a product in packages stated to weigh or contain a certain amount, they are generally required by law to provide at least the stated amount in each package. Since there is always some inherent variation in the amount of product which the manufacturing process deposits in each package, the manufacturer has to understand this variation and set up the process so that no packages or only a very small fraction of packages contain less than the required amount.

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

In the previous section, we noted that we collect data (consisting of measurements on variates $x, y, z, \ldots$ of interest for units in the sample) when we study a population or process. 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 for the questions of interest and, since they are summaries, we need to make sure that they are not misleading.

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-value 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.

There are two classes of summaries: graphical and numerical. First we describe some simple numerical summaries.

Numerical Summaries

Some common numerical summaries, useful for describing features of a single measured variate in a data set, are:

- the *average* (also called the sample average): $\bar{y} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$
- the *(sample) variance*: $s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(y_i - \bar{y})^2$
- the *(sample) standard deviation*: $s = \sqrt{s^2}$
- the *(sample) percentiles* and *(sample) quantiles*: the pth quantile (also called the 100pth percentile) is a value, call it $q(p)$, such that a fraction p of the y values in the data set are less than or equal to $q(p)$. The values $q(0.5)$, $q(0.25)$ and $q(0.75)$ are called the *median*, the *lower quartile*, and the *upper quartile* respectively. Depending on the size of the data set, quantiles are not uniquely defined for all values of p. For example, what is the median of the values $\{1, 2, 3, 4, 5, 6\}$? What is the lower quartile? There are different conventions for defining quantiles in these cases; if the sample size is large, the differences in the quantiles from the various definitions are small.

We can easily understand what the average, quantiles and percentiles tell us about the variate values in a data set. The variance and standard deviation measure the variability or spread of the variate values in a data set. We prefer the standard deviation because it has the same scale as the original variate. Another way to measure variability is to find the difference between a low and high quantile, for example the *interquartile range* $q(0.75) - q(0.25)$.

Example 1.3.1 Comparison of Body Mass Index

In a longitudinal study (i.e. 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. Their initial *Body Mass Index* (BMI) was also calculated. BMI is used to measure obesity or severely low weight. It is defined as follows:

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

There is some variation in what different guidelines refer to as “overweight”, “underweight”, etc. We present one such classification in Table 1.3.1

Table 1.3.1 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 stored in the file *ch1example131.txt* available on the course web page. For statistical analysis of the data, it is convenient to record the data in row-column format. Here are the first few rows of the file

Table 1.3.4 First Rows of the File ch1example131.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 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. See Appendix 2: Data for a listing of the file. We use the software package *R* (see Section 1.6 and Appendix 1) to get the following numerical summaries of the BMI variate for each sex.

Table 1.3.2 Summary of BMI by Sex

sex	First Quartile	Median	Average	Third Quartile	Sample Standard Deviation
Female	23.4	26.8	26.9	29.7	4.60
Male	24.7	26.7	27.1	29.1	3.56

From Table 1.3.2, we see that there are only small differences in any of the summary measures except for the standard deviation which is substantially larger for females. In other words, there is more variation in the BMI 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 1.3.3 BMI Relative Frequency Table by Sex

	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

From Table 1.3.3, we see that the reason for the larger standard deviation for females is that there is a greater proportion of females in the extreme classes.

Graphical Summaries

We consider several types of plots for a data set $\{y_1, y_2, \ldots, y_n\}$ of numerical values.

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 for $j = 1, \ldots, k$

$$f_j = \text{number of values from } \{y_1, \ldots, y_n\} \text{ that are in } \mathrm{I}_j.$$

The f_j are called the observed *frequencies* for $I_1, \ldots, I_k$; note that $\sum_{j=1}^{k} f_j = n$. A *histogram* is a graph in which a rectangle is placed above each interval; the height of the rectangle for I_j is chosen so that the area of the rectangle is proportional to f_j. Two main types of histogram are used. The second is preferred.

(a) a “standard" histogram where the intervals I_j are of equal length. The height of the rectangle is the frequency f_j. This type of histogram is similar to a bar chart.

(b) a “relative frequency” histogram, where the I_j may or may not be of equal length. The height of the rectangle for I_j is chosen so that its area equals f_j/n, the *relative frequency* for I_j. We use *density* as the label for the vertical axis. Note that in this case the sum of the areas of the rectangles in the histogram is equal to one.

We can make the two types of histograms visually comparable by using the same intervals and the same scaling on both axes. If the sample sizes in two groups differ, it is important to use relative frequencies and standardized scales for the axes. To construct a histogram, we have to choose the number and location of the intervals. The intervals are typically selected in such a way that each interval contains at least one y-value from the sample (that is, each $f_j \geq 1$). We can use software packages to produce histograms (see Section 1.6) and they will either automatically select the intervals for a given data set or allow the user to specify them.

Example 1.3.1 continued

In Figure 1.1, we see relative frequency histograms for BMI by sex. We often say that histograms show the *distribution* of the data, in this case for males and females. Here the *shape* of the two distributions is similar, each resembling a Gaussian distribution.

Example 1.3.2

A histogram can have many different shapes. Figure 1.2 shows a histogram of the lifetimes

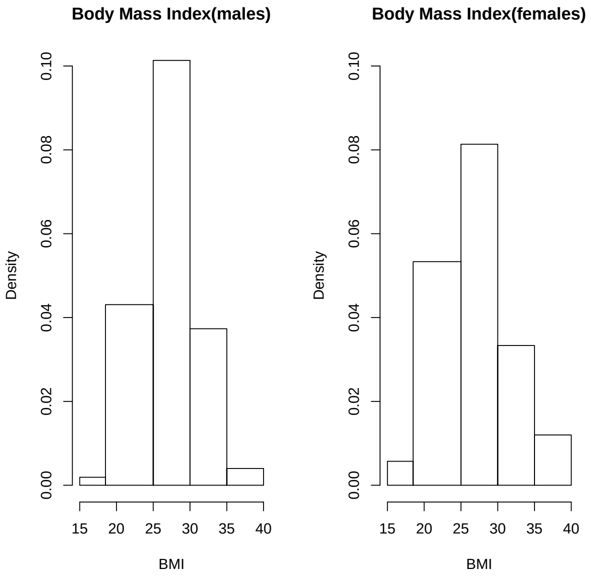


Figure 1.1: Histograms of BMI by sex

(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 *ch1example132.txt* available on the course web page and are listed in Appendix 2. Notice that the distribution has a very different shape compared to the BMI histograms. The brake pad lifetimes have a long right tail. The high degree of 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.

Cumulative frequency plots

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 more concisely the *empirical c.d.f.*

$$\hat{F}(y) = \frac{\text{number of values in } \{y_1, y_2, ..., y_n\} \text{ which are } \le y}{n}. \tag{1.1}$$

To construct $\hat{F}(y)$, it is convenient to first order the y_i's $(i = 1, \ldots, n)$ to give the ordered values $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$. Then, we note that $\hat{F}(y)$ is a step function with a jump at each of the ordered observed values $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$. 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 $1/n$. Between the observed variate values, the empirical cumulative distribution function is constant.

Example 1.3.4 Suppose that $n = 4$ and the y-values (ordered for convenience) are $\{1.5, 2.2, 3.4, 5.0\}$. Then

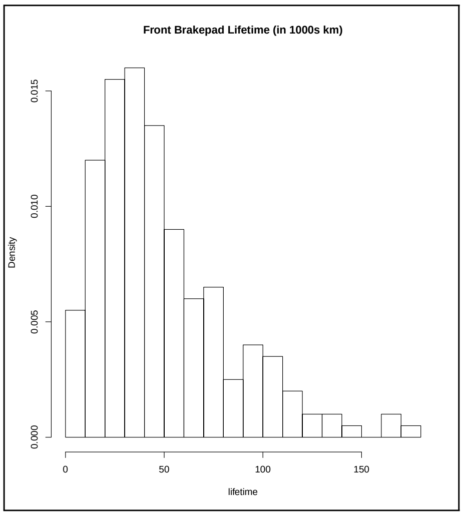


Figure 1.2: Histogram of Brake Pad Lifetimes

$$\hat{F}(y) = \begin{cases} 0 & y < 1.5 \\ 0.25 & 1.5 \leq y < 2.2 \\ 0.50 & 2.2 \leq y < 3.4 \\ 0.75 & 3.4 \leq y < 5.0 \\ 1.00 & y \geq 5.0 \end{cases}$$

Example 1.3.1 continued
Figure 1.3 shows the empirical cumulative distribution function for male and female heights overlaid on the same plot.The plot of the empirical cumulative distribution function does not show the shape of the distribution as clearly does the histogram. However, it shows us 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 m on the horizontal axis where $\hat{F}(m) = 0.5$). For example, we see from Figure 1.3 that the median height for females is about 1.60m and for males about 1.73m.

Box plots
In many situations, we want to compare the values of a variate for two or more groups, as in Example 1.3.1 where we compared BMI values and heights for males versus females. Especially when the number of groups is large (or the sample sizes within groups are small), side-by-side *box plots* are a convenient way to display the data. Box plots are also called *box and whisker plots.*

The box plot is (usually) displayed vertically.The center line in each box corresponds to

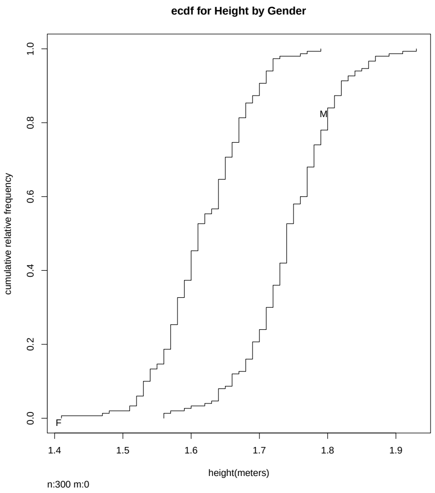


Figure 1.3: Empirical c.d.f. of height by sex

the median and the lower and upper sides of the box correspond to the lower quartile $q(0.25)$ and the upper quartile $q(0.75)$. The so-called whiskers extend down and up from the box to a horizontal line. The lower line is placed at the smallest variate value that is larger than the lower quartile minus 1.5 times the interquartile range, i.e. $q(0.25) - 1.5 \times [q(0.75) - q(0.25)]$. Similarly the upper line is placed at the largest variate value that is smaller than the upper quartile plus 1.5 times the interquartile range, i.e. $q(0.75) + 1.5 \times [q(0.75) - q(0.25)]$. Any values beyond the whiskers (often called outliers) are plotted as open circles.

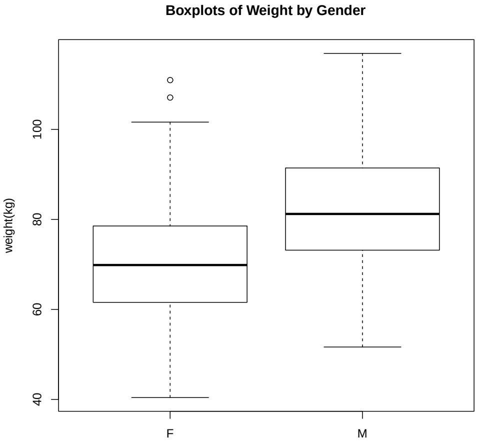


Figure 1.4: Boxplots of Weight by Sex

Figure 1.4 is side-by-side boxplots of male and female weights from Example 1.3.1. We can see for this sample that males are generally heavier than females but that the spread of the two distributions is about the same.

All of the numerical and graphical summaries discussed to this point deal with a single variate. We are often interested in relationships between two variates. A scatterplot can be used to demonstrate this relationship.

Scatterplots

Suppose we have data on two or more variates for each unit in the sample. For example, we might have the heights x and weights y for a sample of individuals. The data can then be represented as n pairs, $\{(x_i, y_i),\ i = 1, \ldots, n\}$ where x_i and y_i are the height and weight of the ith person in the sample.

When we have two such variables, a useful plot is a *scatterplot*, an $x-y$ plot of the points (x_i, y_i), $i = 1, \ldots, n$. The scatterplot shows whether x_i and y_i tend to be related in some way. Figure 1.5 is a scatterplot (with different symbols for males and females) of weight versus height for the data in Example 1.3.1. As expected, we see that there is a tendency for weight to increase as height increases for both sexes. What might be surprising is the variation in weights for heights that are close in value.

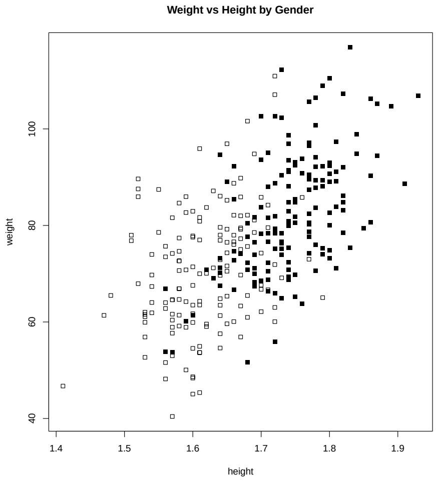


Figure 1.5: Scatterplot of Weight vs Height by Gender

1.4 Probability Distributions and Statistical Models

Probability 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 probability 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 (i.e. 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

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.)

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

Here the parameter θ represents the unknown proportion of smokers in the population, one attribute of interest in the study.

Example 1.4.2 An Exponential Distribution

In Example 1.3.2, 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.)

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

Here 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

$$E(Y) = \int_0^\infty y f(y; \theta) dy = \theta.$$

To model the sampling procedure, we let $Y_1, \ldots, Y_{200}$ be 200 independent copies of Y. We can use the model 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

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

If we model the selection of a data set $(y_1, \ldots, y_n)$ as the realization of n independent copies 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 average $\bar{y}$ corresponds to μ, the expected value of Y
- the sample median corresponds to the solution m of the equation $F(m) = 0.5$ where $F(y) = P(Y \leq y)$ is the cumulative distribution function of Y.
- the sample standard deviation corresponds to σ, the standard deviation of Y, where $\sigma^2 = E[(Y - \mu)^2]$
- the histogram (with the y-axis on the density scale) corresponds to the probability density function of Y

Example 1.4.2 Gaussian Distributions

Earlier, we described an experiment where the goal was to see if there is a relationship between a measure of 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 $(y_1, x_1), \ldots, (y_{40}, x_{40})$. A model for $Y_1, \ldots, Y_{40}$ should depend on the temperatures x_i and one possibility is to let $Y_1, ..., Y_{40}$ be independent random variables with Y_i having the Gaussian distribution $G(\beta_0 + \beta_1 x_i, \sigma)$, $i = 1, \ldots, 40$. In this model, the mean of Y is a linear function of the temperature x_i. The parameter σ allows for variation in performace among chips operating at the same temperature. We will consider such models based on Gaussian random variables in Chapter 6.

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, prediction problems* and *hypothesis testing 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.

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.

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.

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 1.5.1, 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 a two-way frequency table such as the following:

	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 proportion of girls who smoke is $82/250 = 32.8\%$ and the corresponding proportion for males is $71/250 = 28.4\%$. 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 $\hat{\theta} = 153/500 = 30.6\%$. Then using the model and the estimate, we calculate the probability of such a large difference in the observed rates. In this case, we would see such a large difference about 20% of the time (if we selected samples over and over and the hypothesis is true) so there is no evidence of a difference in smoking rates. We examine the logic and details of such a formal procedure in Chapter 5.

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. Suppose that every hour, one can is selected from the new machine and one can from the old machine over a period of 40 hours. You can find measurements of the amounts of liquid in the cans in the file *ch1example152.txt* and also listed in Appendix 1. The variates (column headings) are hour, machine (new = 1, old = 2) and volume (ml). We display the first few rows of the file below. The complete file is listed in the Appendix.

Hour	Machine	Volume
1	1	357.8
1	2	358.7
2	1	356.6
2	2	358.5
3	1	357.1
3	2	357.9

First we examine if the behaviour of the two machines is stable over time. In Figure 1.6, we show a *run chart* of the volumes over time for each machine. 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 average 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. In Figure 1.7 we show side-by-side histograms of the volumes. Since the histograms are "bell-shaped", we also overlaid Gaussian probability density functions with the mean equal to the average and standard deviation equal to the sample standard deviation.

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 models to estimate the long term

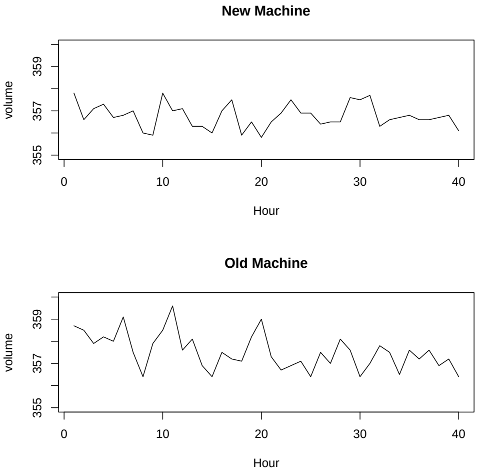


Figure 1.6: Run Charts of Volume by Machine

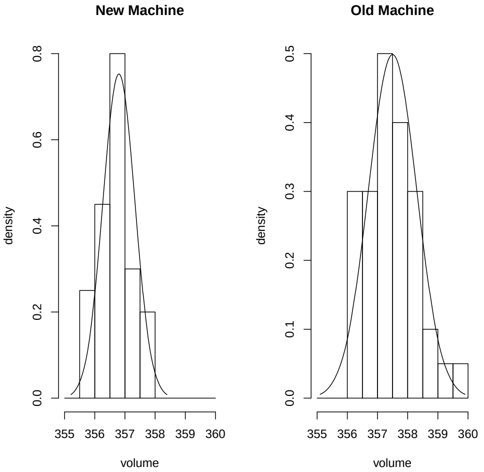


Figure 1.7: Histograms of Volume by Machine

proportion of cans that fall below the required volume. For the new machine, we find that if $V \sim G(356.8, 0.53)$ then $P(V \leq 355) = 0.0003$ so about 3 in $10,000$ cans will be underfilled. The corresponding rate for the old machine is about 9 in $10,000$ cans. 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.

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

1.6 Statistical Software

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 a s EXCEL is also useful.

In this course we use the *R* software system. It is an open source package that has extensive statistical capabilities and very good graphics procedures. The R home page is www.r-project.org where a free download is available for most common operating systems.

Some of the basics of *R* are described in the Appendix at the end of this chapter. We use *R* for several purposes: to manipulate and graph data, to fit and check statistical models, to estimate attributes or test hypotheses, to simulate data from probability models. All of the calculation and plots in this chapter were made with *R*.

1.7 Appendix 1: Using *R*

Lots of help is available in *R*. You can use a search engine to find the answer to most questions. For example, if you search for "*R* tutorial", you will find a number of excellent introductions to *R* that explain how to carry out the above list of tasks. Within *R*, you can find help for a specific function using the command help(function name) but it is often easier to look externally using a search engine.

Here we show how to use *R* on a Windows machine. You should have *R* open as you read this material so you can play along.

Some R Basics

R is command-line driven. For example, if you want to define a quantity x , use the assignment function $< -$ (i.e. $<$ followed by $-$).

$$x < -15$$

or, (a slight complication)

$$x < -c(1, 3, 5)$$

so x is a column vector with elements $1, 3, 5$.

A few general comments

- If you want to change x, you can up-arrow to return to the assignment and make the change you want, followed by a carriage return.
- If you are doing something more complicated, you can type the code in Notepad or some other text editor (Word is not advised!) and cut and paste the code into *R*.
- You can save your session and, if you choose, it will be restored the next time you open *R*.
- You can add comments by entering # with the comment following on the same line.

Vectors

Vectors can consist of numbers or other symbols; we will consider only numbers here. Vectors are defined using the function $c\,()$. For example,

$$x < -c(1, 3, 5, 7, 9)$$

defines a vector of length 5 with the elements given. You can display the vector by typing x and carriage return. Vectors and other objects possess certain attributes. For example, typing

$$length(x)$$

will give the length of the vector x.

You can cut and paste comma- delimited strings of data into the function $c\,()$. This is one way to enter data into *R*. See below to learn how you can read a file into *R*.

Arithmetic

R can be used as a calculator. Enter the calculation after the prompt > and hit return as shown below.

```
 > 7+3
[1] 10
> 7*3
[1] 21
> 7/3
[1] 2.333333
> 2^3
[1] 8
```

You can save the result of the calculation by assigning it to a variable such as y<-7+3

Some Functions

There are many functions in *R*. Most operate on vectors in a transparent way, as do arithmetic operations. (For example, if x and y are vectors then $x + y$ adds the vectors element-wise; if x and y are different lengths, *R* may do surprizing things! Some examples, with comments, follow.

```
> x<- c(1,3,5,7,9)   # Define a vector x
> x            # Display x
[1] 1 3 5 7 9
> y<- seq(1,2,.25)    #A useful function for defining a vector whose
                      elements are an arithmetic progression
> y
[1] 1.00 1.25 1.50 1.75 2.00
> y[2]   # Display the second element of vector y
[1] 1.25
> y[c(2,3)]    # Display the vector consisting of the second and
                  third elements of vector y.
[1] 1.25 1.50
> mean(x)      #Computes the average of the elements of vector x
[1] 5
> summary(x)    # A useful function which summarizes features of
                  a vector x
 Min. 1st Qu. Median Mean 3rd Qu. Max.
    1       3      5    5       7    9
> sd(x)    # Computes the (sample) standard deviation  of the elements of x
[1] 10
> exp(1)    # The exponential function
[1] 2.718282
> exp(y)
[1] 2.718282 3.490343 4.481689 5.754603 7.389056
> round(exp(y),2)   # round(y,n) rounds the elements of vector y to
                       n decimals
[1] 2.72 3.49 4.48 5.75 7.39
> x+2*y
[1]  3.0  5.5  8.0 10.5 13.0
```

As we have seen we often want to compare summary statistics of variate values by group (such as sex). We can use the *by*() function. For example,

```
> y<-rnorm(100)   # y is a vector of length 100 with entries generated at
                 # random from G(0,1)
> x<-c(rep(1,50),rep(2,50))   # x is a vector of length 100 with 50 1s
```

```
                              # followed by 50 2s.
> by(y,x, summary)  # generates a summary for the elements of y for each
                              #value of the grouping variable x
```

We can replace the function summary() by most other simple functions.

Graphs

Note that in *R*, a graphics window opens automatically when a graphical function is used. A useful way to create several plots in the same window is the function *par*() so, for example, following the command

```
par(mfrow=c(2,2))
```

the next 4 plots will be placed in a 2 × 2 array within the same window.

There are various plotting and graphical functions. Three useful ones are

```
plot(y~x)  # Gives a scatterplot of y versus x; thus x and y must
              be vectors of the same length.

hist(y)    # Creates a frequency histogram based on the values in
              the vector y. To get a relative frequency histogram
              (areas of rectangles sum to one) use hist(x,prob=T).
boxplot(y~x)   #Creates side-by-side boxplots of the values of y
               # for each value of x.
```

You can control the axes of plots (especially useful when you are making comparisons) by including $xlim = c(a,b)$ and $ylim = c(d,e)$ as arguments separated by commas within the plotting function. Also you can label the axes by including $xlab =$ “*yourchoice*” and$ylab =$ “*yourchoice*”. A title can be added using $main =$ “*yourchoice*”. There are many other options. Check out the Html help “An Introduction to *R*” for more information on plotting.

To save a graph, you can copy and paste into a Word document for example or alternately use the “Save as” menu to create a file in one of several formats.

Probability Distributions

There are functions which compute values of probability functions or probability density functions, cumulative distribution functions, and quantiles for various distributions. It is also possible to generate random samples from these distributions. Some examples follow for the Gaussian distribution. For other distributions, type *help*(*distributionname*) or check the “Introduction to *R*” in the Html help menu.

```
> y<- rnorm(10,25,5)    # Generate 10 random values from the Gaussian
```

```
                  # distribution G(25,5) and store the values in the vector y.
> y    # Display the values
 [1] 22.50815 26.35255 27.49452 22.36308 21.88811 26.06676 18.16831 30.37838
 [9] 24.73396 27.26640
> pnorm(1,0,1)   # Compute P(Y<=1) for a G(0,1) random variable.
[1] 0.8413447
> qnorm(.95,0,1)   # Find the .95 quantile (95th percentile) for G(0,1).
[1] 1.644854
>dnorm(2,1,3)    # calculates the probability density function at y=2 for Y~G(1,3)
[1] 0.1257944
```

Reading data from a file

R stores and retrieves data from the current working directory. You can use the command

```
getwd()
```

to determine the current working directory. To change the working directory, look in the File menu for "*changedir*" and browse until you reach your choice. There are many ways to read data into *R*. The files we used in Chapter 1 are in .txt format with the variate labels in the first row separated by spaces and the corresponding variate values in subsequent rows. We created the files from EXCEL by saving as text files. To read such files, first be sure the file is in your working directory. Then use the commands

```
a<-read.table('filename.txt',header=T)  #enclose the filename in single quotes
attach(a)
```

The "header=T" tells *R* that the variate names are in the first row of the data file. The object a is called a data frame in *R* and the variate names are of the form "$a : v1$" where $v1$ is the name of the first column in the file. The *R* function *attach*(a) allows you to drop the a : from the variate names.

Writing data to a file.

You can cut and paste output generated by *R* in the sessions window although the format is usually messed up. This approach works best for Figures. You can write an *R* vector or other object to a text file through

```
      write(y,file="filename")
```

`To see more about the write function use help(write).`

Example 1.5.2 Two Filling Machines

Here we list the data in Appendix 2: Data of these notes. Here is the *R* code used in Example 1.5.2. In the file *ch1example152.txt*, there are three columns labelled hour, machine and volume. The data are

hour	machine	volume	hour	machine	volume
1	1	357.8	21	1	356.5
1	2	358.7	21	2	357.3
2	1	356.6	22	1	356.9
.	.	.	.	.	.

And here is the *R* code we used.

```
# read data
a<-read.table('ch1example152.txt',header=T)
attach(a)

# calculate summary statistics and standard deviation by machine
by(volume,machine,summary)
by(volume,machine,sd)

# separate the volumes by machine into separate vectors v1 and v2
v1<-volume[seq(1,79,2)] # picks out machine 1 values
v2<-volume[seq(2,80,2)]  # picks out machine 2 values
h<-1:40
```

```
# plot run charts by machine, one above of the other, type='l'joins the points on the plots
```

```
par(mfrow=c(2,1)) # creates two plotting areas, one above the other
plot(v1~h,xlab='Hour',ylab='volume',main='New Machine', ylim=c(355,360),type='l')
plot(v2~h,xlab='Hour',ylab='volume',main='Old Machine', ylim=c(355,360),type='l')
```

```
# plot side by side histograms, overlay gaussian densities for each machine

par(mfrow=c(1,2)) #creates two plotting areas side by side

br<-seq(355,360,0.5) # defines interval endpoints for the histograms
hist(v1,br,freq=F,xlab='volume',ylab='density',main='New Machine')
w1<-356.8+0.538*seq(-3,3,.01) # values where the gaussian density is located
dd1<-dnorm(w1,356.8,0.53)
```

```
points(w1,dd1,type='l')

hist(v2,br,freq=F, xlab='volume',ylab='density',main='Old Machine')
w2<-357.5+0.799*seq(-3,3,.01)
dd2<-dnorm(w2,357.5,0.8)
points(w2,dd2,type='l')
```

1.8 Problems

1. The average and the sample median are two different ways to describe the location or the center of a data set $(y_1, y_2, \ldots, y_n)$. In this exercise we look at some of their properties. Let $\bar{y}$ be the average and m be the median of the data set.

   (a) Suppose we change the location and scale of the data so that $u_i = a + by_i$ for every $i = 1, ..., n$ where a and b are constants with $b \neq 0$. How do the average and sample median change?

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

   (c) Consider the quantities $r_i = y_i - \bar{y}$, $i = 1, \ldots, n$. Show that $\sum\limits_{i=1}^{n} r_i = 0$. Is it true that $\sum\limits_{i=1}^{n} (y_i - m) = 0$?

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

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

   (f) Use the above results to explain why the sample median income of a country might be a more appropriate summary than the average income.

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

   (h) Consider the function $W(\mu) = \sum\limits_{i=1}^{n} |y_i - \mu|$. Show that $W(\mu)$ is minimized when $\mu = m$. 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 $(y_1, y_2, \ldots, y_n)$. Recall that the sample standard deviation is

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1.2}$$

(a) Suppose we change the location and scale of the data so that $u_i = a + by_i$ for every $i = 1, \ldots, n$ where a and b are constants and $b \neq 0$. How do the sample standard deviation and interquartile range change?

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

(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 interquartile range change as y_0 gets large?

3. Mass production of complicated assemblies such as automobiles depend on our ability to manufacture the components to very tight specifications. The component manufacturer tracks performance by measuring a sample of parts and compaing 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 (also available in the file *ch1exercise3.txt*) 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)

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

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

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

(c) Calculate the sample standard deviation s and the interquartile range.

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

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

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

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

(f) 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.

4. In the above exercise, 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 quantites. Then we can use *R* to generate another 50 observations and recalculate *Ppk*. We do this many times. Here is some *R* code. Make sure you replace XX with the appropriate values. average<- XX Replace XX by the observed average from the above question sd<- XX Replace XX with the observed standard deviation temp<-rep(0,1000) creates a vector of length 1000 to hold the *Ppk* values that we generate for (i in 1:1000) starts a loop y<-rnorm(50, average, sd) generates 50 new observations using the model with the appropriate mean and sd avg<-mean(y);s<-sd(y) calculates the average and sd of the data ppk<-min((10-avg)/(3*s),(avg+10)/(3*s)) calculates *Ppk* temp[i]<-ppk stores the value of *Ppk* and loops for 1000 iterations hist(temp) makes a histogram of the *Ppk* values mean(temp) calculates the average *Ppk* value sd(temp) calculates the standard deviation of the *Ppk* values

   (a) Based on the analysis, how reliable is the estimate produced by the initial sample

   (b) Repeat the above exercise but this time use a sample of 300 pistons. Has the reliability of the estimate increased? Why?

5. The data below show the lengths (in cm) of 20 male and female coyotes captured in Nova Scotia. The data are available in the file *ch1exercise5.txt*

Females

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

Males

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

   (a) Plot relative frequency histograms of the lengths for females and males using *R*. Make sure the scales and bins are the same.

   (b) Compute the sample average $\bar{y}$ and sample standard deviation s for the female and male coyotes separately. Assuming $\mu = \bar{y}$ and $\sigma = s$, plot the probability density function for Gaussian distributions $G(\mu, \sigma)$ over top of the histograms for the females and males. (Based on Table 2.3.2 in Wild and Seber 1999)

MODEL FITTING, MAXIMUM LIKELIHOOD ESTIMATION, AND MODEL CHECKING

2.1 Statistical Models and Probability Distributions

A statistical model is a mathematical model that incorporates probability[1] 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 the 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. Chapter 3 discusses these issues. You have already seen some examples in Chapter 1, and have been reminded of material on random variables and probability distributions which you have seen in a previous course on probability.

A preliminary step in probability and statistics is the choice of a probability model 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. Current data set, against which models can be assessed.

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,

[1] 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.

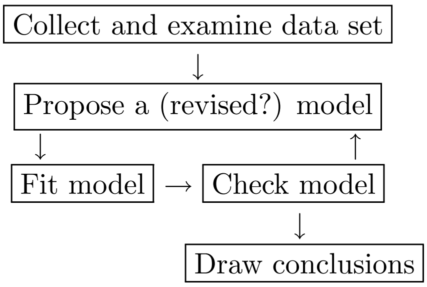
the Binomial distribution was derived as a model for outcomes in repeated independent trials with two possible outcomes on each trial while the Poisson distribution was derived as a model for the random occurrence of events in time or space. The Gaussian or Normal distribution, on the other hand, is 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 let us suppose that for some discrete random variable Y we consider a family whose probability function depends on the parameter θ (which may be a vector of values):

$$P(Y = y; \theta) = f(y; \theta) \text{ for } \ y \in A$$

where A is a countable (i.e. discrete) set of real numbers, the *range* of the random variable Y. In order to apply the model to a specific problem we require a value for θ; the selection of a value based on the data (let us call it $\hat{\theta}$) is often referred to as "fitting" the model or as "estimating" the value of θ. The next section decribes a method for doing this.

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. However, in this course 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 frequently in these notes. See the Stat 230 Notes available on the course webpage. You should also consult the Table of Distributions at the end of these notes for a condensed table of properties of these distributions uncluding their moment generating functions and their moments.

Binomial Distribution

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

$$P(Y = y; \theta) = f(y; \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } y = 0, 1, \ldots, n \tag{2.2}$$

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 (also called a Normal) distribution if its probability density function is of the form

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

where μ and σ are parameters, with $-\infty < \mu < \infty$ 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, $N(\mu, \sigma^2)$, we specify the variance σ^2 for the parameter. 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 size 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 common 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 $Y_1, \ldots, Y_k \quad (k \geq 2)$ have the joint probability function

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

where each y_i, for $i = 1, \ldots, k$, is an integer between 0 and n, and satisfying the condition $\sum_{i=1}^{k} y_i = n$. The elements of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ satisfy $0 < \theta_i < 1$ for $i = 1, \ldots, k$, 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 $1, \ldots, k$), and the probability outcome i occurs is θ_i. If Y_i, $i = 1, \ldots, k$ is the number of times that outcome i occurs in a sequence of n independent trials, then $(Y_1, \ldots, Y_k)$ have the joint probability function above. We write $(Y_1, \ldots, Y_k) \sim \text{Multinomial}(n; \boldsymbol{\theta})$.

Since $\sum_{i=1}^{k} Y_i = n$ we can rewrite $f(y_1, \ldots, y_k; \boldsymbol{\theta})$ using only $k-1$ variables, say $y_1, \ldots, y_{k-1}$ by replacing y_k with $n - y_1 - \ldots - y_{k-1}$. 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 will also consider models that include explanatory variables, or covariates. For example, suppose that the response variable Y is 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, so we might want to study this relationship. A way to do this is to consider females with a given height x (say in meters), and to propose that the distribution of Y, given x is Gaussian, $G(\alpha + \beta x, \sigma)$. That is, we are proposing that the average (expected) weight of a female depends linearly on her height x and we write this as

$$E(Y|x) = \alpha + \beta x$$

Such models are considered in Chapters 6-8.

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, μ and σ in the Gaussian model or θ in an Exponential model).

2.2 Estimation of Parameters (Model Fitting)

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, 10 females from the population, measure their weights $y_1, y_2, \ldots, y_{10}$, and use the average,

$$\hat{\mu} = \bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i \tag{2.3}$$

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 $y_1, \ldots, y_{10}$ 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 $\mu = \frac{1}{10}\sum\limits_{i=1}^{10} y_i$. We introduced a special notation $\hat{\mu}$. This serves a dual purpose, both to remind you that $\frac{1}{10}\sum\limits_{i=1}^{10} y_i$ 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, $i = 1, 2, \ldots, 10$ and *is therefore a random quantity.* A different draw of the sample y_i, $i = 1, 2, \ldots, 10$ will result in a different value for $\hat{\mu}$.

Instead of ad hoc approaches to estimation as in (2.3), 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.

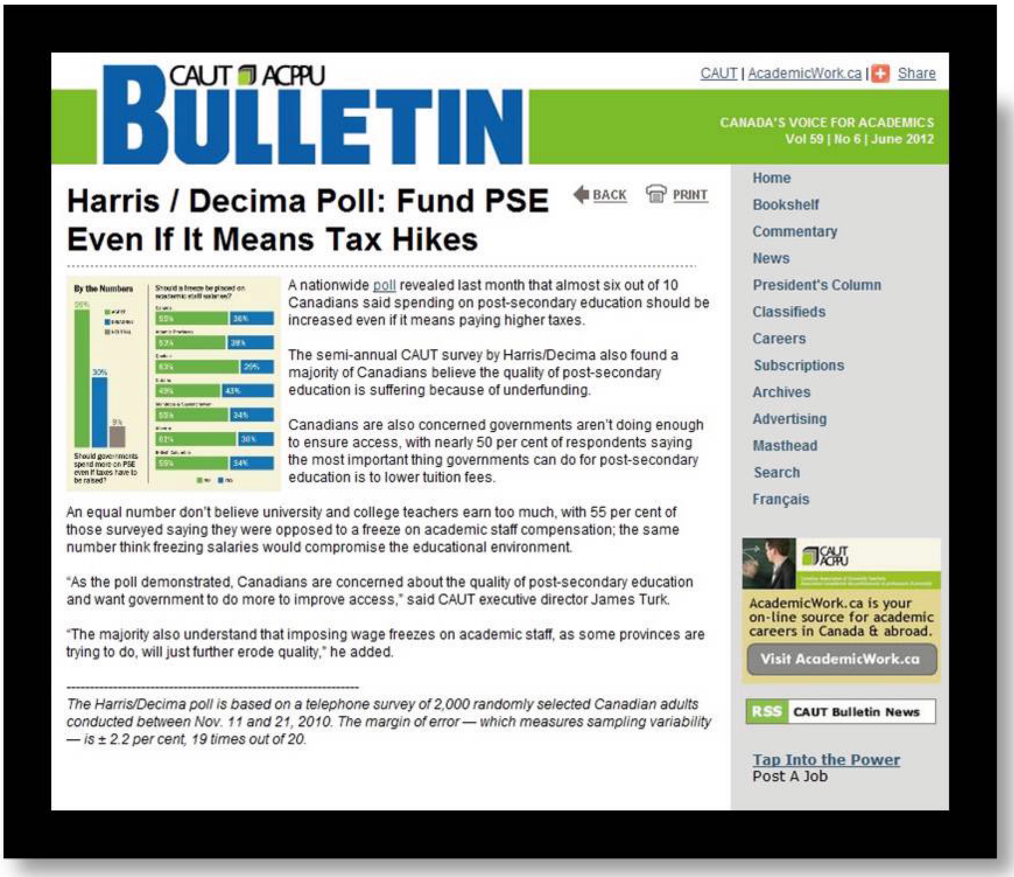
Let the discrete (vector) random variable $\mathbf{Y}$ represent potential data that will be used to estimate θ, and let $\mathbf{y}$ represent the actual observed data that are obtained in a specific application. Note that to apply the method of maximum likelihood, we must know (or make assumptions 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 population units. The *likelihood function* for θ is then defined as

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

where the *parameter space* Ω is the set of possible values for θ. Note that the likelihood function is a function of the parameter θ and the given data $\mathbf{y}$. For convenience we usually write just $L(\theta)$. Also, *the likelihood function is the probability that we observe at random the observation* $\mathbf{y}$, *considered as a function of the parameter* θ. Obviously values of the parameter that make our observation $\mathbf{y}$ more probable would seem more credible or likely than those that make it less probable. Therefore values of θ for which $L(\theta)$ is large are more consistent with the observed data $\mathbf{y}$. The value $\hat{\theta}$ that maximizes $L(\theta)$ for given data $\mathbf{y}$ is called the *maximum likelihood estimate* [2] (m.l. estimate) of θ. This seems like a "sensible" approach, and it turns out to have very good properties.

Example 2.2.1 (a public opinion poll)[3].

We are surrounded by polls. They guide the policies of our political leaders, the products that are developed by manufacturers, and increasingly the content of the media. For example the poll[4] in Figure 2.2 was conducted by Harris/Decima company under contract


CAUT ACPPU
BULLETIN
CAUT | AcademicWork.ca | Share
CANADA'S VOICE FOR ACADEMICS
Vol 59 | No 6 | June 2012
BACK
PRINT
Home
Bookshelf
Commentary
News
President's Column
Classifieds
Careers
Subscriptions
Archives
Advertising
Masthead
Search
Français
By the Numbers
Should a freeze be placed on academic staff salaries?
55%
36%
57%
38%
63%
29%
49%
43%
55%
34%
61%
30%
59%
34%
59%
30%
9%
Should governments spend more on PSE even if taxes have to be raised?
AcademicWork.ca is your on-line source for academic careers in Canada & abroad.
Visit AcademicWork.ca
RSS CAUT Bulletin News
Tap Into the Power
Post A Job

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

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.

Figure 2.2: The CAUT Bulletin

of the CAUT (Canadian Association of University Teachers). This is a semi-annual poll on Post-Secondary Education and Canadian Public Opinion. The poll above was conducted

[2] 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.*

[3] See the corresponding video "harris decima poll and introduction to likelihoods" at www.watstat.ca

[4] http://www.caut.ca/uploads/Decima_Fall_2010.pdf

in November 2010. Harris/Decima uses a telephone poll of 2000 "representative" adults. Twenty-six percent of respondents agreed and 48% disagreed with the following statement: "University and college teachers earn too much".

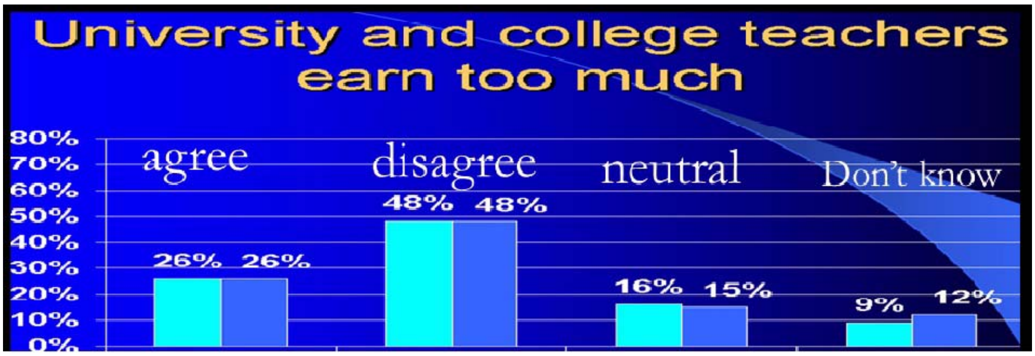


Figure 2.3: Results of the Harris/Decima poll. The two bars are from polls conducted Nov. 9, 2010 and Nov 10, 2010 respectively.

Harris/Decima declared their result to be accurate within ±2.2 percent, 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. It is assumed that Y is closely modelled by a Binomial distribution:

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

where θ represents the fraction of the entire population that agree. In this case, if we select a random sample of n persons and obtain their views we have $\mathbf{Y} = Y$, and the observed data is $\mathbf{y} = y = 520$, the number out of 2000 who were polled that agreed with the statement. Thus the likelihood function is given by

$$L(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } 0 < \theta < 1 \tag{2.4}$$

or for this example

$$\binom{2000}{520} \theta^{520} (1-\theta)^{2000-520} \quad \text{for } 0 < \theta < 1. \tag{2.5}$$

It is easy to see that (2.4) is maximized by the value $\hat{\theta} = y/n$. (You should show this.) For this example the value of this maximum likelihood estimate is 520/2000 or 26%. This is also easily seen from a graph of the likelihood function (2.5) given in Figure 2.4. From the graph it can also be seen that the interval suggested by the pollsters, $26 \pm 2.2\%$ or [23.8, 28.2] is a reasonable interval for the parameter θ since it seems to contain most of

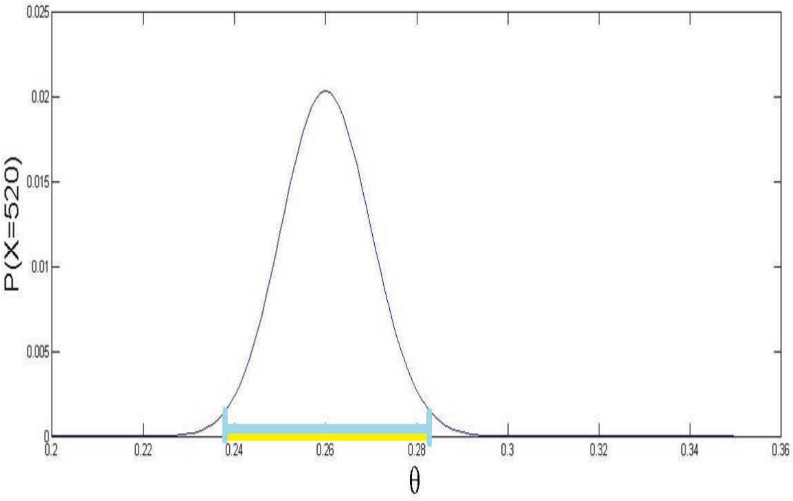


Figure 2.4: Likelihood function for the Harris/Decima poll and corresponding interval estimate for θ

the values of θ with large values of the likelihood $L(\theta)$. We will return to the construction of such interval estimates later.

Example 2.2.2

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. Again assume that Y is modelled by a Binomial distribution:

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

where θ represents the fraction 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

$$L(\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \quad \text{for } 0 < \theta < 1 \tag{2.6}$$

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

Note that the likelihood function's basic properties, for example, where its maximum occurs and its shape, are not affected if we multiply $L(\theta)$ 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 is with the parameter θ_1 versus θ_2. The ratio $L(\theta_1)/L(\theta_2)$ is also unaffected if we multiply $L(\theta)$ by a constant. In view of this we might define the likelihood as $P(\mathbf{Y} = \mathbf{y}; \theta)$ or any constant multiple of it, so, for example, we could drop the term $\binom{n}{y}$ in (2.6) and define $L(\theta) = \theta^y(1-\theta)^{n-y}$. This function and (2.6) 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. This rescaled version is called the **relative likelihood function**

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

It is also convenient to define the **log likelihood function**,

$$\ell(\theta) = \log L(\theta) \quad \text{for } \theta \in \Omega.$$

Note that $\hat{\theta}$ also maximizes $\ell(\theta)$. (Why?) Because functions are often (but not always!) maximized by setting their derivatives equal to zero[5], we can usually obtain $\hat{\theta}$ by solving the equation

$$\frac{d\ell}{d\theta} = 0.$$

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

$$\frac{d\ell}{d\theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta}.$$

Solving $d\ell/d\theta = 0$ gives $\hat{\theta} = y/n$.

In many applications the data set $\mathbf{Y}$ are assumed to consist of a random sample $Y_1, \ldots, Y_n$ from some process or population, where each Y_i has the probability function (or probability density function) $f(y;\theta)$, $\theta \in \Omega$. In this case $\mathbf{y} = (y_1, \ldots, y_n)$ and

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

(You should recall from probability that if $Y_1, \ldots, Y_n$ are independent random variables then their joint probability function is the product of their individual probability functions.)

In addition, if for estimating θ we have two data sets $\mathbf{y}_1$ and $\mathbf{y}_2$ from two independent studies, 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 $L_j(\theta) = P(\mathbf{Y}_j = \mathbf{y}_j; \theta)$, $j = 1, 2$.

[5] 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 setting $f'(x) = 0$?

Likelihood for Continuous Distributions.

Recall that we defined likelihoods for discrete random variables as the probability of the observed values, 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 is always zero. However in the continuous case, we define likelihood similarly but with the probability function $P(\mathbf{Y} = \mathbf{y}; \theta)$ replaced by the joint probability density function evaluated at the observed values. For *independent* observations Y_i, $i = 1, 2, ..., n$ from the same probability density function $f(y; \theta)$, the joint probability density function of $(Y_1, Y_2, ..., Y_n)$ is

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

Consequently this is used in this context for the likelihood function. For n independent observations $y_1, y_2, ..., y_n$ from a continuous probability density function $f(y; \theta)$, the likelihood function is defined as

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

Example 2.2.3

Suppose that the random variable Y represents the lifetime of a randomly selected light bulb in a large population of bulbs, and that Y follows an Exponential distribution with probability density function

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

where $\theta > 0$. If a random sample of light bulbs is tested and the lifetimes $y_1, \ldots, y_n$ are observed, then the likelihood function for θ is, from (2.7),

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

Thus

$$\ell(\theta) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} y_i$$

and solving $d\ell/d\theta = 0$, we obtain

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

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

Example 2.2.2 revisited.

Sometimes the likelihood function for a given set of data can be constructed in more than one way. For the random sample of n persons who are tested for HIV, for example, we could define

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

for $i = 1, \ldots, n$. (Note: $I(A)$ is the indicator function; it equals 1 if A is true and 0 if A is false.) In this case the probability function for Y_i is Binomial$(1; \theta)$ with

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

and the likelihood function is

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

where $y = \sum\limits_{i=1}^{n} y_i$. This is the same likelihood function as we obtained in Example 2.2.1, if we use the fact that $Y = \sum\limits_{i=1}^{n} Y_i$ has a Binomial distribution, Binomial(n, θ).

Example 2.2.4

As an example involving more than one parameter, suppose that the random variable Y has a Gaussian distribution with probability density function

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

The random sample $y_1, ..., y_n$ then gives, with $\boldsymbol{\theta} = (\mu, \sigma)$,

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

where $-\infty < \mu < \infty$ and $\sigma > 0$. Thus

$$\ell(\boldsymbol{\theta}) = \ell(\mu, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 - (n/2)\log(2\pi).$$

We wish to maximize $\ell(\mu, \sigma)$ with respect to both parameters μ and σ. Solving [6] the two equations[7]

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0\\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)^2 = 0,\end{aligned}$$

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

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}\\ \hat{\sigma} &= \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}.\end{aligned}$$

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

Example 2.2.5

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

$$P(Y = y; \theta) = f(y; \theta) = \frac{(\theta v_i)^y}{y!}e^{-\theta v_i} \quad \text{for } y = 0, 1, \ldots \tag{2.8}$$

where θ is the average number of bacteria per millilitre (ml) 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 (2.8),

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

Suppose that n water samples, of volumes $v_1, \ldots, v_n$, are selected. Let $z_1, \ldots, z_n$ be the observed values of the presence indicators. The likelihood function is then

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

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

[7] 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 σ while holding μ constant.

and the log likelihood function is

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

We cannot maximize $\ell(\theta)$ mathematically by solving $d\ell/d\theta = 0$, 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

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

Either by maximizing $\ell(\theta)$ numerically for $\theta > 0$, or by solving $d\ell/d\theta = 0$ numerically, we find the maximum likelihood estimate of θ to be $\hat{\theta} = 0.478$. A simple way to maximize $\ell(\theta)$ is to plot it, as shown in Figure 2.5; the maximum likelihood estimate can then be found by inspection or, for more accuracy, by iteration using Newton's method[8].

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.

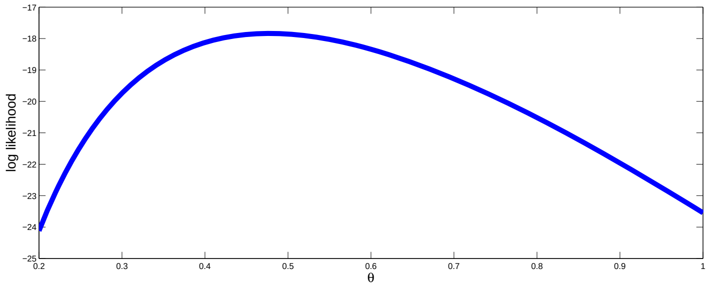


Figure 2.5: **The log likelihood function $\ell(\theta)$ for Example 2.2.**5

[8] You should recall this from your calculus course

2.3 Likelihood Functions For Multinomial Models

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

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

If the θ_i's are to be estimated from data involving n "trials", of which y_i resulted in outcome i, $i = 1, \ldots, k$, then it seems obvious that

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

would be a sensible estimate. This can also be shown to be the maximum likelihood estimate for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.[9]

Example 2.3.1

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 1 to 4 were $y_1 = 172$, $y_2 = 38$, $y_3 = 14$ and $y_4 = 176$ (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 we might get 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 $\theta_1 = 0.448$, $\theta_2 = 0.083$, $\theta_3 = 0.034$, $\theta_4 = 0.436$.)

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

Example 2.3.2

[9] The log likelihood can be taken as (dropping the $n!/(y_1! \ldots y_k!)$ term for convenience) $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{k} y_i \log \theta_i$. This is a little tricky to maximize because the θ_i's satisfy a linear constraint, $\sum \theta_i = 1$. The Lagrange multiplier method (Calculus III) for constrained optimization allows us to find the solution $\hat{\theta}_i = y_i/n$, $i = 1, \ldots, k$.

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. 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 we would estimate α from a random sample of size n giving y_1, y_2 and y_3 persons of types MM, MN and NN by using the likelihood function

$$\begin{aligned} L(\alpha) &= \frac{n!}{y_1!y_2!y_3!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3} \\ &= \frac{n!}{y_1!y_2!y_3!}[\alpha^2]^{y_1}[2\alpha(1-\alpha)]^{y_2}[(1-\alpha)^2]^{y_3} \\ &= c\alpha^{2y_1+y_2}(1-\alpha)^{y_2+2y_3} \quad \text{for } 0<\alpha<1 \ \text{ where } c = \frac{n!}{y_1!y_2!y_3!}2^{y_2}. \end{aligned}$$

Example 2.4.2 in the next section considers some data for this setting.

2.4 Checking Models

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 $y_1,\ldots,y_n$ of y-values from the population or process.

The probability model is supposed to represent the relative frequency of sets of y-values in large samples, so a fundamental check is to compare model probabilities and relative frequencies for a sample. Recall the definition of a histogram in Section 1.3 and let the range of Y be partitioned into intervals $I_j = [a_{j-1}, a_j), \quad j = 1,\ldots,k$. From our model $f(y;\theta)$ we can compute the values

$$\hat{p}_j = P(a_{j-1} \le Y < a_j; \hat{\theta}) \quad \text{for } j = 1,\ldots,k.$$

If the model is suitable, these values should be "close" to the observed relative frequencies $r_j = f_j/n$ in the sample. (Recall that f_j is the number of y-values in the sample that are in the interval I_j). This method of comparison works for either discrete or continuous random variables. An example of each type follows.

Example 2.4.1

Suppose that an Exponential model for a positive-valued continuous random variable Y has been proposed, with probability density function

$$f(y) = 0.01e^{-0.01y} \quad \text{for } y > 0 \tag{2.9}$$

and that a random sample of size $n = 20$ has given the following values $y_1, \dots, y_{20}$ (rounded to the nearest integer):

10	32	15	26	157	99	109	88	39	118
61	104	77	144	338	72	180	63	155	140

For illustration purposes, let us partition $[0, \infty)$, the range of Y, into four intervals $[0, 30)$, $[30, 70)$, $[70, 140)$,$[140, \infty)$. The probabilities $\hat{p}_j$, $j = 1, \dots, 4$ from the model (2.9) are given by

$$\hat{p}_j = \int_{a_{j-1}}^{a_j} 0.01e^{-0.01y} dy = e^{-0.01a_{j-1}} - e^{-0.01a_j}$$

and we find $\hat{p}_1 = 0.261$, $\hat{p}_2 = 0.244$, $\hat{p}_3 = 0.250$, $\hat{p}_4 = 0.247$, (the numbers add to 1.002 and not 1.0 because of round-off error). The relative frequencies $r_j = f_j/20$ from the random sample are $r_1 = 0.15$, $r_2 = 0.25$, $r_3 = 0.30$, $r_4 = 0.30$. These agree fairly well with the model-based values $\hat{p}_j$, but we might wonder about the first interval. We discuss how "close" we can expect the agreement to be following the next example. With a sample of this small a size, the difference between r_1 and $\hat{p}_1$ represented here does **not** suggest that the model is inadequate.

This example is an artificial numerical illustration. In practice we usually want to check a family of models for which one or more parameter values is unknown. When parameter values are unknown we first estimate them using maximum likelihood, and then check the resulting model. The following example illustrates this procedure.

Example 2.4.2

In Example 2.3.2 we considered a model from genetics in which the probability a person is blood type MM, MN or NN is $\theta_1 = \alpha^2$, $\theta_2 = 2\alpha(1-\alpha)$, $\theta_3 = (1-\alpha)^2$, respectively. Suppose a random sample of 100 individuals gave 17 of type MM, 46 of type MN, and 37 of type NN.

The relative frequencies from the sample are $r_1 = 0.17$, $r_2 = 0.46$, $r_3 = 0.37$, where we use the obvious "intervals" $I_1 = \{\text{person is MM}\}$, $I_2 = \{\text{person is MN}\}$, $I_3 = \{\text{person is NN}\}$. (If we wish, we could also define the random variable Y to be 1, 2, 3 according to whether a person is MM, MN or NN.) Since α is unknown, we must estimate it before we can check the family of models given above. From Example 2.3.2, the likelihood function for α from the observed data (ignoring the constant c) is

$$L(\alpha) = [\alpha^2]^{17}[\alpha(1-\alpha)]^{46}[(1-\alpha)^2]^{37} \quad \text{for } 0 < \alpha < 1.$$

Collecting terms, we find

$$\ell(\alpha) = \log L(\alpha) = 80 \log \alpha + 120 \log(1 - \alpha)$$

and $d\ell/d\alpha = 0$ gives the maximum likelihood estimate $\hat{\alpha} = 0.40$. The model-based probabilities for I_1, I_2, I_3 are thus

$$\hat{\theta}_1 = \hat{\alpha}^2 = 0.16, \ \hat{\theta}_2 = 2\hat{\alpha}(1 - \hat{\alpha}) = 0.48, \ \hat{\theta}_3 = (1 - \hat{\alpha})^2 = 0.36$$

and these agree quite closely with $r_1 = 0.17$, $r_2 = 0.46$, $r_3 = 0.37$. On this basis the model seems satisfactory.

The method above suffers from some arbitrariness in how the I_j's are defined and in what constitutes "close" agreement between the model-based probabilities $\hat{p}_j$ and the relative frequencies $r_j = f_j/n$. Some theory that provides a formal comparison will be given later in Chapter 7, but for now we will just rely on the following simple guideline. If we consider the random variables F_j, $j = 1, \ldots, k$ corresponding to the observed frequencies f_j then $(F_1, \ldots, F_k)$ have a Multinomial$(n; p_1, \ldots, p_k)$ distribution, where p_j is the "true" value of $P(a_{j-1} \le Y < a_j)$ in the population. In addition, any single F_j has a Binomial(n, p_j) distribution. This means we can assess how variable either f_j or $r_j = f_j/n$ is likely to be, in a random sample. From the Central Limit Theorem, if n is large enough, the distribution of F_j is approximately normal, $N(np_j,\, np_j(1 - p_j))$. It follows that

$$P\left(np_j - 1.96\sqrt{np_j(1 - p_j)} \le F_j \le np_j + 1.96\sqrt{np_j(1 - p_j)}\right) \approx 0.95$$

and thus (dividing by n and rearranging)

$$P\left(-1.96\sqrt{\frac{p_j(1 - p_j)}{n}} \le R_j - p_j \le 1.96\sqrt{\frac{p_j(1 - p_j)}{n}}\right) \approx 0.95 \tag{2.10}$$

where $R_j = F_j/n$. This allows us to get a rough idea for what constitutes a large discrepancy between an observed relative frequency r_j and a true probability p_j. For example when $n = 20$ and p_j is about 0.25, as in Example 2.4.1, we get from (2.10) that

$$P\left(-0.19 \le R_j - p_j \le 0.19\right) \approx 0.95$$

so it is quite common for r_j to differ from p_j by up to 0.19. The discrepancy between $r_1 = 0.15$ and $p_1 = 0.261$ in Example 2.4.1 is consequently not unusual and does not suggest the model is inadequate.

For larger sample sizes, r_j will tend to be closer to the true value p_j. For example, with $n = 100$ and $p_j = 0.5$, (2.10) gives

$$P\left(-0.10 \le R_j - p_j \le 0.10\right) \approx 0.95.$$

Thus in Example 2.4.2, there is no indication that the model is inadequate. (We are assuming here that the model-based values $\hat{p}_j$ are like the true probabilities as far as (2.10) is concerned. This is not quite correct but (2.10) will still serve as a rough guide. We are also ignoring the fact that we have picked the largest difference between r_j and $\hat{p}_j$, as the Binomial distribution is not quite correct either. Chapter 7 shows how to develop checks of the model that get around these points.)

Graphical Checks

A graph that compares relative frequencies and model-based probabilities provides a nice picture of the "fit" of the model to the data. Two plots that are widely used are based on histograms and the empirical cumulative distribution function $\hat{F}(y)$ which were both discussed in Chapter 1.

The histogram plot for a continuous random variable Y is as follows. Plot a relative frequency histogram of the random sample $y_1, \ldots, y_n$ and superimpose on this a plot of the probability density function $f(y;\theta)$ for the proposed model. The area under the probability density function between values a_{j-1} and a_j equals $P(a_{j-1} \leq Y < a_j)$ so this should agree well with the area of the rectangle over $[a_{j-1}, a_j)$. The plots in Figure 1.7 for the can-filling data in Chapter 1 are of this type.

For a discrete random variable Y we plot a probability histogram for the probability distribution $f(y;\theta)$ and superimpose a relative frequency histogram for the data, using the same intervals I_j in each case.

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, $P(Y \leq y; \theta) = F(y;\theta)$. If the model is suitable, the two curves should not be too far apart. An illustration is given in the next example.

Example 2.4.3

For the data on female heights in Chapter 1 and using the results from Example 2.2.4 we obtain $\hat{\mu} = 1.62$, $\hat{\sigma} = 0.0637$ as the maximum likelihood estimates of μ and σ. Figure 2.6 shows (a) a relative frequency histogram for these data with the $G(1.62, 0.0637)$ probability density function superimposed and (b) a plot of the empirical cumulative distribution function with the $G(1.62, 0.0637)$ cumulative distribution 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 - probability density function 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. Neither plot suggests strongly that the Gaussian model is unsatisfactory. Both plots were created using R.

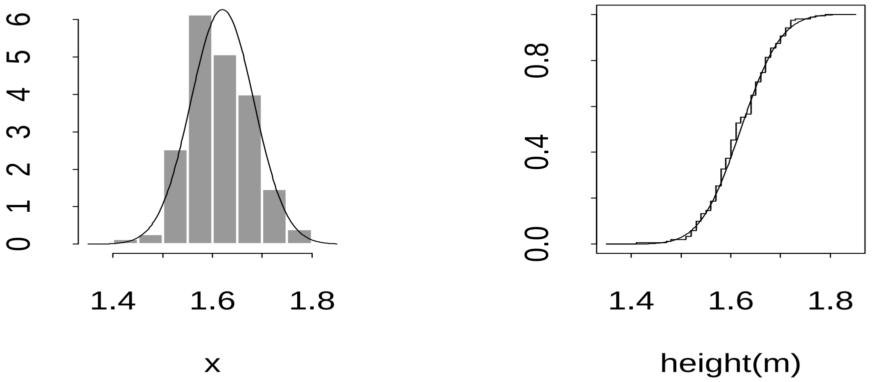


Figure 2.6: **Model and Data Comparisons for Female Heights**

2.5 Problems

1. 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.)

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 probability of the observed results (as a function of θ) for the two experiments. Thus obtain the likelihood function for θ for each experiment and show that they are proportional. Show the maximum likelihood estimate $\hat{\theta}$ is the same in each case. What is the maximum likelihood estimate of θ?

(b) Suppose n people came to a blood donor clinic. Assuming $\theta = 0.10$, how large should n be to ensure that the probability of getting 10 or more B- type donors is at least 0.90? (The R functions *gbinom()* or *pbinom()* can help here.)

3. Consider Example 2.3.2 on M-N blood types. If a random sample of n individuals gives y_1, y_2, and y_3 persons of types MM, MN, and NN respectively, find the maximum likelihood estimate $\hat{\alpha}$ in the model in terms of y_1, y_2, y_3.

4. 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

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

(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 does this give if $n = 50$ with $n_1 = 16$, $n_2 = 16$, $n_3 = 18$?

(c) Does the model appear to fit the data well?

5. 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?

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

$$\begin{aligned} P(Y = k; \alpha) &= \alpha^k \qquad\qquad k = 1, 2, \ldots \\ P(Y = 0; \alpha) &= \tfrac{1-2\alpha}{1-\alpha} \end{aligned}$$

Here α is an unknown parameter with $0 < \alpha < \frac{1}{2}$.

(a) Suppose that n families are selected at random and that f_y is the number of families with y children ($f_0 + f_1 + \ldots = n$). Determine the maximum likelihood estimate of α.

(b) Consider a different type of sampling wherein a single child is selected at random and the size of family the child comes from is determined. Let Y represent the number of children in the family. Show that

$$P(Y = y; \alpha) = cy\alpha^y \quad \text{for } y = 1, 2, \ldots$$

and determine c.

(c) Suppose that the type of sampling in part (b) was used and that with $n = 33$ the following data were obtained:

y:	1	2	3	4
f_y:	22	7	3	1

Determine the maximum likelihood estimate of α. Also estimate the probability a couple has no children.

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

7. 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 instead of knowing the y_i's, we know only whether or not there was one or more particles emitted in each time interval. Making a suitable assumption, give the likelihood function for θ based on these data, and describe how you could find the maximum likelihood estimate of θ.

8. **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 then as follows:

- k (where $0 \leq k \leq n$) pieces fail, at times $y_1, \ldots, y_k$.
- $n - k$ pieces are still working after time C.

(a) If Y has an Exponential(θ) distribution, show that

$$P(Y > C; \theta) = e^{-C/\theta} \quad \text{for } C > 0.$$

(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 400 hours, when the test was terminated, with $\sum_{i=1}^{5} y_i = 450$ hours. Find the maximum likelihood estimate of θ.

9. **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 $\theta(x) = \exp(\alpha + \beta x)$, 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?

10. 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?

11. The course web page has data on the lifetimes of the right front disc brakes pads for a specific car model. The lifetimes y are in km driven, and correspond to the point at which the brake pads in new cars are reduced to a specified thickness. The data on $m = 92$ randomly selected cars are contained in the file brakelife.text.

(a) Assuming a $G(\mu, \sigma)$ model for the lifetimes, determine the maximum likelihood estimates of μ and σ based on the data. How well does the Gaussian model fit the data?

(b) Another model for such data is given by

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left[-\frac{1}{2}\left(\frac{\log y - \mu}{\sigma}\right)^2\right], \quad \text{for } y > 0.$$

(Note: Show using methods you learned in your course on probability that if $X \backsim G(\mu, \sigma)$ then $Y = \log X$ has the probability density function given above.) Using this model determine the maximum likelihood estimates of μ and σ based on the data. How well does this model fit the data? Which of the two models describes the data better?

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 any systematic way. It is the object of this chapter to do that. Well-conducted studies are needed to produce maximal information within existing cost and time constraints. Conversely, a poorly planned or executed study can be worthless or even misleading.

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. Throughout the course, we will develop each of the five steps in more detail and see many applications of PPDAC's use. We identify the steps in the following example.

Example 3.1
The following newspaper article was taken from the Kitchener-Waterloo Record, December

1, 1993. It describes an empirical investigation in the field of medicine. There are thousands of studies in this field every year conducted 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.

K-W Record, December 1, 1993

Fats raise risk of lung cancer in non-smokers

WASHINGTON (AP) – Add lung cancer to the growing list of diseases that seem to be influenced by diets high in fat. A study by the U.S. National Cancer Institute of non-smoking women in Missouri found that those who eat diets with 15 percent or more saturated fat are about six times more likely to develop lung cancer than those whose meals have 10 per cent or less of fat. "We found that as you increase the amount of saturated fat, you increase the amount of lung cancer," Michael Alavanja, an institute researcher, said Tuesday. A report on the study is to be published Friday in the Journal of the National Cancer Institute. Earlier studies have linked high-fat diets with cancers of the colon, prostate and breast. High-fat diets also are thought to increase the risk of heart disease. Alavanja said his research compared the diets of 429 non-smoking women who had lung cancer with the diets of 1021 non-smoking women who did not have lung cancer. The women all lived in Missouri, were of about the same age and represented "a typical American female population." The women filled out forms that asked about their dietary habits. They were then divided into 5 groups based on the amount of fat and other nutrients they consumed. Alavanja said the study found that those with diets with the lowest amount of saturated fat and the highest amount of fruits, vegetables, beans and peas were the least likely to develop lung cancer. At the other end of the scale, 20 per cent of the women in the study with the highest consumption of fat and diets lowest in fruits, vegetables, beans and peas had about six times more lung cancer. For a specific type of lung cancer, adenocarcinoma, there was an 11-fold difference between those on lowest-fat diets and those on the highest-fat diets. Adenocarcinoma is a form of lung cancer that is less often associated with smoking. "The leading contributors of dietary saturated fat were hamburgers, cheeseburgers, and meat loaf ... followed by weekly consumption of cheeses and cheese spreads, hot dogs, ice cream and sausages," the study said. Alavanja said that these foods, by themselves cannot be considered good or bad, but that they appear to create a lung cancer risk with the represent 15 percent or more of the calories in the diet.

Here are the five steps:

- **Problem**

  Does a high level of dietary saturated fat cause an increased risk of lung cancer in non-smoking women?

  Plan

  Find a set of non-smoking women with lung cancer and another set of non-smoking women without lung cancer. The women are to be comparable in age. Measure the level of saturated fat in the diet of each individual woman.

  Data

  Collect dietary fat levels for 429 lung cancer patients and 1021 other women.

  Analysis

Define five categories of dietary fat intake (low to high) and determine the number of women in the study that fall in each category. Then calculate the proportion of women in each category that have lung cancer.

Conclusion

The lowest level of fat intake category has the lowest rate of lung cancer and the highest level of fat intake category has the highest rate of lung cancer. It appears that increased dietary fat intake increases the risk of lung cancer in non-smoking women.

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.

You will learn to use PPDAC 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, govenrment policy statements and various business reports. If you see the phrase "evidence based decison" or "evidence based management", look for an empirical study.

In the rest of this chapter, we discuss the steps of PPDAC in more detail. To do so we introduce a numer of technical terms. Every subject has its 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. We have written the new terms in italics where they are defined.

3.2 The 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?

The first step is to define the *units* and the *target population* or *target process.* 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. In this example 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.

Note that the target population is a collection of units. 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?

We define a *variate* as a characteristic of every 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.

We specify the questions in terms of attributes of the target population. An *attribute* is a function of the variates over the target population. In the smoking example, one important attribute is the proportion of teenagers in the target population. In the can-filling example, we are interested in 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.

In most cases, we cannot calculate the attributes of interest directly because we can only examine a *sample* 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.

We will later consider a special class of problems that have a *causative aspect*. These problems are common and of critical importance. For example, we might ask questions such as :

"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 compulsory driver training reduce the incidence of accidents among new drivers?"

All of these questions are about causation. Each corresponding Problem has a causative aspect. We will see in Chapter 8 how we must be exceedingly careful in the Plan and Analysis of an empirical study in order to answer such causative questions.

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

3.3 The Plan

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 must depend on the output from the Problem step.

We begin with the terms *study units* and *study population.* The study units are those available to be included in the study. In most cases, the study units are elements of the target population (as in the teenage smoking survey) but, for example, in many medical applications, we must use animals as study units when the target population consists of people. The study population or study process is the set of study units that could possibly be included in the investigation. 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. In many cases (but not always!) the study population is a subset of the target population. For example, 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.

We noted above that the study population is usually not identical to the target population. The attributes of interest in the study population may differ from those specified in the Problem and we call this difference *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.

As part of the Plan, we specify the *sampling protocol* which is the procedure we use to select a sample of units from the study population. In other words we need to determine how we will select the sample. In the 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 rigourous methods are used. Often we "take what we can get".

Even with random sampling, we are looking at only a subset of the units in the study population and hence the sample attributes may well differ from those in the study population. We call this difference *sampling error.* 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 sampling 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 also need to determine the *sample size*, i.e. the number of study units sampled from the study population. 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.

We must decide which variates we are going to measure or determine for the units in the sample. For any attibutes 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 detemined for a given unit, errors are often introduced by the measurement system which determines the value. The observed value and the "true" value are usually not identical and we call the unknown difference between the measured and true value the *measurement error*. These errors then become part of the analysis. In practice, we need to ensure that the measurement systems used do not contribute substantial error to our Conclusions. We may have to study the measurement systems used separately to ensure that this is so.

Figure 3.2 shows the steps in the Plan and the sources of error

A person using PPDAC for an empirical study should, by the end of the Plan stage, 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 alos pay attention to the various types of error that may occur and how they might impact the conclusions.

3.4 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.

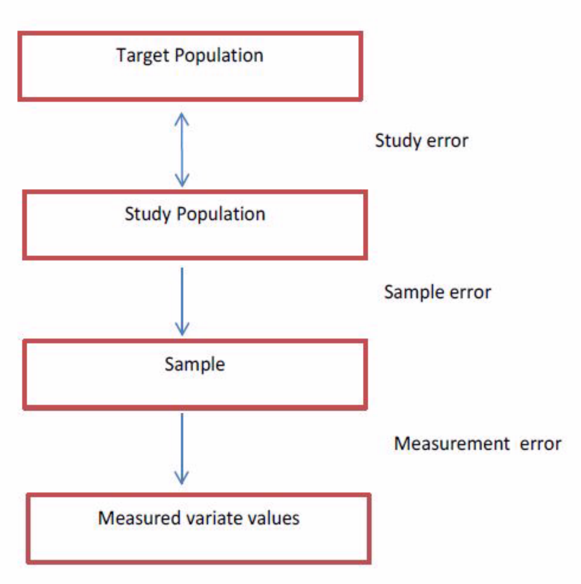


Figure 3.2: Steps in the plan

- 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.

3.5 Analysis and Conclusion

In Chapters 1 and 2, we discussed both formal and informal analysis methods. A key step in formal analyses is the selection of an appropriate model that can describe the data and how we collected it. 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.

In the Conclusion step, we answer the questions posed in the Problem. In other words, the Analysis and Conclusion are directed by the Problem. We also try to quantify (or at least discuss) potential errors as described in the Plan step.

We end this chapter with a case study that demonstrates the use of PPDAC.

3.6 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 3.3 shows the diameter in question.

The journal is a "cylindrical" part of the crankshaft. The diameter of the journal must be defined since the cross-section of the journal is not perfectly round and there may be taper along the axis of the cylinder. The gauge measures the maximum diameter as the crankshaft is rotated at a fixed distance from the end of the cylinder.

The specification for the diameter is -10 to $+10$ units with a target of 0. The measurements are re-scaled automatically by the gauge to make it easier to see deviations from the target. If the measured diameter is less than -10, the crankshaft is scrapped and a cost is incurred. If the diameter exceeds $+10$, the crankshaft can be reworked, again at considerable cost. Otherwise, the crankshaft is judged acceptable.

Overall Project

A project is planned 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

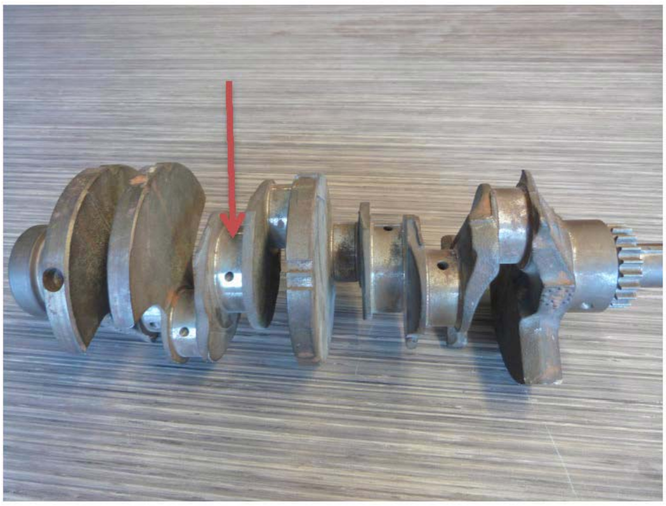


Figure 3.3: Crankshaft with arrow

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 (note that a unit is the act of taking a measurement, not the result) 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 3.4. In such a diagram, we list *explanatory variates* organized by the major "bones" that might be responsible for variation in the response variate, here the measured journal diameter. We can use the diagram in formulating the Plan.

Note that the measurement system includes the gauge itself, the way the part is loaded into the gauge, who loads the part, the calibration procedure (every two hours, a master part is put through the gauge and adjustments are made based on the measured diameter of the master part; that is “the gauge is zeroed”), and so on.

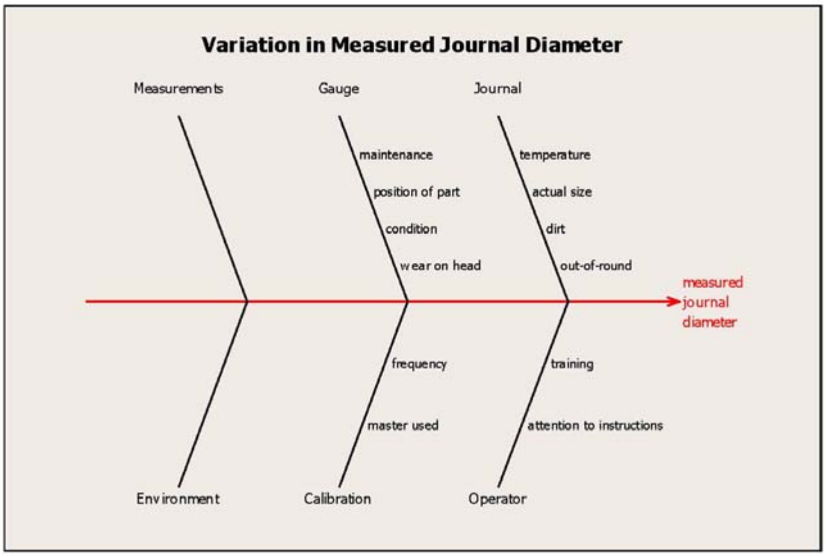


Figure 3.4: Fishbone diagram

Plan

To determine the properties of the measurement errors we must measure crankshafts with known diameters. “Known” implies that the diameters were measured by an off-line measurement system that is very reliable. For any measurement system study in which bias is an issue, there must be a reference measurement system which is known to have negligible bias and variability which is much smaller than the system under study.

There are many issues in establishing a study process or a study population. For convenience, we want to conduct the study quickly using only a few parts. However, this restriction may lead to study error if the bias and variability of the measurement system change as other explanatory variates change over time or parts. We guard against this latter possibility by using three crankshafts with known diameters as part of the definition of the study process. Since the units are the taking of measurements, we define the study population as all measurements that can be taken in one day on the three selected crankshafts. These crankshafts were selected so that the known diameters were spread out over the range of diameters normally 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.)

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

$$Y_{ij} = \mu_i + R_{ij}, \quad R_{ij} \sim G(0, \sigma_m) \quad \text{independent} \tag{3.2}$$

where $i = 1$ to 3 indexes the three crankshafts and $j = 1, \ldots, 10$ 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 2.2.4 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 (3.2). Recall that if $Y_{ij} \backsim G(\mu_i, \sigma_m)$ then

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

Therefore if we repeatedly measure the same journal diameter, then about 95% of the time we would expect to see the observations vary by about $\pm 2(1.37) = \pm 2.74$.

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 (sampling error). Since the measurements were not taken in random order, another source of sampling 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 3.5. 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.

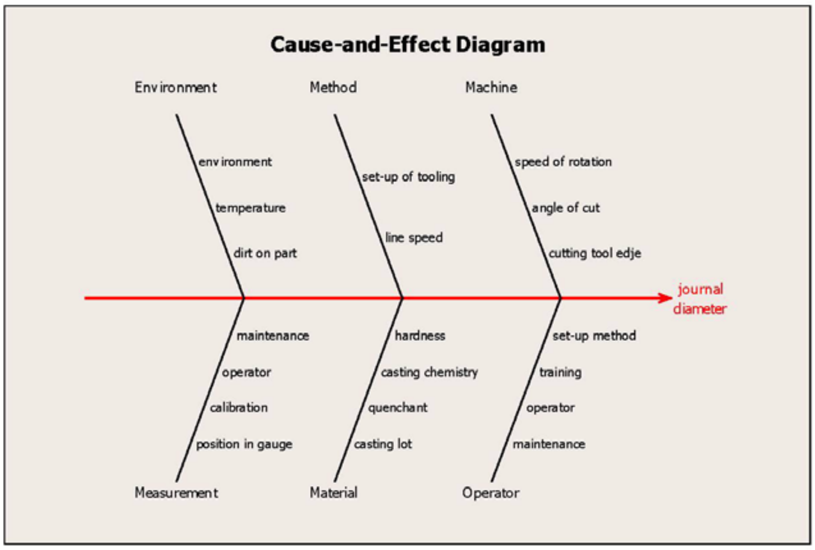


Figure 3.5: Fishbone diagram

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 (sampling 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 3.6. From this, we can see that the measured diameters vary from -14 to $+16$.

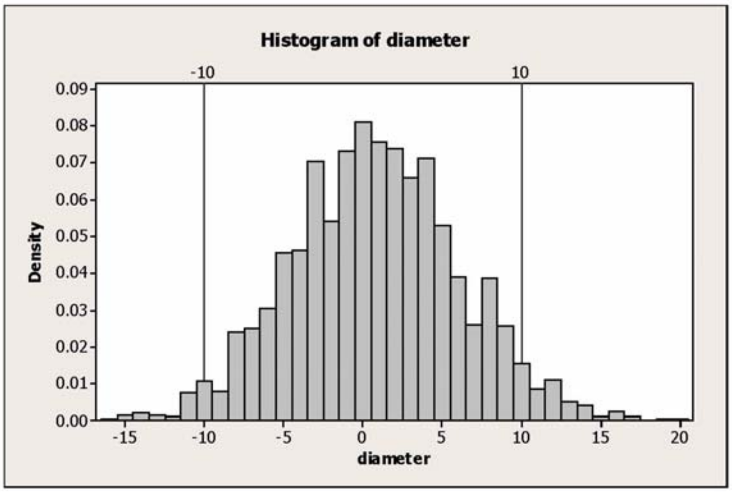


Figure 3.6: Histogram of 2000 measured values from the gauge memory

Analysis

A model for these data is given by

$$Y_i = \mu + R_i, \quad R_i \sim G(0, \sigma) \quad \text{independently for } i = 1, ..., 2000$$

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, 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 sampling 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. The measurments versus the times of measureurement are plotted in Figure 3.7 using a plot called a run chart. In the first study the standard deviation was estimated to be 1.37. In a sample of observations from a $G(0, 1.37)$ distribution we would expect approximately 95% of the observations to lie in the interval $[0 - 2(1.37), \ 0 + 2(1.37)] = [-2.74, \ 2.74]$ which is obviously not true for the data displayed in the run chart. These data have a much larger variability. This was a shocking result for the people in charge of the process.

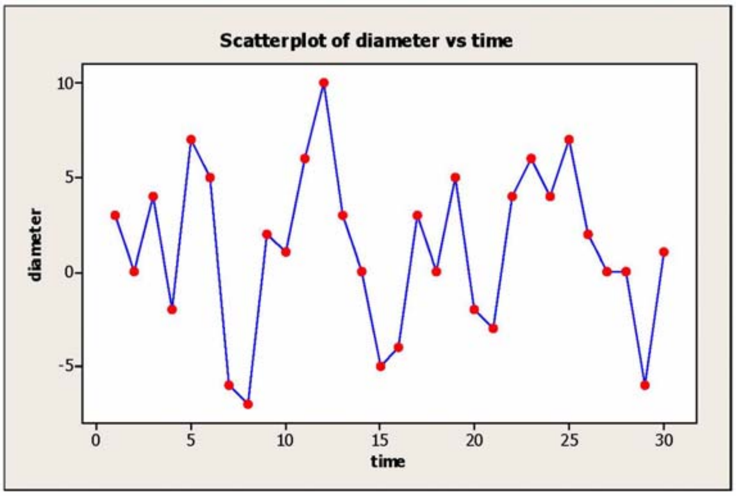


Figure 3.7: Scatter plot of diameter versus time

Comments

Study 3 revealed that the measurement system had a serious long term problem. At first, it was suspected that the cause of the variability was the fact that the gauge was not calibrated over the course of the study. Study 3 was repeated with a calibration before each measurement. A pattern similar to that for Study 3 was seen. A detailed examination of the gauge by a repairperson from the manufacturer revealed that one of the electronic components was not working properly. This was repaired and Study 3 was repeated. This study showed variation similar to the variation of the short term study (Study 1) so that the overall project could continue. When Study 2 was repeated, the overall variation and the number of scrap and reworked crankshafts was substantially reduced. The project was considered complete and long term monitoring showed that the scrap rate was reduced to about 0.7% which produced an annual savings of more than $100,000.

As well, three similar gauges that were used in the factory were put through the “long term” test. All were working well.

Summary

- An important part of any Plan is the choice and assessment of the measurement system.
- The measurement system may contribute substantial error that can result in poor decisions (e.g. scrapping good parts, accepting bad parts).
- We represent systematic measurement error by bias in the model. The bias can be assessed only by measuring units with known values, taken from another reference measurement system. The bias may be constant or depend on the size of the unit being measured, the person making the measurements, and so on.
- Variability can be assessed by repeatedly measuring the same unit. The variability may depend on the unit being measured or any other explanatory variates.
- Both bias and variability may be a function of time. This can be assessed by examining these attributes over a sufficiently long time span as in Study 3.

3.7 Problems

1. 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.

2. 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.

3. 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?

4. For a study like the one in Example 1.3.1, 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.

ESTIMATION

4.1 Introduction

Many statistical problems involve the estimation of some quantity or attribute. For example: the fraction of North American women age 16-25 who smoke; the 10th, 50th and 90th percentiles of body-mass index (BMI) for Canadian males age 21-35; the probability a sensor will classify the colour of an item correctly. The statistical approach to estimation is based on the following idea:

Develop a model for variation in the population or process you are considering, in which the attribute or quantity you want to estimate is included, and a corresponding model for data collection.

As we will see, this leads to powerful methods for estimating unknown attributes and, importantly, for determining the uncertainty in the estimates.

We have already seen in Chapter 2, that attributes can be expressed as parameters θ in a statistical model (probability distribution) and that they can be estimated using the method of maximum likelihood. Let us consider the following example and make some important observations.

Example 4.1.1. Suppose we want to estimate attributes associated with BMI for some population of individuals (e.g. 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 average BMI in the population which, in terms of the model, is $\mu = E(Y)$.

(ii) The median BMI in the population which, in terms of the model, is $\mu = E(Y)$ since the Gaussian distribution is symmetric about its mean.

(iii) For the BMI population, the 0.1 (population) quantile, $Q\,(0.1) = \mu - 1.28\sigma$, which satisfies $P\,(Y \leq q) = 0.1$. (To see this, note that $P(Y \leq \mu{-}1.28\sigma) = P(Z \leq -1.28) = 0.1$, where $Z = (Y - \mu)/\sigma \sim G(0,1)$.)

(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.

Thus, if we collected a random sample of, say, 150 individuals and calculated the maximum likelihood estimates as $\hat{\mu} = 27.1$, $\hat{\sigma} = 3.56$ then estimates of the attributes in (i)-(iv) would be: (i) and (ii) $\hat{\mu} = 27.1$, (iii) $\hat{Q}(0.1) = \hat{\mu} - 1.28\hat{\sigma} = 22.54$ and (iv) $\hat{p} = 0.0132$.

The preceding example raises several issues.

- Where do we get our probability distribution? 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 will be considered later in the course. 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 1.3.2 it was not clear that a Gaussian model was suitable.

- 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 the BMI example it has been assumed that BMI was measured without error for a random sample of units (persons) from the population. Here we typically assume that the data came 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.

- The estimate $\hat{\mu} = 27.1$ is an estimate of μ, the average BMI in the population but not usually equal to it. How far away from μ is $\hat{\mu}$ likely to be? If we take a sample of only $n = 50$ persons, would we expect the estimate $\hat{\mu}$ to be as "good" as $\hat{\mu}$ based on 150 persons? (What does "good" mean?)

We focus on the third point in this chapter; we assume that we can deal with the first two points with ideas introduced in Chapters 1 and 2.

4.2 Estimators and Sampling Distributions

Suppose that some attribute or parameter θ is to be estimated. We assume that a random sample $y_1, \ldots, y_n$ can be drawn from the population or process in question, from which θ

can be estimated. In general terms a **point estimate** of θ, denoted as $\hat{\theta}$, is some function of the observed sample $y_1, \ldots, y_n$,

$$\hat{\theta} = g(y_1, \ldots, y_n). \tag{4.2}$$

For example

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

is a point estimate of 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 is how to determine or quantify the uncertainty in an estimate. We do this using **sampling distributions**, which are based on the following idea. If we select random samples on repeated occasions, then the estimates $\hat{\theta}$ obtained from the different samples will vary. For example, five separate random samples of $n = 50$ persons from the same male population described in Example 1.3.1 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 the distribution of the estimator is called the **sampling distribution**.

More precisely, we define this idea as follows. Let the random variables $Y_1, \ldots, Y_n$ represent the observations in a random sample, and associate with the estimate $\hat{\theta}$ given by (4.2) a random variable

$$\tilde{\theta} = g(Y_1, \ldots, Y_n).$$

For example

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

is a random variable. We call $\tilde{\theta}$ the **estimator** of θ corresponding to $\hat{\theta}$. (We will always use $\hat{\theta}$ to denote an estimate, i.e. a numerical value, and $\tilde{\theta}$ to denote the corresponding estimator, the random variable.) We can think of $\tilde{\theta}$ as describing an estimation **procedure** or how to process the data to obtain an estimate, and the numerical value $\hat{\theta}$ as the value obtained from this procedure for a particular data set. The distribution of $\tilde{\theta}$ is called the **sampling distribution** of the estimator.

Since $\tilde{\theta}$ is a function of the random variables $Y_1, \ldots, Y_n$ we can find its distribution, at least in principle. Two ways to do this are (i) using mathematics and (ii) by computer simulation. Once we know the sampling distribution of an estimator $\tilde{\theta}$ then we are in the position to express the uncertainty in an estimate. The following example illustrates how this is done: **we examine the probability that the estimator $\tilde{\theta}$ is "close" to θ.**

Example 4.1.2

Suppose we want to estimate the mean $\mu = E(Y)$ of a random variable, and that a Gaussian distribution $Y \sim G(\mu, \sigma)$ describes variation in Y in the population. Let $Y_1, \ldots, Y_n$ represent a random sample from the population, and consider the estimator

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

for μ. Recall that if the distribution of Y_i is $G(\mu, \sigma)$ then the distribution of $\bar{Y}$ is Gaussian, $G(\mu, \sigma/\sqrt{n})$. Let us now consider the probability that the random variable $|\tilde{\mu} - \mu|$ is less than or equal to some specified value Δ. We have

$$P\left(|\tilde{\mu} - \mu| \le \Delta\right) = P\left(\mu - \Delta \le \bar{Y} \le \mu + \Delta\right) = P\left(\frac{-\Delta\sqrt{n}}{\sigma} \le Z \le \frac{\Delta\sqrt{n}}{\sigma}\right). \tag{4.3}$$

where $Z = (\bar{Y} - \mu)/(\sigma/\sqrt{n}) \sim G(0, 1)$. Clearly, as n increases, the probability (4.3) approaches one. Furthermore, if we know σ (even approximately) then we can find the probability for any given Δ and n. For example, suppose Y represents the height of a male (in meters) in the population of Example 1.3.1, and that we take $\Delta = 0.01$. That is, we want to find the probability that $|\tilde{\mu} - \mu|$ is no more than 0.01 meters. Assuming $\sigma = s = 0.07$ (meters), (4.1.3) gives the following results for sample sizes $n = 50$ and $n = 100$:

$$\begin{aligned} n = 50: &\quad P(|\tilde{\mu} - \mu| \le 0.01) = P(-1.01 \le Z \le 1.01) = 0.688 \\ n = 100: &\quad P(|\tilde{\mu} - \mu| \le 0.01) = P(-1.43 \le Z \le 1.43) = 0.847 \end{aligned}$$

This indicates that a large sample is "better" in the sense that the probability is higher that $\tilde{\mu}$ will be within 0.01m of the true (and unknown) average height μ in the population. It also allows us to express the uncertainty in an estimate $\hat{\mu} = \bar{y}$ from an observed sample $y_1, \ldots, y_n$ by indicating the probability that any single random sample will give an estimate within a certain distance of μ.

Example 4.1.3

In the preceding example we were able to work out the variability of the estimator mathematically, using results about Gaussian probability distributions. In some settings we might not be able to work out the distribution of an estimator mathematically; however, we could use simulation to study the distribution[10]. For example, suppose we have a random sample $y_1, \ldots, y_n$ which we have assumed comes from an Exponential(θ) distribution. The maximum likelihood estimate of θ is $\hat{\theta} = \bar{y}$. (Can you show this?) What is the sampling

[10] This approach can also be used to study sampling from a finite population of N values, $\{y_1, \ldots, y_N\}$, where we might not want to use a continuous probability distribution for Y.

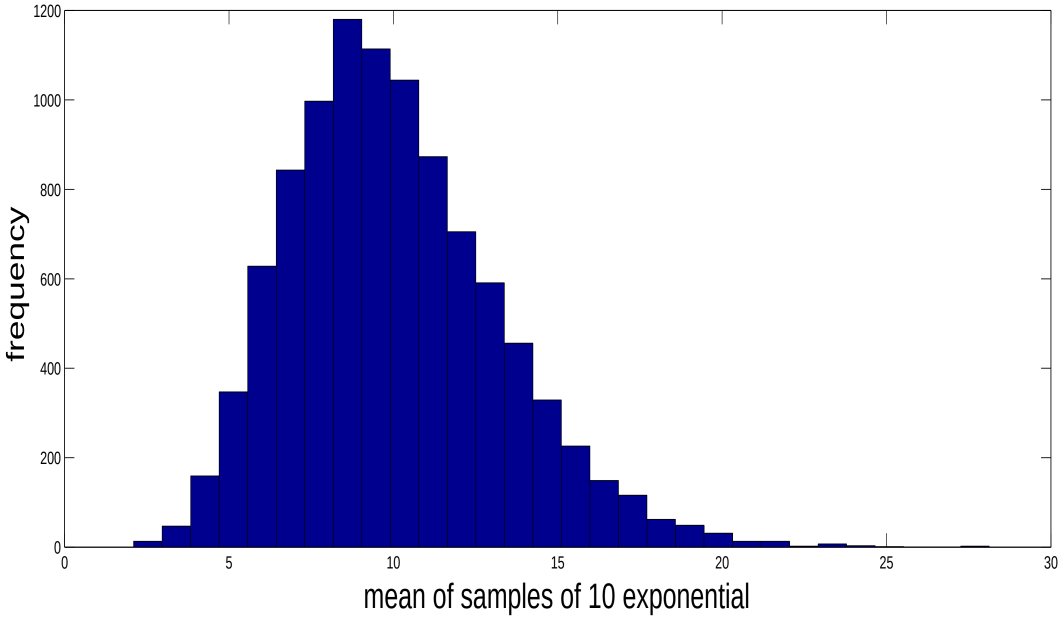


Figure 4.2:

distribution for $\tilde{\theta} = \bar{Y}$ in this case? We can examine the sampling distribution by taking repeated samples of size n, $y_1, \ldots, y_n$, giving (possibly different) values of $\bar{y}$ for each sample. We can investigate the distribution of the random variable $\tilde{\theta}$ by simulation, as follows:

1. Generate a sample of size n; in R this is done using the statement

   $$y \leftarrow \text{rexp}(n, 1/\theta).$$

   (Note that in R the parameter is specified as $1/\theta$.)

2. Compute $\hat{\mu} = \bar{y}$ from the sample; in R this is done using the statement

   $$ybar \leftarrow mean(y)$$

   We then repeat this, say k times. The k values $\bar{y}_1, \ldots, \bar{y}_k$ can then be considered as a sample from the distribution of $\tilde{\theta}$, and we can study the distribution by plotting a histogram or other plot of the values.

The histogram above was obtained by drawing $k = 10000$ samples of size $n = 10$ from an Exponential(10) distribution, calculating the values $\bar{y}_1, \ldots, \bar{y}_{10000}$ and then plotting the frequency histogram. What do you notice about the distribution particularly with respect to symmetry? Does the distribution look like a Gaussian distribution?

The approach illustrated in the preceding example can be used more generally. The main idea is that, for a given estimator $\tilde{\theta}$, we need to determine its sampling distribution in order to be able to compute probabilities of the form $P(|\tilde{\theta} - \theta| \leq \Delta)$ so that we can quantify the uncertainty of the estimate. We now review some results from probability and derive a few other results that will be used in estimation.

4.3 Some Distribution Theory

In the probability course you have taken have learned the basic discrete distributions: the Discrete Uniform, Binomial, Poisson, and Hypergeometric, as well as the simpler continuous distributions: the Uniform, Exponential and Normal or Gaussian. You should review this material. You should also review the Gamma function

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx \quad \text{for } \alpha > 0$$

and its properties.

There are a few new important distributions that will be introduced in this course, including the Chi-squared distribution and the Student t distribution. The first distribution we consider arises when we consider the distribution of a scaled verison of the relative likelihood function over repeated samples.

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

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

and the probability density function of W is

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

which is a member of the Chi-squared family of distributions. This result is revisted in Theorem 4.2.6 using moment generating functions.

The χ^2 (chi-squared) Distribution

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 is a parameter of the distribution taking values in the set $\{1, 2, \ldots\}$. We write $X \sim \chi^2(k)$. The parameter k is referred to as the "degrees of freedom" (d.f.) parameter. In Figure 4.3 you see the characteristic shapes of the Chi-squared probability density functions. For degrees of freedom $k = 2$, the density is the Exponential(1) density but for $k > 2$, the probability density function is unimodal with maximum value at $x = k - 2$. For large values of k, the probability density function resembles that of a Normal distribution with mean k and variance $2k$.

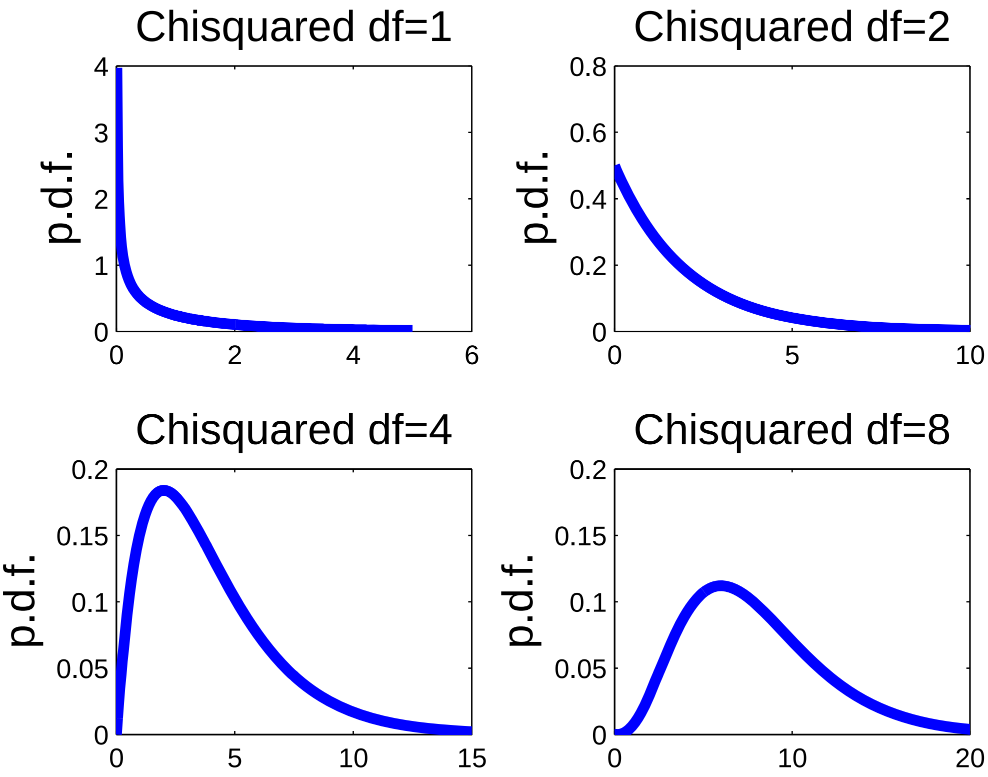


Figure 4.3: Chi-squared probabilities densities with degrees of freedom 1,2,4 and 8.

Problem 8 at the end of the chapter gives some results for the $\chi^2\,(k)$ distribution, including the fact that its moment generating function is

$$M(t) = (1-2t)^{-k/2} \quad \text{for } t < \frac{1}{2}, \tag{4.5}$$

and that its mean and variance are $E(X) = k$ and $Var(X) = 2k$. The cumulative distribution function, $F(x; k)$, can be given in closed algebraic form for even values of k. In R the functions $dchisq(x, k)$ and $pchisq(x, k)$ give the probability density function $f(x; k)$ and cumulative distribution function $F(x; k)$ for the $\chi^2(k)$ distribution. A table with selected values is given at the end of these notes.

We now state a pair of important results. The first shows that when we add independent Chi-squared random variables, the sum also has a Chi-squared distribution, and the degrees of freedom are added.

Theorem 1 [11] *Let $W_1, \ldots, W_n$ be independent random variables with $W_i \sim \chi^2(k_i)$. Then $S = \sum\limits_{i=1}^{n} W_i \sim \chi^2(\sum\limits_{i=1}^{n} k_i)$.*

We have already shown that the Chi-squared distribution arises as the square of a standard Normal random variable.

Theorem 2 *If $Z \sim G(0,1)$ then the distribution of $W = Z^2$ is $\chi^2(1)$.*

Furthermore if we add together the squares of several independent standard Normal random variables then we are adding independent Chi-squared random variables. The result can only have Chi-squared distribution.

Corollary 3 :[12] *If $Z_1, \ldots, Z_n$ are mutually independent $G(0,1)$ random variables and $S = \sum\limits_{i=1}^{n} Z_i^2$,then $S \sim \chi^2(n)$.*

Interval Estimation Using Likelihood Functions

The estimates and estimators discussed in Section 4.2 are often referred to as **point estimates** and **point estimators**. This is because they consist of a single value or "point". The discussion of sampling distributions shows how to address the uncertainty in an estimate, but we nevertheless prefer in most settings to also indicate explicitly the uncertainty in the estimate. This leads to the concept of an **interval estimate**[13], which takes the form

$$\theta \in [L(\mathbf{y}), U(\mathbf{y})] \quad \text{or} \quad L(\mathbf{y}) \leq \theta \leq U(\mathbf{y}),$$

where $L(\mathbf{y})$ and $U(\mathbf{y})$ are functions of the observed data $\mathbf{y}$. Notice that this provides an interval with endpoints L and U both of which depend on the data. If we let $L(\mathbf{Y})$ and

[11] **Proof:** W_i has m.g.f. $M_i(t) = (1-2t)^{-k_i/2}$. Thus $M_s(t) = \prod\limits_{i=1}^{n} M_i(t) = (1-2t)^{-\sum\limits_{i=1}^{n} k_i/2}$ and this is the m.g.f. of a χ^2 distribution with degrees of freedom $\sum\limits_{i=1}^{n} k_i$.

[12] **Proof**: By the theorem, each X_i^2 has a $\chi^2(1)$ distribution. Theorems 4.2.5 and 4.2.6 then give the result.

[13] See the video *What is a confidence Interval?* at *watstat.ca*

$U(\mathbf{Y})$ represent the associated random varibles then $[L(\mathbf{Y}),\ U(\mathbf{Y})]$ is a random interval. If we were to draw many random samples from the same population and each time we constructed the interval $[L(\mathbf{y}), U(\mathbf{y})]$ how often would the statement $L(\mathbf{y}) \leq \theta \leq U(\mathbf{y})$ be true? There is a specific probability (hopefully large) that this statement is correct in general, i.e. that the parameter will fall in this random interval, and this probability is $P[L(\mathbf{Y}) < \theta < U(\mathbf{Y})]$. This probability gives an indication how good the rule is by which the interval estimate was obtained. For example $P[L(\mathbf{Y}) < \theta < U(\mathbf{Y})] = 0.95$, means that 95% of the time (i.e. 95% of the different samples we might draw), the parameter falls in the interval $[L(\mathbf{y}), U(\mathbf{y})]$ constructed from the data set $\mathbf{y}$. This means we can be reasonably safe in assuming, on this occasion, and for this data set, it does so. In general, uncertainty in an estimate is explicitly stated by giving the interval estimate along with the probability $P(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})])$.

The likelihood function can also 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 $-\infty < \mu < \infty$ but the standard deviation must be positive, i.e. $\sigma > 0$. 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 4 *Suppose θ is scalar and that some observed data (say a random sample $y_1, \ldots, y_n$) have given a likelihood function $L(\theta)$. The **relative likelihood function** $R(\theta)$ is then defined as*

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

where $\hat{\theta}$ is the maximum likelihood estimate (obtained by maximizing $L(\theta)$) and Ω is the parameter space. Note that

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

Definition 5 *A $100p\%$ likelihood interval*[14] *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 larger values of $L(\theta)$ (and hence $R(\theta)$) are the most plausible in

[14] or a "p" likelihood interval

the 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 in Figure 2.4, the interval that the pollsters provided, i.e. 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

Suppose θ is the proportion of people in a large population who have a specific characteristic. If n persons are randomly selected and Y is the number who have the characteristic, then $Y \sim \text{Binomial}(n, \theta)$ is a reasonable model and the observed data y gives the likelihood function

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

Maximizing $L(\theta)$ we find $\hat{\theta} = y/n$ and

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

Figure 4.4 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**,

$$r(\theta) = \log R(\theta) = \ell(\theta) - \ell(\hat{\theta}) \quad \text{for } \theta \in \Omega$$

where $\ell(\theta) = \log L(\theta)$ is the log likelihood function. It is often 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 $r(\theta) \geq \log p$. While both plots are unimodal and have idential 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 have desirable properties. One is that they become narrower as the sample size increases, thus indicating that larger samples contain more information about θ. They are also easy to obtain, since all we really have to do is plot $R(\theta)$ or $r(\theta) = \log R(\theta)$.

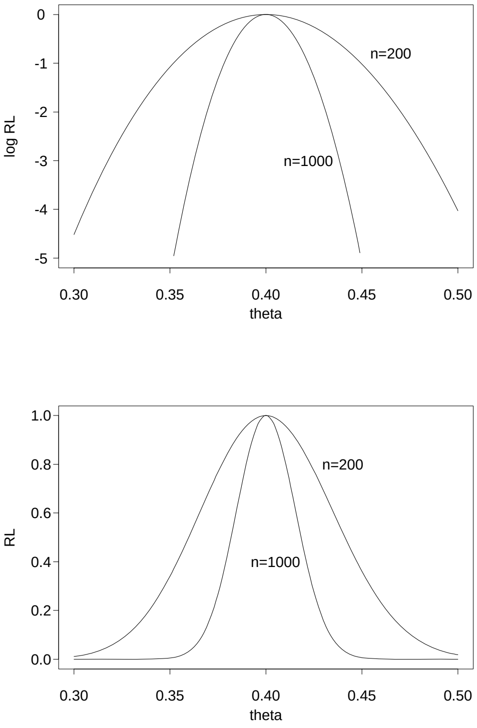


Figure 4.4: **Relative Likelihood and log Relative Likelihood Functions for a Binomial Parameter**

This approach can also be extended to deal with vector parameters, in which case $R(\boldsymbol{\theta}) \leq P$ gives likelihood "regions" for $\boldsymbol{\theta}$.

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 4.1 to the interval estimates, as discussed at the start of this section. This leads to the concept of confidence intervals, which we describe next.

4.4 Confidence Intervals for a Parameter

In general, a likelihood interval or any other interval estimate for θ based on observed data $\mathbf{y}$ takes the form $[L(\mathbf{y}), U(\mathbf{y})]$. Suppose we assume that the model chosen is correct and θ_0 is the true (unknown) value of the parameter. It is not certain that the statement

$\theta_0 \in [L(\mathbf{y}), U(\mathbf{y})]$ is true. To quantify the uncertainty in the interval estimate we look at an important property of the correponding interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$ called the **coverage probability** which is defined as follows.

Definition 6 *The value*

$$C(\theta_0) = P\left[L(\mathbf{Y}) \leq \theta_0 \leq U(\mathbf{Y})\right] \tag{4.6}$$

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 in (4.6). The parameter θ_0 is an unknown constant associated with the population, but it is a fixed constant, NOT a random variable and therefore does not have a distribution. The statement (4.6) 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_0. Suppose also that we knew that we would construct an interval of the form $[L(\mathbf{y}), U(\mathbf{y})]$ once we had collected the data. Then the probability that θ_0 will be contained in this new interval is $C(\theta_0)$[15].

How then does $C(\theta_0)$ assist in the evaluation of interval estimates? In practice, we try to find intervals for which $C(\theta_0)$ 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.*

Definition 7 *A* $100p\%$ ***confidence interval***[16] *for a parameter is an interval estimate* $[L(\mathbf{y}), U(\mathbf{y})]$ *for which*

$$P\left[L(\mathbf{Y}) \leq \theta_0 \leq U(\mathbf{Y})\right] = p \tag{4.7}$$

where p is called the confidence coefficient.

If $p = 0.95$, for example, then (4.7) indicates that 95% of the samples $\mathbf{Y}$ that we would draw from this model result in an interval $[L(\mathbf{Y}), U(\mathbf{Y})]$ which includes the parameter θ_0 (and of course 5% do not). This gives us some confidence that for a particular sample, such as the one at hand, the true value of the parameter is contained in the interval.

To show that confidence intervals exist, and that the confidence coefficient can sometimes not depend on the unknown parameter θ_0, we consider the following simple example.

Example 4.4.1

[15] When we use the observed data y, $L(y)$ and $U(y)$ are numerical values not random variables. We do not know whether or not $L(y) \leq \theta_0 \leq U(y)$. $P\left[L(y) \leq \theta_0 \leq U(y)\right]$ makes no more sense than $P\left(1 \leq \theta_0 \leq 3\right)$ since $L(y), \theta_0, U(y)$ are all numerical values: there is no random variable to which the probability statement can refer.

[16] See the video at www.watstat.com called ”what is a confidence interval”

Suppose $Y_1, \ldots, Y_n$ is a random sample from a $G(\mu_0, 1)$ distribution. That is, $\mu_0 = E(Y_i)$ 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 $\overline{Y} = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i$ is the sample mean. Since $\overline{Y} \sim G(\mu_0, 1/\sqrt{n})$, then

$$\begin{aligned}
&P\left(\overline{Y} - 1.96/\sqrt{n} \leq \mu_0 \leq \overline{Y} + 1.96/\sqrt{n}\right) \\
&= P\left[-1.96 \leq \sqrt{n}\left(\overline{Y} - \mu_0\right) \leq 1.96\right] \\
&= P\left(-1.96 \leq Z \leq 1.96\right) \\
&= 0.95
\end{aligned}$$

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. This is an example in which the confidence coefficient does not depend on the unknown parameter, an extremely desirable feature of an interval estimator.

We repeat the very important interpretation of a $100p\%$ confidence interval (since so many people get the interpretation incorrect!): *If the procedure is used repeatedly then in a fraction p of cases the constructed intervals will contain the true value of the unknown parameter.* 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 the probability that $\mu_0 \in [9.91,\ 10.89]$ is 0.95, but we have a high degree of **confidence** (95%) that the interval $[9.91,\ 10.89]$ contains μ_0.

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, and we show a bit later that likelihood intervals are a type of confidence interval.

Recall that the coverage probability for the interval in the above example did not depend on the unknown parameter, a highly desirable property because we'd like to know the coverage probability while not knowing the value of the unknown parameter. We next consider a general method for finding confidence intervals which have this property.

Pivotal Quantities and Confidence Intervals

Definition 8 *A **pivotal quantity** $Q = g(\mathbf{Y}; \theta)$ is a function of the data $\mathbf{Y}$ and the unknown parameter θ such that the distribution of the random variable Q is fully known. That is, probability statements such as $P\,(Q \geq a)$ and $P\,(Q \leq b)$ depend on a and b but not θ or any other unknown information.*

The motivation for this definition is the following. Suppose we can begin with a statement such as $P\left[a \leq g(\mathbf{Y};\theta) \leq b\right] = 0.95$ where $g(\mathbf{Y};\theta)$ is a pivotal quantity whose distribution is completely known. Suppose also 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

$$0.95 = P\left[a \leq g(\mathbf{Y};\theta) \leq b\right] = P\left[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})\right],$$

the confidence coefficient for the confidence interval $\left[L\left(\mathbf{y}\right),\ U\left(\mathbf{y}\right)\right]$ is equal to 0.95 which does not depend on the unknown parameter θ. $\left[L\left(\mathbf{y}\right), U\left(\mathbf{y}\right)\right]$ is a confidence interval for θ with confidence coefficient equal to 0.95, a value which does not depend on the parameter θ. The confidence coefficient does depend on a and b, but these are determined by the known distribution of $g(\mathbf{Y};\theta)$.

Example 4.4.2

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a random sample from the $G(\mu, \sigma_0)$ distribution where $E\left(Y_i\right) = \mu$ is unknown but $sd\left(Y_i\right) = \sigma_0$ is known. Since

$$Q = Q\left(\mathbf{Y};\mu\right) = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim G(0,1)$$

and $G(0,1)$ is a completely known distribution, Q is a pivotal quantity. (For simplicity we just write μ instead of μ_0 for the unknown true value which is to be estimated.) To obtain a 95% confidence interval for μ we need to find values a and b such that $P(a \leq Q \leq b) = 0.95$. Now

$$\begin{aligned} 0.95 &= P\left(a \leq \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \leq b\right) \\ &= P\left(\overline{Y} - b\sigma_0/\sqrt{n} \leq \mu \leq \overline{Y} - a\sigma_0/\sqrt{n}\right), \end{aligned}$$

so that

$$\bar{y} - b\sigma_0/\sqrt{n} \leq \mu \leq \bar{y} - a\sigma_0/\sqrt{n} \ \text{ or } \ \left[\bar{y} - b\sigma_0/\sqrt{n},\ \bar{y} - a\sigma_0/\sqrt{n}\right]$$

is a 95% confidence interval for μ based on the observed data $\mathbf{y} = (y_1, \ldots, y_n)$. Note that there are infinitely many pairs (a, b) giving $P(a \leq Q \leq b) = 0.95$. A common choice for the standard normal is to pick points symmetric about zero, $a = -1.96$, $b = 1.96$; this gives the interval $[\bar{y} - 1.96\sigma_0/\sqrt{n},\ \bar{y} + 1.96\sigma_0/\sqrt{n}]$ or $\bar{y} \pm 1.96\sigma_0/\sqrt{n}$ which turns out to be the narrowest possible 95% confidence interval. The interval $[\bar{y} - 1.96\sigma_0/\sqrt{n},\ \bar{y} + 1.96\sigma_0/\sqrt{n}]$ is often referred to as a "two-sided" confidence interval. Note also that this interval takes the form

point estimate $\pm\, a \times$ standard deviation of the estimator.

Many "two-sided" confidence intervals in this course will take this form.

Another choice for a and b would be $a = -\infty$, $b = 1.645$, which gives the interval $[\overline{y} - 1.645\sigma_0/\sqrt{n}, \infty)$. The interval $[\overline{y} - 1.645\sigma_0/\sqrt{n}, \infty)$ 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 distributions it is not possible to find "exact" pivotal quantities or confidence intervals for θ whose coverage probabilities do not depend somewhat on the true value of θ. However, in general we can find quantities $Q_n = g(Y_1, ..., Y_n, \theta)$ 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. Polls

Consider Example 4.3.1 discussed earlier, where $Y \sim \text{Binomial}(n, \theta)$. 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

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

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

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

Thus

$$\hat{\theta} \pm 1.96\left[\frac{\hat{\theta}(1-\hat{\theta})}{n}\right]^{1/2} \tag{4.8}$$

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 $n = 100$, $y = 18$ in a poll. Then (4.8) becomes $0.18 \pm 1.96\,[0.18(0.82)/100]^{1/2}$ or $0.115 \le \theta \le 0.255$ or $[0.115,\ 0.255]$.

Remark: It is important to understand that confidence intervals may vary quite a lot when we take repeated samples. For example, in Example 4.4.3, ten samples of size $n = 100$ which were simulated for a population where $\theta = 0.25$ gave the following approximate 95% confidence intervals for θ:

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

For larger samples (larger n), the confidence intervals are narrower and will have better agreement. For example, try generating a few samples of size $n = 1000$ and compare the confidence intervals for θ.

Likelihood-Based Confidence Intervals

Likelihood intervals are approximate confidence intervals and sometimes they are exact confidence intervals. Recall the relative likelihood $R(\theta) = L(\theta)/L(\hat{\theta})$ and define the quantity

$$\Lambda = \Lambda(\theta) = -2\log R(\theta) = 2\ell(\tilde{\theta}) - 2\ell(\theta)$$

where $\tilde{\theta}$ is the maximum likelihood estimator. Then Λ, which is a random variable, is called the **likelihood ratio statistic**. The following result can be proved:

Proposition 9 *If $L(\theta)$ is based on $\mathbf{Y} = (Y_1, \ldots, Y_n)$, 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 Λ converges to $\chi^2(1)$ as $n \to \infty$.*

This means that Λ can be used as an approximate pivotal quantity in order to get confidence intervals for θ. Because highly plausible values of θ are ones for which $R(\theta)$ is close to one (i.e. Λ is close to zero), we obtain approximate $100p\%$ confidence intervals for θ by working from the probability $P(W \leq c) = p$ where $W \backsim \chi^2(1)$. Since

$$p = P(W \leq c) \approx P(\Lambda \leq c)$$

an approximate $100p\%$ confidence intervals for θ is obtained by finding all θ values such that $2\ell(\hat{\theta}) - 2\ell(\theta) \leq c$ where $\hat{\theta}$ is the maximum likelihood estimate, i.e.

$$\left\{\theta : 2\ell(\hat{\theta}) - 2\ell(\theta) \leq c\right\} = \left\{\theta : R(\theta) \geq e^{-c/2}\right\}$$

is an approximate $100p\%$ confidence interval for θ. Usually this interval must be found numerically.

Example 4.4.4

Consider the Binomial model in Examples 4.3.1 and 4.4.3. The likelihood ratio statistic (show it!) is

$$\Lambda(\theta) = 2n\tilde{\theta}\log(\tilde{\theta}/\theta) + 2n(1-\tilde{\theta})\log\left(\frac{1-\tilde{\theta}}{1-\theta}\right)$$

where $\tilde{\theta} = Y/n$ is the maximum likelihood estimator of θ. To get an approximate 95% confidence interval for θ we note that $P(W \leq 3.841) = 0.95$ where $W \backsim \chi^2(1)$. To find the approximate confidence interval we need to find all θ values satisfying

$$2n\hat{\theta}\log(\hat{\theta}/\theta) + 2n(1-\hat{\theta})\log\left(\frac{1-\hat{\theta}}{1-\theta}\right) \leq 3.841$$

where $\hat{\theta} = y/n$. This must be done numerically, and depends on the observed data y. For example, suppose that for $n = 100$, we observe $y = 40$ so that $\hat{\theta} = 0.40$. Let $\lambda(\theta)$ be the observed value of the random variable $\Lambda(\theta)$ for these data so that

$$\lambda(\theta) = 80 \log(0.4/\theta) + 120 \log\left(\frac{0.6}{1-\theta}\right).$$

Figure 4.5 shows a plot of $\lambda(\theta)$ and the horizontal line $\lambda = 3.841$ from which the approximate 95% confidence interval can be determined. Solving $\lambda(\theta) \leq 3.841$, we obtain $0.307 \leq \theta \leq 0.496$ or $[0.307,\ 0.496]$ is the approximate 95% confidence interval.

We could also use the approximate 95% confidence interval (4.8) from Example 4.4.3 for this situation. It gives the interval is $0.304 \leq \theta \leq 0.496$ or $[0.304,\ 0.496]$. The two confidence intervals differ slightly (they are both based on approximations) but are extremely close.

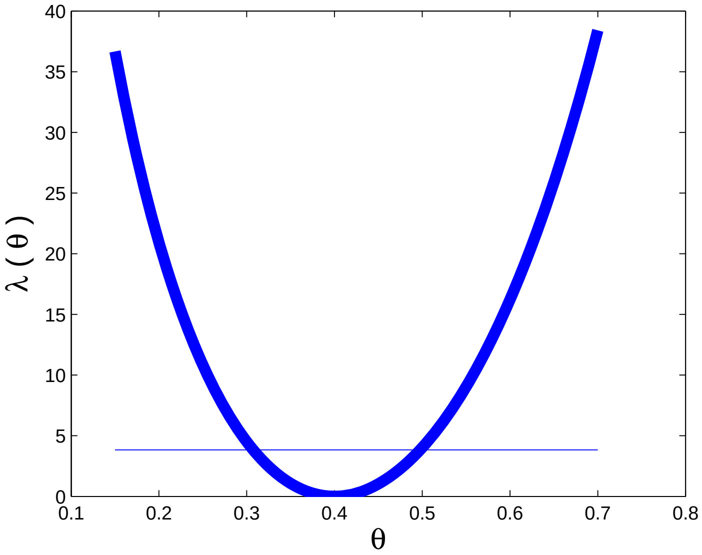


Figure 4.5: **Likelihood Ratio Statistic for Binomial Parameter**

We can now see that a likelihood interval is also a confidence interval. We first note that the 100p% likelihood interval defined by $\{\theta; R(\theta) \geq p\}$ is equivalent to

$$\left\{\theta : \lambda(\theta) = 2\ell(\hat{\theta}) - 2\ell(\theta) \leq 2\log p\right\}.$$

The confidence coefficient for this interval is $P[\Lambda(\theta) \leq -2\log p]$ which can be approximated by

$$P[\Lambda(\theta) \leq -2\log p] \approx P(W \leq -2\log p)$$

where $W \backsim \chi^2(1)$. If we take $p = 0.1$ then since

$$P[\Lambda(\theta) \leq -2\log(0.1)] \approx P[W \leq -2\log(0.1)] = 0.968$$

a 10% likelihood interval is an approximate 96.8% confidence interval.

Conversely since $P\left(W \leq 3.841\right) = 0.95$ this implies that the set of values

$$\left\{\theta : \lambda\left(\theta\right) = 2\ell(\hat{\theta}) - 2\ell(\theta) \leq 3.841\right\} = \{\theta : R(\theta) \geq 0.147\}$$

represents an approximate 95% confidence interval. Therefore an approximate 95% confidence interval for θ is given by a 14.7% likelihood interval. What likelihood intervals would correspond to approximate 90% and 99% confidence intervals?

4.4.3 Choosing a Sample Size

We have seen in examples in this chapter 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 Estimation of a Binomial Probability

Suppose we want to estimate the probability θ from a Binomial experiment in which the response variable Y has a Binomial(n, θ) distribution. We will use the approximate pivotal quantity

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

introduced in Example 4.4.3 which has approximately the $G(0, 1)$ distribution. This will be used to obtain confidence intervals for θ. (Using the likelihood ratio statistic leads to a more difficult derivation and in any case, for large n, confidence intervals constructed using the likelihood ratio statistic are very close to those based on Q.) 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 $2\left(0.03\right)$. Let us see why this is used and where it leads. From Example 4.4.3, we know that (see (4.4.2))

$$\hat{\theta} \pm 1.96\left[\hat{\theta}(1-\hat{\theta})/n\right]^{1/2}$$

is an approximate 0.95 confidence interval for θ and the width of this interval is

$$2\left(1.96\right)\left[\hat{\theta}(1-\hat{\theta})/n\right]^{1/2}.$$

To make this confidence interval narrower that $2\left(0.03\right)$ (or even narrower, say $2\left(0.025\right)$), we need n large enough so that

$$1.96\left[\hat{\theta}(1-\hat{\theta})/n\right]^{1/2} \leq 0.03$$

or

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

Of course don't know what $\hat{\theta}$ is because we have not taken a sample, but we note that the worst case scenario occurs 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 $n = 480$ (show this). In many situations this might not be sufficiently accurate for the purpose of the study, however.

Exercise: Show that to ensure that width of the approximate 95% confidence interval is $2\,(0.02)$ or smaller, you need $n = 2401$. What should n be to make a 99% confidence interval less than $2\,(0.02)$ or less?

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 Chapter 6. Many of the tools of this section can also be extended to the muli-parameter[17] setting but we will not discuss this further here.

[17]

Models With Two or More Parameters. When there is a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ of unknown parameters, we may want to get interval estimates for individual parameters θ_j, $j = 1, \ldots, k$ or for functions $\psi = h(\theta_1, \ldots, \theta_k)$. For example, with a Gaussian $G(\mu, \sigma)$ model we might want to estimate μ and σ. In some problems there are pivotal quantities which are functions of the data and (only) the parameter of interest. We will use such quantities in Chapter 6, where we consider estimation and testing for Gaussian models.

4.5 A Case Study: Testing Reliability of Computer Power Supplies

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

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 4.5.1 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 4.5.1 the asterisks denote the censored observations. Note the data have been organized so that the lifetimes are listed first followed by the cenored 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 $t = 40°$. The

There also exist approximate pivotal quantities based on the likelihood function and maximum likelihood estimates. These are mainly developed in more advanced followup courses to this one, but we will briefly consider this approach later in the notes.

It is also possible to construct **confidence regions** for two or more parameters. 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. Then we can define the relative likelihood function $R(\theta_1, \theta_2) = L(\theta_1, \theta_2)/L(\hat{\theta}_1, \hat{\theta}_2)$ as in the scalar case. The set of pairs (θ_1, θ_2) which satisfy $R(\theta_1, \theta_2) \geq p$ is then called a **100p% likelihood region** for (θ_1, θ_2). The concept of confidence intervals can similarly be extended to **confidence regions**.

data are $y_1, \ldots, y_{25}$ 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 have let $\mu = \theta_{40}$ for the moment for convenience. The contribution to the likelihood function for an observed lifetime y_i we know is simply

$$f(y_i; \mu) = (1/\mu) \exp\left(-y_i/\mu\right).$$

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

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

the contribution to the likelihood function of each censored observation is $e^{-600/\mu}$. Therefore the likelihood function for μ based on the data $y_1, \ldots, y_{25}$ is

$$L(\mu) = \left[\prod_{i=1}^{22} \frac{1}{\mu} e^{-y_i/\mu}\right] \left[\prod_{i=23}^{25} e^{-y_i/\mu}\right] = \mu^{-k} \exp\left(-s/\mu\right)$$

where $k = 22 =$ the number of uncensored obervations and $s = \sum_{i=1}^{25} y_i =$ 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 $\hat{\theta}_{40} = s/k$.

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

$$L(\theta_t) = (\theta_t)^{-k_t} \exp\left(-s_t/\theta_t\right)$$

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 discussed in Chapter 1 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

$$\theta_t = \exp\left(\alpha + \frac{\beta}{t + 273.2}\right) \tag{4.9}$$

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

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

where θ_t is given by (4.9). (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 its about 100 hours. If the model (4.9) 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.

Table 4.5.1 Lifetimes (in hours) from an accelerated life test experiment in PC power supplies

Temperature			
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.

4.6 Problems

1. Consider the data on heights of adult males and females from Chapter 1. (The data are on the course web page.)

   (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).

2. When we measure a quantity we are in effect estimating the true value of the quantity; measurements of the same variate on different occasions are usually not equal. A chemist has two ways of measuring a particular quantity; one has more random error than the other. For method I, measurements $X_1, X_2, \ldots$ follow a normal distribution with mean μ and variance σ_1^2, whereas for method II, measurements $Y_1, Y_2, \ldots$, have a normal distribution with mean μ and variance σ_2^2.

(a) Suppose that the chemist has n measurements $X_1, \ldots, X_n$ of a quantity by method I and m measurements, $Y_1, \ldots, Y_m$ by method II. Assuming that σ_1^2 and σ_2^2 are known, write down the combined likelihood function for μ, and show that

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

where $w_1 = n/\sigma_1^2$ and $w_2 = m/\sigma_2^2$. Why does this estimator 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) Determine the standard deviation of $\tilde{\mu}$ and of $(\bar{X} + \bar{Y})/2$ under the conditions of part (b). Why is $\tilde{\mu}$ a better estimator?

3. Suppose that a fraction p of a large population of persons over 18 years of age never drink alcohol. In order to estimate p, a random sample of n persons is to be selected and the number y who do not drink determined; the maximum likelihood estimate of p is then $\hat{p} = y/n$. We want our estimate $\hat{p}$ to have a high probability of being close to p, and want to know how large n should be to achieve this.Consider the random variable Y and estimator $\tilde{P} = Y/n$.

(a) Describe how you could work out the probability that $-0.03 \leq \tilde{P} - p \leq 0.03$, if you knew the values of n and p.

(b) Suppose that $p = 0.40$. Using an approximation determine how large n should be in order to ensure

$$P\left(-0.03 \leq \tilde{P} - p \leq 0.03\right) = 0.95.$$

4. Let n and k be integers. Suppose that blood samples for $n \times k$ people are to be tested to obtain information about θ, the fraction of the population infected with a certain virus. In order to save time and money, **pooled testing** is used: 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) Give an expression for the probability that x 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 x.

(c) Suppose $n = 100$, $k = 10$ and $x = 89$. Give the maximum likelihood estimate $\hat{\theta}$, the relative likelihood function, and find a 10% likelihood interval for θ.

(d) Discuss (or do it) how you would select an "optimal" value of k to use for pooled testing, if your objective was not to estimate θ but to identify persons who are infected, with the smallest number of tests. Assume that you know the value of θ and the procedure would be to test all k persons individually each time a pooled sample was positive. (Hint: Suppose a large number n of persons must be tested, and find the expected number of tests needed.)

(a) For the data in Problem 4 of Chapter 2, plot the relative likelihood function $R(\alpha)$ and determine a 10% likelihood interval. 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 and maximum likelihood estimate of 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.

5. 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

16	14	25	19	23	12
22	28	19	15	18	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

13	7	12	9	15	17
10	13	8	10	12	14

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). Do you think the government's decision was a good one, as far as the reliability of the machines is concerned?

(c) Use the likelihood ratio statistic $\Lambda(\theta)$ as an approximate pivotal quantity to obtain an approximate 95% confidence intervals for θ for each company.

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

6. The lifetime T (in days) of a particular type of lightbulb is assumed to have a distribution with probability density function

$$f(t;\theta) = \frac{\theta^3 t^2 e^{-\theta t}}{2} \quad \text{for } t > 0 \text{ and } \theta > 0.$$

(a) Suppose $t_1, t_2, \ldots, t_n$ is a random sample from this distribution. Show that the likelihood function for θ is equal to

$$c \times \theta^{3n} \exp\left(-\theta \sum_{i=1}^{n} t_i\right) \quad \text{for } \theta > 0$$

where c is constant with respect to θ.

(b) Find the maximum likelihood estimate $\hat{\theta}$ and the relative likelihood function $R(\theta)$.

(c) If $n = 20$ and $\sum_{i=1}^{20} t_i = 996$, graph $R(\theta)$ and determine the 10% likelihood interval for θ. What is the approximate confidence level associated with this interval?

(d) Suppose we wish to estimate the mean lifetime of a lightbulb. Show $E(T) = 3/\theta$. (Recall that $\int_0^\infty x^{n-1} e^{-x} dx = \Gamma(n) = (n-1)!$ for $n = 1, 2, \cdots$). Find a 95% confidence interval for the mean.

(e) The probability p that a lightbulb lasts less than 50 days is

$$p = p(\theta) = P(T \leq 50; \theta) = 1 - e^{-50\theta}[1250\theta^2 + 50\theta + 1].$$

(Can you show this?) Thus $\hat{p} = p(\hat{\theta}) = 0.580$ and we can find a 95% confidence interval for p from a confidence interval for θ. In the data referred to in part (c), the number of lightbulbs which lasted less than 50 days was 11 (out of 20). Using a Binomial model, we can also obtain a 95% confidence interval for p (see Examples 4.4.3 and 4.4.4). Find both intervals. What are the pros and cons of the second interval over the first one?

7. **The χ^2 (Chi-squared) distribution.** Suppose $Y \sim \chi^2(k)$ with probability density function given by

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

(a) Show that this probability density function integrates to one for any $k \in \{1, 2, \ldots.\}$.

(b) Show that the moment generating function of Y is given by

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

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

(c) Plot the probability density function for $k = 5$, $k = 10$ and $k = 25$ on the same graph. What do you notice?

8. 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 $\sum\limits_{i=1}^{30} x_i = 11,400$ days. It is known that survival times were approximately Exponentially distributed with mean θ days.

(a) Write down the likelihood function for θ and obtain the likelihood ratio statistic. Use this to obtain an approximate 90% confidence interval for θ.

(b) Show that $m = \theta \ln 2$ is the median survival time. Give a approximate 90% confidence interval for m.

9. Let X have an Exponential distribution with probability density function

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

where $\theta > 0$.

(a) Show that $Y = 2X/\theta$ has a $\chi^2(2)$ distribution. (Hint: compare the probability density function of Y with (4.4).

(b) If $X_1, \ldots, X_n$ is a random sample from the Exponential distribution above, prove that

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

(You may use results in Section 4.2.) U is therefore a pivotal quantity, and can be used to get confidence intervals for θ.

(c) Refer to Problem 9. 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 interval found in 9(a). Which interval is preferred here? (Why?)

10. Two hundred adults are chosen at random from a population and each is asked whether information about abortions should be included in high school public health sessions. Suppose that 70% say they should.

(a) Obtain a 95% confidence interval for the proportion p of the population who support abortion information being included.

(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).

11. Consider the height data discussed in Problem 1 above. If heights Y are $G(\mu, \sigma)$ and $\tilde{\mu} = \bar{Y}$ and $\tilde{\sigma}^2 = \sum_{i=1}^{n} (Y_i - \tilde{\mu})^2/n$ are the maximum likelihood estimators based on a sample of size n then it can be shown that when n is large, the random variable

$$Z = \frac{\sqrt{n}(\bar{Y} - \mu)}{\tilde{\sigma}}$$

has approximately a $G(0, 1)$ distribution and so Z is an approximate pivotal quantity. Use Z to obtain approximate 99% confidence intervals for μ for males and for females.

12. 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 (anonymized) 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.

13. [18] 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 $X_1, X_2, \cdots$ 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 $X_1, \ldots, X_n$ is said to be *consistent* for θ if $\tilde{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

(i) Let $X_1, \ldots, X_n$ be independent and identically distributed Uniform$(0,\theta)$ random variables. Show that $\tilde{\theta}_n = \max(X_1, \ldots, X_n)$ is consistent for θ.

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

14. [19] Refer to the definition of consistency in Problem 14(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

$$X_{i1}, X_{i2} \sim N(\mu_i, \sigma^2) \quad \text{for } i = 1, \cdots, n$$

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. (To do this you have to find the maximum likelihood estimators for $\mu_1, \ldots, \mu_n$ as well as for σ^2.)

(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?

15. [20] **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.

[18] Challenge problem: optional

[19] Challenge problem: optional

[20] Challenge problem: optional

(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 $\left[1 + \frac{t^2}{2n} + 0(n)\right]$, where $0(n)$ 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 $\left[1 + \frac{t^2}{2n} + 0(n)\right]^n$. 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 $n \to \infty$.)

TESTS OF HYPOTHESES

5.1 Introduction

What can it mean to test a hypothesis in the light of observed data or information? 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 the 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 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* the first hypothesis, "*the defendant is innocent*", 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. We may simply conclude at the end of the proceedings that there was insufficient evidence for a finding of guilty and the defendant is then exonerated. Of course there are also 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, and so the hypothesis is less credible. This default hypothesis is often referred to as the "null" hypothesis, denoted by H_0 ("null" is used because it often means a new treatment has no effect). Of course, there is an alternative, not always specified, because in many cases it is 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 (my spouse) 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 sort of values of observed value of Y are highly inconsistent with H_0 and what sort of values are compatible with H_0?

3. What observed values of Y would lead to accepting H_0 and what observed values would lead to rejecting H_0?

In answer to 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 2 we note that clearly observed values of Y that are very small (e.g. $0 - 10$) or very large (e.g. $90 - 100$) would lead us to to believe that H_0 may be false, whereas values near 50 are perfectly consistent with H_0. This leads naturally to the concept of a **test statistic** (also called a **discrepancy measure**) which is some function of the data $D = g(\mathbf{Y})$ that is constructed to measure the degree of "agreement" between the data $\mathbf{Y}$ and the hypothesis H_0. It is conventional to define D so that $D = 0$ represents the best possible agreement between the data and H_0, and so that the larger D is, the poorer the agreement. Methods of constructing test statistics will be described later, 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 $d = 2$. 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 $P(D \geq d)$ assuming that H_0 is true. We can compute this easily in the our given example. If H_0 is true then $Y \sim$ Binomial$(n, 0.5)$ and

$$\begin{aligned}
P\left(D \geq d\right) &= P\left(|Y - 50| > |52 - 50|\right) = P\left(|Y - 50| > 2\right) \\
&= 1 - P(49 \leq Y \leq 51) \\
&= 1 - \binom{100}{49}(0.5)^{100} - \binom{100}{50}(0.5)^{100} - \binom{100}{51}(0.5)^{100} \\
&\approx 0.76435.
\end{aligned}$$

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 (i.e. result in at least as large a value of D as the observed value of 2) as 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 evidently 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. This probability, 0.76453 in this example, is called the observed significance level or the $p-value$ of a test.

We now proceed to a more formal treatment of hypothesis tests. Two types of hypotheses that a statistician or scientist might be called upon to test in the light of observed data are:

(1) assuming a family of distributions, say having probability density function $f(\mathbf{y};\theta)$ for the data $\mathbf{Y}$, that the parameter θ has some specified value θ_0; we denote this as $H_0 : \theta = \theta_0$.

(2) that a random variable Y has a specified probability distribution, say with probability density function $f_0(y)$; we denote this as $H_0 : Y \sim f_0(y)$.

The above test of ESP is an example of the first of these. For the second, you might question whether a "default" hypothesis should be that a random variable follows a specific distribution such as the normal distribution and certainly this is not appropriate unless we have very good reasons, practical or theoretical, for this assumption.

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 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. To test H_0, we now calculate the observed **p-value** (also called the **observed significance level**), defined as

$$p - value = P(D \geq d; H_0), \tag{5.2}$$

where the notation "$; H_0$" means "assuming H_0 is true". If the p-value is close to zero then we are inclined to doubt that H_0 is true, because **if it is true** the **probability of getting agreement as poor or worse than that observed** is small. This makes the alternative explanation, H_0 is false, more appealing. In other words, we must accept that one of the following two statements is correct:

(a) H_0 is true but *by chance* and we have observed $\mathbf{Y}$ that indicates poor agreement with H_0, or

(b) H_0 is false.

The p-value indicates how small that chance is in (a) above. If it is large, there is no evidence for (b). If it is less than about 0.05, we usually interpret that as providing moderately strong evidence against H_0 in light of the observed data. If it is very small, for example 0.001, this is taken as very strong evidence against H_0 in light of the observed data.

Example 5.1.1 Testing a binomial probability

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 $\theta = P\,(\text{die turns up one})$ 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. Then $\mathbf{Y} = Y$ and 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), $D = \max((Y - n/6), 0)$.

Suppose that $n = 180$ tosses gave $y = 44$. Using $D = \max((Y - n/6), 0)$, we get $d = \max((44 - 180/6), 0) = 14$ and

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

which provides strong evidence against H_0, and suggests that θ is bigger than 1/6.

Example 5.1.2 Suppose that in the experiment in Example 5.1.1 we observed $y = 35$ ones in $n = 180$ tosses. Now the p-value is

$$\begin{aligned} p - value &= P(Y \geq 35; \theta = 1/6) \\ &= \sum_{y=35}^{180} \binom{180}{y} \left(\frac{1}{6}\right)^y \left(\frac{5}{6}\right)^{180-y} \\ &= 0.183 \end{aligned}$$

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 strong 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.

Example 5.1.3. 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 sensible test of $H_0 : \mu = 1$ could be based on the test statistic

$$D = |\overline{Y} - 1|$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Since $\overline{Y} \sim G(\mu, \sigma/\sqrt{n})$, where $\mu = E(Y)$ and $\sigma^2 = Var(Y)$, we can compute the p-value (at least approximately) using a Gaussian distribution. Since we don't know σ^2 we will estimate it by the sample variance s^2 in the calculations below. Of course if we substitute an estimate of the variance in place of the true variance this may (does!) make a difference to the distribution of the test statistic, and this is a refinement of this test that we will deal with in Section 6.2, but for the present, to keep things simple, let us pretend that our true variance is identical to the estimated one.

The samples from scales A and B above give us

$$\begin{aligned} A: &\quad \bar{y} = 1.0061, \quad s = 0.0230, \quad d = 0.0061 \\ B: &\quad \bar{y} = 0.9810, \quad s = 0.0170, \quad d = 0.0190. \end{aligned}$$

The p-value for A is (pretending $\sigma = s = 0.0230$)

$$\begin{aligned} p - value &= P(D \geq 0.0061;\ \mu = 1) \\ &= P(|\overline{Y} - 1| \geq 0.0061) \\ &= P\left(\left| \frac{\overline{Y} - 1}{0.0230/\sqrt{10}} \right| \geq \frac{0.0061}{0.0230/\sqrt{10}} \right) \\ &= P(|Z| \geq 0.839) \quad \text{where} \quad Z \sim G(0,1) \\ &= 0.401 \end{aligned}$$

and thus there is no evidence of bias (that is, no evidence that $H_0 : \mu = 1$ is false) for scale A.

For scale B, however, we get, (again pretending pretending $\sigma = s = 0.0170$)

$$\begin{aligned} p - value &= P\left(\left| \frac{\overline{Y} - 1}{0.0170/\sqrt{10}} \right| \geq \frac{0.0190}{0.0170/\sqrt{10}} \right) \\ &= P(|Z| \geq 3.534) \\ &= 0.0004 \end{aligned}$$

and thus there is very strong evidence against $H_0 : \mu = 1$, suggesting strongly that scale B is biased.

Finally, note that just because 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 get an approximate 95% confidence interval for $\mu = E(Y)$ for scale B by using the approximate pivotal quantity

$$Z = \frac{\overline{Y} - \mu}{s/\sqrt{10}} \text{ is approximately } G(0,1).$$

In section 6.2 we will find the exact distribution of $\frac{\overline{Y}-\mu}{S/\sqrt{10}}$ but for the present we will be satisfied using the Gaussian approximation. Since for a standard Gaussian random variable Z, $P(-1.96 \leq Z \leq 1.96) = 0.95$, we get the approximate 95% confidence interval $\bar{y} \pm 1.96s/\sqrt{10}$, or 0.981 ± 0.011, or $0.970 \leq \mu \leq 0.992$. Evidently scale B consistently understates the weight but the bias in measuring the 1 kg weight is likely fairly small (about $1\% - 3\%$). It is important to keep in mind in general that although we might be able to find evidence against a given hypothesis, this does not mean that the differences found are of practical significance. For example a patient person willing to toss a particular coin one million times can almost certainly find evidence against $H_0 : P(\text{heads}) = \frac{1}{2}$. This does not mean that in a game involving a few dozens or hundreds of tosses that H_0 is not a tenable and useful approximation. 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.

The approach to testing hypothesis described above is very general and straightforward, but a few points should be stressed:

1. If the p-value is small (close to 0) then the test indicates **strong evidence against** H_0 in light of the observed data; this is often termed "statistically significant" evidence against H_0. Rough rules of thumb are that $p - value < 0.05$ provides moderately strong evidence against H_0 and that $p - value < 0.01$ provides strong evidence.

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. 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 5.1.1 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 ("highly statistically significant" evidence) against a hypothesis H_0, there is no implication about how "wrong" H_0 is. For example in Example 5.3.1 there was strong evidence that scale B was biased (that is, strong evidence against $H_0 : bias = 0$), but the relative magnitude $(1 - 3\%)$ of the bias is apparently small. In practice, we try to supplement a significant test with an interval estimate that indicates the magnitude of the departure from H_0. This is how we check whether a result is **" scientifically" significant** as well as **statistically significant**.

4. So far we have not refined the conclusion when we do find strong evidence against the null hypothesis. Often we have in mind an "alternative" hypothesis. For example if the standard treatment for pain provides relief in about 50% of cases, and we test, for patients medicated with an alternative $H_0 : P(\text{relief}) = \frac{1}{2}$ we will obviously wish to know, if we find strong evidence against H_0, in what direction that evidence lies. If the probability of relief is greater than $\frac{1}{2}$ we might consider further tests or adopting the drug, but if it is less, then the drug will be abandoned for this purpose. We will try and adapt to this type of problem with our choice of discrepancy measure D.

A drawback with the approach to testing described so far is that we are not told how to construct the test statistic or discrepancy measure D. Often there are "intuitively obvious" statistics that can be used; this is the case in most examples in this section. However, In the next section we show 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 (5.1.1) for the observed data. Calculating probabilities involving D brings us back to distribution theory: in most cases the exact probability (5.1.1) is hard to determine mathematically, and we must either use an approximation or use computer simulation. Fortunately, for the tests in the next section we can use approximations based on χ^2 distributions.

5.2 Likelihood Ratios and Testing Statistical Hypotheses

Likelihood Ratios and Testing a Hypothesis for a Single Parameter

In Chapter 2 we used likelihood functions to gauge the plausibility of parameter values in the light of the obverved data. It should seem natural, then, to base a test of hypothesis on a likelihood or, in comparing the plausibility of two values, a ratio of the likelihoods. 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 at hand, it would seem reasonable to attempt to settle this argument using the ratio of the likelihood at these two values, i.e.

$$L(\theta_0)/L(\theta_1). \tag{5.3}$$

As usual we define the likelihood function $L(\theta) = L(\theta; y) = f(y; \theta)$ where $f(y; \theta)$ is the probability density function or probability function of the random variable Y representing the data and y is the observed value of the data. Let us now consider testing the plausibility of my hypothesized value θ_0 against an unpecified alternative. In this case it is natural to replace θ_1 in (5.3) by the value which appears most plausible given the data, i.e. its maximum likelihood estimate $\hat{\theta}$ for which

$$L(\hat{\theta}) = \max_{\theta \in \Omega} L(\theta).$$

The resulting likelihood ratio we recognise as the value of the relative likelihood function at θ_0,

$$R(\theta_0) = L(\theta_0)/L(\hat{\theta}).$$

If $R(\theta_0)$ is close to one, then θ_0 is plausible in the light of the observed data, but if $R(\theta_0)$ is very small and close to 0, then θ_0 is not plausible in the light of the observed data and this suggests evidence against H_0. Therefore the corresponding random variable, $L(\theta_0)/L(\tilde{\theta})$, [21] 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 $L(\theta_0)/L(\tilde{\theta})$ since it leads to a well-known distribution, the chi-squared distribution. We use the likelihood ratio statistic which was introduced in Chapter 4:

$$\Lambda = \Lambda(\theta_0) = -2\log\left[L(\theta_0)/L(\tilde{\theta})\right] = 2\ell(\tilde{\theta}) - 2\ell(\theta_0). \tag{5.4}$$

We choose this particular function because, when $H_0 : \theta = \theta_0$ is true, Λ has a approximately a chi-squared distribution with 1 degree of freedom. 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 $H_0 : \theta = \theta_0$. To determine the p-value we first calculate the observed value of $\Lambda(\theta_0)$, denoted by λ and given by

$$\lambda = \lambda(\theta_0) = -2\log\left[L(\theta_0)/L(\hat{\theta})\right] = 2\ell(\hat{\theta}) - 2\ell(\theta_0)$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ based on the observed data. The approximate p-value is then

$$p - value \approx P\left[W > \lambda(\theta_0)\right] \tag{5.5}$$

where $W \sim \chi^2(1)$.

Let us summarize the contruction 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 $f(y;\theta)$ 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

[21] 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 $\tilde{\theta} = g(\mathbf{Y})$.

values of Λ correspond to a disagreement between the hypothesis H_0 and the data and so provide evidence against H_0. Moreover it can be shown that Λ has approximately a $\chi^2(1)$ distribution so the 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 appraoches $\chi^2(1)$ as the size of the data set becomes large.

General Case: Multidimensional parameter θ

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,

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}).$$

Similary 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 $\boldsymbol{\theta}$ contrained to lie in the set $\Omega_0 \subset \Omega$) so that

$$L(\hat{\boldsymbol{\theta}}_0) = \max_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta}).$$

Now consider the corresponding statistic (random variable)

$$\Lambda = 2\ell(\tilde{\boldsymbol{\theta}}) - 2\ell(\tilde{\boldsymbol{\theta}}_0) = -2\log\left[\frac{L(\tilde{\boldsymbol{\theta}}_0)}{L(\tilde{\boldsymbol{\theta}})}\right] \tag{5.6}$$

and let

$$\lambda = 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\hat{\boldsymbol{\theta}}_0) = -2\log\left[\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right]$$

denote an observed value of Λ. If the observed value λ is very small then there is evidence against H_0 (again you should determine why this is true). In this case it can be shown that under H_0, the distribution of Λ 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

$$p - value = P(\Lambda \geq \lambda;\ H_0) \approx P(W \geq \lambda) \tag{5.7}$$

where $W \sim \chi^2(k-p)$.

Some Examples

The likelihood ratio test covers a great many different types of examples, but we only provide a few here.

Example 5.2.1. Lifetimes of light bulbs; A single parameter exponential model. Test $H_0 : \theta = \theta_0$ (a given value) based on a random sample $y_1, ..., y_n$. Thus $\Omega = \{\theta : \theta > 0\}$, $\Omega_0 = \{\theta_0\}$; $k = 1$, $p = 0$ and

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta) \quad \text{for } \theta > 0.$$

The variability in lifetimes of light bulbs (in hours, say, of operation before failure) is often well described by an Exponential distribution with probability density function

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

where $\theta = E(Y) > 0$ is the average (mean) lifetime. Here $\Omega = \{\theta : \theta > 0\}$ is the set of possible values for the parameter θ since mean lifetimes must be positive. A manufacturer claims that the mean life of a particular brand of bulbs is 2000 hours. We can examine that claim by testing the hypothesis

$$H_0 : \theta = 2000$$

assuming that the Exponential model applies.

Suppose for illustration that a random sample of $n = 20$ light bulbs was tested over a long period and that the total of the lifetimes $y_1, \ldots, y_{20}$ was observed to be $\sum_{i=1}^{20} y_i = 38,524$ hours. (It turns out that for the test below we need only the value of $\sum_{i=1}^{20} y_i$ and not the individual lifetimes $y_1, \ldots, y_{20}$ so we haven't bothered to list them. They would be needed, however to check that the exponential model was satisfactory.) Let us carry out a likelihood ratio test of H_0. The likelihood function based on a random sample $y_1, \ldots, y_n$ is

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

Note that in terms of our general theory the parameter space of θ is $\Omega = \{\theta : \theta > 0\}$ and the parameter space under H_0 is the single point $\Omega = \{2000\}$[22]. We use the likelihood ratio statistic Λ of (5.4) as our test statistic D. To evaluate this we first write down the log likelihood function (noting that $n = 20$ and $\sum_i = 1^2 0 y_i = 38524$)

$$\ell(\theta) = -20 \log \theta - \frac{38524}{\theta} \quad \text{for } \theta > 0.$$

Next, we obtain $\hat{\theta}$ by maximizing $\ell(\theta)$: this gives

$$\hat{\theta} = \frac{38524}{20} = 1926.2 \text{ hours.}$$

[22] The dimensions of Ω and Ω_0 are 1 and 0, respectively.

Now we can compute the observed value of Λ from (5.4) as

$$\begin{aligned}
\lambda &= 2\ell(\hat{\theta}) - 2\ell(2000) \\
&= -40\log(\hat{\theta}/2000) - \frac{77048}{\hat{\theta}} + \frac{77048}{2000} \\
&= 0.028
\end{aligned}$$

The final computational step is to compute the p-value, which we do using the χ^2 approximation (5.5). This gives

$$\begin{aligned}
p - value &= P(\Lambda \geq 0.028) \quad \text{assuming } H_0 \text{ is true} \\
&\approx P(W \geq 0.028) \quad \text{where } W \sim \chi^2(1) \\
&= 0.87
\end{aligned}$$

The p-value is not close to zero so we conclude that there is no evidence against H_0 and no evidence against the manufacturer's claim that θ is 2000 hours. Although the maximum likelihood estimate $\hat{\theta}$ was under 2000 hours (1926.2) it was not sufficiently under to give evidence against $H_0 : \theta = 2000$.

Example 5.2.2 Comparison of two parameters: two Poisson means. In problem 6 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

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

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 $(y_1, \ldots, y_m)$ 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

$$L_1(\mu_A) = \prod_{i=1}^{n} f(x_i; \mu_A) = \prod_{i=1}^{n} \frac{\mu_A^{x_i} \exp(-\mu_A)}{x_i!} \quad \text{for } \mu_A > 0$$

and the likelihood function for μ_B only based on $(y_1, \ldots, y_m)$ is

$$L_2(\mu_B) = \prod_{i=1}^{m} f(y_i; \mu_B) = \prod_{j=1}^{m} \frac{\mu_B^{y_j} \exp(-\mu_B)}{y_j!} \quad \text{for } \mu_B > 0.$$

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

$$\begin{aligned} L(\boldsymbol{\theta}) &= L(\mu_A, \mu_B) = L_1(\mu_A) \times L_2(\mu_B) \\ &= \prod_{i=1}^{n} \frac{\mu_A^{x_i} \exp(-\mu_A)}{x_i!} \prod_{j=1}^{m} \frac{\mu_B^{y_j} \exp(-\mu_B)}{y_j!} \\ &= c \times \exp(-n\mu_A - m\mu_B)\, \mu_A^{\sum\limits_{i=1}^{n} x_i} \mu_B^{\sum\limits_{j=1}^{m} y_j} \quad \text{for } (\mu_A, \mu_B) \in \Omega \end{aligned}$$

and the log likelihood function for $\boldsymbol{\theta} = (\mu_A, \mu_B)$ is

$$l(\boldsymbol{\theta}) = -n\mu_A - m\mu_B + \left(\sum_{i=1}^{n} x_i\right) \log \mu_A + \left(\sum_{j=1}^{m} y_j\right) \log \mu_B + \log c \tag{5.8}$$

where

$$c = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \left(\prod_{j=1}^{m} \frac{1}{x_j!}\right)$$

does not depend on $\boldsymbol{\theta}$.

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

$$\ell(\boldsymbol{\theta}) = \ell(\mu_A, \mu_B) = -12\mu_A + 240 \log \mu_A - 12\mu_B + 140 \log \mu_B + \log c \quad \text{for } (\mu_A, \mu_B) \in \Omega$$

The values of μ_A and μ_B which maximize $\ell(\mu_A, \mu_B)$ are obtained by solving the two equations[23]

$$\frac{\partial \ell}{\partial \mu_A} = 0, \qquad \frac{\partial \ell}{\partial \mu_B} = 0,$$

which gives two equations in two unknowns:

$$\begin{aligned} -12 + \frac{240}{\mu_A} &= 0 \\ -12 + \frac{140}{\mu_B} &= 0 \end{aligned}$$

The maximum likelihood estimates of μ_A and μ_B (unconstrained) are $\hat{\mu}_A = 240/12 = 20.0$ 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 $\ell(\mu_A, \mu_B)$ under the constraint $\mu_A = \mu_B$. To do this we merely let $\mu = \mu_A = \mu_B$ in (5.8) to obtain

$$\begin{aligned} \ell(\mu, \mu) &= -12\mu + 240 \log \mu - 12\mu + 140 \log \mu \\ &= -24\mu + 380 \log \mu \quad \text{for } \mu > 0. \end{aligned}$$

Solving $\partial \ell(\mu, \mu)/\partial \mu = 0$, 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 to observed value of the likelihood ratio statistic, which from (5.6) is

$$\begin{aligned} \lambda &= 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\hat{\boldsymbol{\theta}}_0) \\ &= 2\ell(20.0, 11.667) - 2\ell(15.833, 15.833) \\ &= 2\,(682.92 - 669.60) \\ &= 26.64 \end{aligned}$$

Finally, we compute the approximate p-value for the test, which by (5.7) is

$$\begin{aligned} &P(\Lambda \geq 26.64;\ H_0 \text{ is true}) \\ &\quad \approx P(W \geq 26.64) \quad \text{where } W \sim \chi^2\,(1) \\ &\quad = 0.25 \times 10^{-7} \end{aligned}$$

[23] think of this as maximizing over each parameter with the other parameter fixed.

Our conclusion is that there is very strong evidence against the hypothesis $H_0 : \mu_A = \mu_B$; these data, through the lens of a likelihood ratio test, 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 $\mu_A - \mu_B$; 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 confidence intervals in order to express the uncertainty in such estimates.)

Example 5.2.3 Likelihood ratio tests of hypotheses for the Gaussian distribution: Suppose $Y \sim G(\mu, \sigma)$ with probability density function

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

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 a sample $Y_1, Y_2, ..., Y_n$ from this distribution is

$$L(\mu) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right] \quad \text{for } \mu > 0$$

and the log likelihood function is

$$\ell(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + c$$

where

$$c = \log\left[(2\pi)^{-n/2}\sigma^{-n}\right]$$

does not depend on μ. In order to maximize the log likelihood with respect to μ, we need only minimize the quantity

$$\sum_{i=1}^{n}(y_i - \mu)^2$$

and differentiating this with respect to μ and setting the derivative equal to 0 gives

$$-2\sum_{i=1}^{n}(y_i - \mu) = 0.$$

Solving this for μ gives the maximum likelihood estimate $\hat{\mu} = \bar{y}$ and

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

the correponding maximum likelihood estimator of μ. Note that the log likelihood can be written as

$$\begin{aligned}\ell(\mu) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + c \\ &= \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right] + c\end{aligned}$$

where we have used the algebraic identity[24]

$$\sum_{i=1}^{n}(y_i-\mu)^2=\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-\mu)^2.$$

To test the hypothesis $H_0:\mu=\mu_0$ we use the likelihood ratio statistic

$$\begin{aligned}\Lambda&=2\ell(\tilde{\mu})-2\ell(\mu_0)\\&=\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu_0)^2-\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\tilde{\mu})^2\\&=\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu_0)^2-\sum_{i=1}^{n}(Y_i-\tilde{\mu})^2-n(\bar{Y}-\tilde{\mu})^2\right]\qquad(5.9)\\&=\frac{1}{\sigma^2}n(\bar{Y}-\mu_0)^2\qquad\text{since }\tilde{\mu}=\bar{Y}\\&=\left(\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}\right)^2.\qquad(5.10)\end{aligned}$$

The purpose for writing the likelihood ratio statistic in the form (5.10) is to draw attention to the fact that it is the square of the standard normal random variable $\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}$ and is therefore has exactly a chi-squared distribution with degrees of freedom equal to 1. 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 Λ is clearly $\chi^2(1)$ (not only asymptotically but for any value of n) from a basic property of this distribution.

We now proceed to a more interesting and more practical example for the normal distribution, which involves testing a hypothesis for one of the parameters when both are unknown. Consider for example a test of $H_0:\sigma=\sigma_0$ based on a random sample $y_1,y_2,...,y_n$. 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

$$L(\boldsymbol{\theta})=L(\mu,\sigma)=\prod_{i=1}^{n}f(Y_i;\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{Y_i-\mu}{\sigma}\right)^2}$$

and the log likelihood function is

$$\ell(\mu,\sigma)=-n\log(\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2+c$$

where

$$c=\log\left[(2\pi)^{-n/2}\right]$$

[24] 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

does not depend on μ or σ. The maximum likelihood estimators of (μ, σ) in the unconstrained case are

$$\begin{aligned}\tilde{\mu} &= \bar{Y}\\ \tilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.\end{aligned}$$

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

$$\begin{aligned}\Lambda &= 2\ell(\bar{Y}, \tilde{\sigma}) - 2\ell(\bar{Y}, \sigma_0)\\ &= -2n\log(\tilde{\sigma}) - \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2n\log(\sigma_0) + \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\\ &= 2n\log(\sigma_0/\tilde{\sigma}) + \left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}^2}\right)n\tilde{\sigma}^2\\ &= n\left[\log(\sigma_0^2/\tilde{\sigma}^2) + \left(\frac{\tilde{\sigma}^2}{\sigma_0^2} - 1\right)\right].\end{aligned}$$

This is not as obviously a chi-squared random variable as in the last case but it is, as one might expect, a function[25] of 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 Λ increases as the quantity $\tilde{\sigma}^2/\sigma_0^2$ gets further away from 1 in either direction. The test proceeds by obtaining the observed value of Λ:

$$\lambda = n\left[\log(\sigma_0^2/\hat{\sigma}^2) + \left(\frac{\hat{\sigma}^2}{\sigma_0^2} - 1\right)\right]$$

and then obtaining and interpreting the p-value

$$P(W > \lambda)$$

where $W \backsim \chi^2(1)$.

Example: Testing for multinomial probabilities. Consider a random vector $\mathbf{Y} = (Y_1, ..., Y_m)$ with Multinomial probability function:

$$f(y_1, ..., y_m; \theta_1, ..., \theta_m) = \frac{n!}{y_1! \cdots y_m!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_m^{y_m} \quad \text{for } 0 \le y_j \le n \text{ where } \sum_{j=1}^{m} y_j = n.$$

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 $\dim(\alpha) = p < m - 1$. Thus, the parameter in the unconstrained model is $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$ and the parameter space $\Omega = \{(\theta_1, ..., \theta_m) : 0 \le \theta_j \le 1, \text{ where } \sum_{j=1}^{m}\theta_j = 1\}$ has dimension

[25] $g(x) = x - 1 - \log(x)$

$m-1$ and in the constrained model $\Omega_0 = \{(\theta_1(\boldsymbol{\alpha}), .., \theta_m(\boldsymbol{\alpha})) : \text{for all } \boldsymbol{\alpha}\}$ has dimension p . The likelihood ratio statistic is constructed from the likelihood function:

$$L(\boldsymbol{\theta}) = f(y_1, ..., y_m; \boldsymbol{\theta}).$$

We will give more specific examples of the Multinomial in Chapter 6. We conclude with some short remarks about the relationship between tests of hypothesis and interval estimation.

5.3 Hypothesis Testing and Interval Estimation

Hypothesis tests for hypotheses of the form $H_0 : \theta = \theta_0$, where θ is a scalar parameter, are very closely related to interval estimates for θ. For likelihood ratio tests the connection is immediately obvious, because the likelihood ratio statistic is

$$\Lambda = 2\ell(\tilde{\theta}) - 2\ell(\theta_0)$$

is used for both tests and confidence intervals. For a test of $H_0 : \theta = \theta_0$, the p-value is approximately given by

$$p - value = P\left[W \geq \lambda(\theta_0)\right] \tag{5.11}$$

where $\lambda(\theta_0) = 2\ell(\hat{\theta}) - 2\ell(\theta_0)$ and $W \backsim \chi^2\,(1)$. We write $\lambda(\theta_0)$ to remind ourselves that we are testing $H_0 : \theta = \theta_0$. On the other hand, to get an approximate $100q\%$ confidence interval for θ we find by all values of θ such that

$$\lambda(\theta_0) = 2\ell(\hat{\theta}) - 2\ell(\theta_0) \leq c \tag{5.12}$$

where $P\,(W \leq c) = q$ or $P\,(W > c) = 1-q$ and $W \backsim \chi^2\,(1)$. For example for an approximate 95% confidence interval we use $c = 3.84$.

We now see the following by comparing (5.11) and (5.12): The parameter value θ_0 is inside an approximate $100q\%$ confidence interval given by (5.12) if and only if for the test of $H_0 : \theta = \theta_0$ we have by (5.11) that the $p - value$ is greater than or equal to $1 - q$.

For example, θ_0 is **inside** the approximate 95% confidence interval if and only if the p-value for $H_0 : \theta = \theta_0$ satisfies $p - value \geq 0.05$. To see this note that

$$\begin{gathered} p - value \geq 0.05 \\ \text{if and only if } P\left[W \geq \lambda(\theta_0)\right] \geq 0.05 \\ \text{if and only if } \lambda(\theta_0) \leq 3.84 \\ \text{if and only if } \theta_0 \text{ is inside the approximate 95\% confidence interval.} \end{gathered}$$

The connection between tests and confidence intervals can also be made when other test statistics beside the likelihood ratio statistic are used. If D is a test statistic for testing $H_0 : \theta = \theta_0$ then we can obtain a 95% confidence interval for θ by finding all values θ_0 such that $p - value \geq 0.05$.

5.4 Problems

1. 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.

2. Refer back to Problem 1 in Chapter 1. Frame this problem as a hypothesis test. What test statistic is being used? What are the significance levels from the data in parts (b) and (c)?

3. 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 $y_1, \cdots, y_n$.

   (a) Give a test statistic which could be used to test that the y_i's $(i = 1, \cdots, n)$ are consistent with a random sample from Uniform$(0, 1)$.

   (b) Generate 1000 y_i's and carry out the test in (a).

4. 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 $y_1, \ldots, y_{20}$ 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 Y_i are known to be approximately Exponential with mean θ.

(a) Use a likelihood ratio test to test the hypothesis that $\theta = 1000$ hours. Is there any evidence that the company's power systems do not meet the contracted standard?

(b) If you were a personal computer manufacturer using these power systems, would you like the company to perform any other statistical analyses besides testing $H_0 : \theta = 1000$? Why?

5. 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 $\theta_j = P($digit j 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 a 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: You need to consider $P($largest of 6 independent Λ's is $\geq \lambda)$.)

6. **Testing a genetic model**. Recall the model for the M-N blood types of people, discussed in Examples 2.3.2 and 2.5.2. In a study involving a random sample of n persons the numbers Y_1, Y_2, Y_3 $(Y_1 + Y_2 + Y_3 = n)$ who have blood types MM, MN and NN respectively has a Multinomial distribution with joint probability function

$$f(y_1, y_2, y_3) = \frac{n!}{y_1!, y_2!, y_3!} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3} \quad \text{for } y_j = 0, 1, \ldots; \quad \sum_{j=1}^{3} y_j = n$$

and since $\theta_1 + \theta_2 + \theta_3 = 1$ the parameter space $\Omega = \{(\theta_1, \theta_2, \theta_3) : \theta_j \geq 0, \ \sum_{j=1}^{3} p_j = 1\}$ has dimension 2. The genetic model discussed earlier specified that $\theta_1, \theta_2, \theta_3$ can be expressed in terms of only a single parameter α, $0 < \alpha < 1$, as follows:

$$\theta_1 = \alpha^2, \quad \theta_2 = 2\alpha(1 - \alpha), \quad \theta_3 = (1 - \alpha)^2 \tag{5.13}$$

Consider (5.13) as the hypothesis H_0 to be tested. In this case, the dimension of the parameter space for $(\theta_1, \theta_2, \theta_3)$ under H_0 is 1, and the general methodology of likelihood ratio tests can be applied. This gives a test of the adequacy of the genetic model.

Suppose that a sample with $n = 100$ persons gave observed values $y_1 = 18$, $y_2 = 50$, $y_3 = 32$. Test the hypothesis (5.13) and state your conclusion.

7. **Likelihood ratio test for a Gaussian mean.** Suppose that a random variable Y has a $G(\mu, \sigma)$ distribution and that we want to test the hypothesis $H_0 : \mu = \mu_0$, where μ_0 is some specified number. The value of σ is unknown.

   (a) Construct the likelihood ratio statistic Λ for this hypothesis. (Note that the parameter space is $\Omega = \{\boldsymbol{\theta} = (\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$.) Assume that a random sample $y_1, \ldots, y_n$ is available.

   (b) Show that Λ can be expressed as a function of $T = \sqrt{n}(\bar{Y} - \mu_0)/S$, where

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

   is the sample variance and $\bar{Y}$ is the sample mean. Note: you will want to use the identity

   $$\sum_{i=1}^{n}(Y_i - \mu_0)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2.$$

8. 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 $P_1, \ldots, P_K$ and let θ_j, $j = 1, \ldots, K$ 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 $y_1, \ldots, y_K$ for a given t-year period. Describe how to test the hypothesis that $\theta_1 = \theta_2 = \ldots = \theta_K$.

   (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.

y_j:	2025	1116	3210	1687	2840
P_j:	27	18	41	29	31

GAUSSIAN RESPONSE MODELS

6.1 Introduction

A "response" variable Y is one whose distribution has parameters which depend on the value of other variables. We have already seen many examples of such forms of dependence in the data in these notes such as heights and body-mass index measurements of people. Many problems involve explanatory variables x (which may be a vector) that are related to a response Y which is often assumed to have a Normal or Gaussian distribution, and we will call these models Gaussian response models. These are by far the most common models in applications of statistics.

For many of the models we have studied in these notes, we assumed that we had a random sample $Y_1, Y_2, ..., Y_n$, i.e. independent random variables 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 some other vector $\mathbf{x}_i$ of covariates (explanatory variables which we measure). In other words we will usually assume that

$$Y_i \sim G\left(\mu\left(\mathbf{x}_i\right), \sigma\right) \quad \text{for } i = 1, \ldots, n \ \text{ independently}$$

where $\mu(\mathbf{x})$ is some function of the covariate $\mathbf{x}$. Notice that the assumed model is such that the mean of Y_i depends on the covariate $\mathbf{x}_i$ corresponding to the response Y_i but the standard deviation σ does not depend on $\mathbf{x}_i$; it is the same for all values of i. While this last assumption is not essential, it does make the models easier to analyze so we will generally use it here.

So the difference between various Gaussian response models is in the choice of covariates and the function $\mu(\mathbf{x})$. Often $\mathbf{x}$ consists of a vector of covariates and it is natural and simple to assume that $\mu(\mathbf{x})$ is a linear function of the components of this vector. The choice of $\mu(\mathbf{x})$ is guided by past information and on current data from the population or process in question.

Here are some examples of settings where Gaussian response models can be used.

Example 6.1.1 The soft drink bottle filling process of Example 1.4.2 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: $Y \sim G(\mu, \sigma)$.

In this case we can think of the machines as being like a covariate, with μ and σ differing for the two machines. We could write

$$\begin{aligned} Y &\sim G(\mu_O, \sigma_O) \text{ for observations from the old machine} \\ Y &\sim G(\mu_N, \sigma_N) \text{ for observations from the new machine.} \end{aligned}$$

In this case there is no formula relating μ and σ to the machines; they are simply different. Notice that an important feature of a machine is the variability of its production so we have, in this case, permitted the two variance parameters to be different.

Example 6.1.2 Price versus Size of Commercial Buildings [26]

Ontario property taxes are based on "market value", which is determined by comparing a property to the price of those which have recently been sold. The value of a property is separated into components for land and for buildings. Here we deal with the value of the buildings only but a similar analysis could be conducted for the value of the property.

A manufacturing company was appealing the assessed market value of its property, which included a large building. Sales records were collected on the 30 largest buildings sold in the previous three years in the area. The data are given in Table 6.1.1 and plotted in Figure 6.2 in a **scatter plot**, which is a plot of the points (x_i, y_i). 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 variable 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.

Table 6.1.1 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

The scatter plot shows that price (y) is roughly inversely proportional to size (x) but there is obviously variability in the price of buildings having the same area (size). In this

[26] This reference can be found in earlier course notes for Oldford and MacKay, STAT 231 Ch. 16

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

$$Y_i \sim G(\beta_0 + \beta_1 x_i, \sigma) \quad \text{for } i = 1, \ldots, n \ \text{ independently}$$

where β_0 and β_1 are parameters. Again we assumed a common standard deviation σ for the observations.

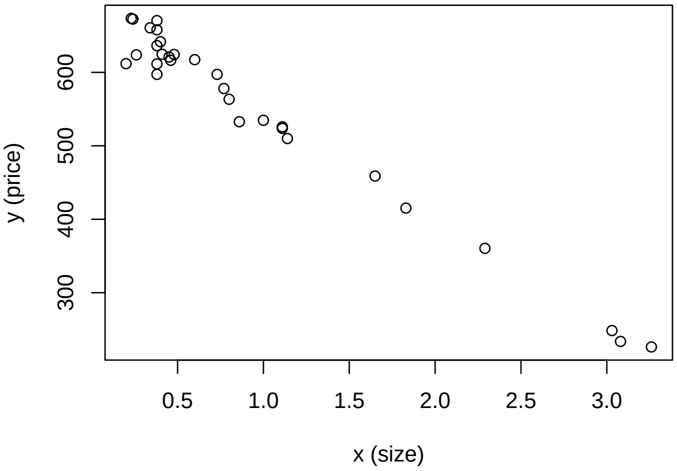


Figure 6.2: **Scatter Plot of Size vs. Price for 30 Buildings**

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: Two bolts of the same dimension and material will generally break at different forces. Understanding the distribution of breaking strengths is very important in construction and other areas.

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 6.3

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 scatter plot 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 $Y \sim G(g(x), \sigma)$. 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 $g(x)$ would be; the relationship looks slightly nonlinear so we might try a quadratic function

$$g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

or some other nonlinear function. The parameters are the constants $\beta_0, \beta_1, \beta_2$ which are unknown and to be estimated from the data.

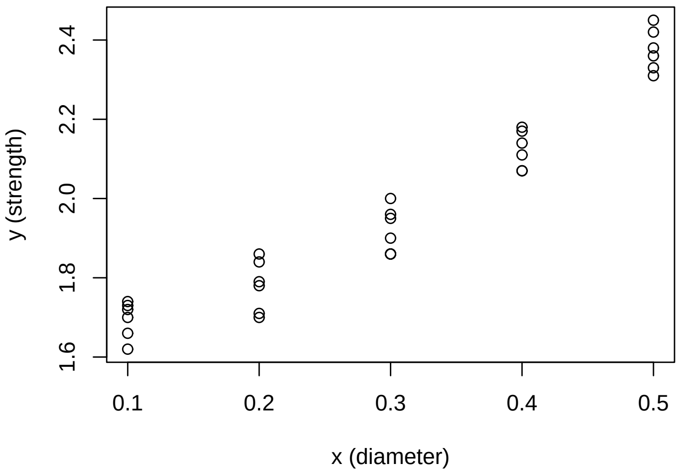


Figure 6.3: **Scatter Plot of Diameter vs. Strength for Steel Bolts.**

Definition 2 *A **Gaussian response model** is one for which the distribution of the response variable Y, **given** the associated vector of covariates $\mathbf{x} = (x_1, x_2, ..., x_k)$ for an* *individual unit, is of the form*

$$Y \sim G(\mu\,(\mathbf{x})\,, \sigma(\mathbf{x})) \tag{6.2}$$

where here we allow the more general case in which the standard deviation can depend on the covariates as well.

If observations are made on n randomly selected units we often write this as

$$Y_i \sim G\left(\mu(\mathbf{x}_i), \sigma(\mathbf{x}_i)\right) \quad \text{for } i = 1, \ldots, n \text{ independently}$$

In most examples here, we will use models where $\sigma(\mathbf{x}_i) = \sigma$ is constant and $g(\mathbf{x}_i)$ is a **linear function** of the covariates. These models are called **Gaussian linear models** and can be written

$$Y_i \sim G\left(\mu(\mathbf{x}_i), \sigma\right) \text{ for } i = 1, \ldots, n \text{ independently} \tag{6.3}$$
$$\text{with } \mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij},$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is the vector of 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**[27] **models**, and the β_j's are called the **regression coefficients**.
Remark: Sometimes the model (6.3) 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.

The model (6.3) describes many situations well. The following are some illustrations.

1. $Y_i \sim G(\mu, \sigma)$, where Y_i is the height of a random female, corresponds $\mu\,(\mathbf{x}_i) = \beta_0 = \mu$, $i = 1, 2, ..., n$.

2. The model in Example 6.1.2 had $\mu(\mathbf{x}_i) = \mu(x_i) = \beta_0 + \beta_1 x_i$ where x_i is the size of the ith building, $i = 1, 2, ..., n$.

3. The bolt strength model in Example 6.1.3 had $\mu(\mathbf{x}_i) = \mu(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ where x_i is the diameter of the ith bolt, $i = 1, 2, ..., n$..

We know consider estimation and testing procedures for these Gaussian response models. We begin with models that have no covariates so that the observations are all from the same Gaussian distribution.

[27] The term "regression" is used because it was introduced in the 19th century in connection with these models, but we will not explain why it was used here. It is called "linear" because it is linear in the parameters β_i.

6.2 Inference for a single sample from a Gaussian Distribution

Suppose that $Y \sim G(\mu, \sigma)$ models a response variable y in some population or process. A random sample $Y_1, \ldots, Y_n$ is selected, and we want to estimate the model parameters and possibly to test hypotheses about them. We have already seen in Section 2.2 that the maximum likelihood estimators of μ and σ^2 are

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

A closely related point estimator of σ^2 is the sample variance[28],

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

whch 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 $\tilde{\sigma}^2$. We now consider interval estimation and tests of hyptheses for μ and σ.

6.2.1 Confidence Intervals and Tests for μ and σ

If σ were known then, as discussed in Chapter 4,

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

would be a pivotal quantity and could be used to get 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 which we will denote as T. We will write T in terms of S since the formulas below look a little simpler in this case, so T is defined as

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \tag{6.4}$$

Since S,unlike σ, is a random variable in (6.4) the distribution of T is not exactly the $G(0, 1)$. It turns out that its distribution is what is known as a **Student t** (or just "t") distribution. We will digress briefly to present this distribution and show how it arises.

Student t Distribution.

[28] The sample variance has the advantage that it is "unbiased" i.e. that $E(S^2) = \sigma^2$. To see this note that $E(S^2) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(Y_i - \mu)^2 - n(\bar{Y} - \mu)^2\right] = \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2$

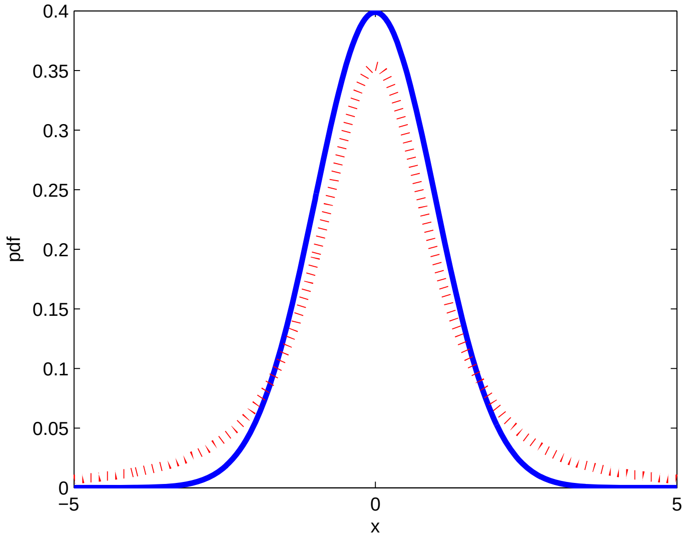


Figure 6.4: p.d.f.'s of the $t\,(2)$ distribution (dotted red) and the $G\,(0,1)$ distribution (solid blue)

The Student t distribution[29] (t distribution for short) has probability density function

$$f(x;k) = c_k \times (1+\frac{x^2}{k})^{-(k+1)/2} \quad \text{for } -\infty < x < \infty \quad \text{and } k = 1, 2, \ldots$$

The parameter k is called the degrees of freedom. The constant c_k is

$$c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma(\frac{k}{2})}$$

We write $T \sim t\,(k)$ to denote that the random variable T has a Student t distribution with k degrees of freedom.

In Figure 6.4 the probability density function $f(x;k)$ for $k = 2$ is plotted together with the $G\,(0,1)$ probability density function.Obviously the Student t probability density function is similar to that of the $G\,(0,1)$ distribution in several respects: it is symmetric

[29] This distribution arises when we consider independent random variables $Z \sim G(0,1)$ and $U \sim \chi^2\,(k)$ and then define the new random variable $T = \frac{Z}{(U/k)^{1/2}}$. Then T has a **student -t distribution with k degrees if freedom**. In this case

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

(ii) $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2\,(n-1)$

(iii) $\bar{Y}$ and S^2 (and therefore U) are independent.

We will not prove these results here.

about the origin, it is unimodal, and indeed for large degrees of freedom k, the graph of the probability density function $f(x;k)$ is indistinguishable from that of the $G(0,1)$ probability density function. The primary difference, for small degrees of freedom such as the one plotted, is in the tails of the density. The Student t density has larger "tails" or more area in the extreme left and right tails, which means it is more prone to large or small values than is the standard normal. Problem 1 at the end of this chapter considers some properties of $f(x;k)$.

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

There is one fundamental reason that the t distribution is an essential tool of any statistician, and it is that in the Gaussian case, (6.4) has a t distribution.

Theorem 3 *Suppose $Y_1, \ldots, Y_n$ is a random sample from a common Gaussian distribution $Y_i \sim G(\mu, \sigma)$ having sample mean $\bar{Y}$ and sample variance S^2. Then the statistic (6.4) has the t distribution with $k = n-1$ degrees of freedom.*

Confidence Intervals for μ

Since $T = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom which is a completely known distribution, T is a pivotal quantity and we can use it to construct a $100p\%$ confidence interval for μ. First we obtain constants a_1 and a_2 such that $P(a_1 \leq T \leq a_2) = p$ using t tables or *R*, and then we solve the inequality

$$a_1 \leq \frac{\bar{y}-\mu}{s/\sqrt{n}} \leq a_2 \tag{6.5}$$

for *the only unknown, μ*. (Note that if we attempted to use the Gaussian distributed random variable $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ to build a confidence interval we would have two unknowns in the inequality since both μ and σ are unknown.) Solving (6.5) we obtain the $100p\%$ confidence interval for μ as

$$\bar{y} - a_2 s/\sqrt{n} \leq \mu \leq \bar{y} - a_1 s/\sqrt{n}$$

and this is justified since

$$P\left(\bar{Y} - a_2 S/\sqrt{n} \leq \mu \leq \bar{Y} - a_1 S/\sqrt{n}\right) = p.$$

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. Notice that since the t distribution is symmetric about zero, we can choose $-a_1 = a_2 = a$. The $100p\%$ confidence interval (usually referred to as a "two-sided" interval) is then

$$\bar{y} \pm a\left(\frac{s}{\sqrt{n}}\right) \tag{6.6}$$

where $P(T > a) = p/2$ and $T \sim t(n-1)$. We note that this interval is of the form $\bar{y} - as/\sqrt{n} \leq \mu \leq \bar{y} - as/\sqrt{n}$ 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 6.2.1 Scores Y for an IQ test administered to ten year olds 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.

We wish to estimate the parameter μ for this school based on these data. We obtain confidence intervals for the average IQ test score μ in the population by using the pivotal quantity

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

Since $P(-2.262 \leq T \leq 2.262) = 0.95$ for $T \sim t(9)$, a 95% confidence interval for μ is $\bar{y} \pm 2.262s/\sqrt{10}$. For the given data $\bar{y} = 104.6$ and $s = 8.57$, so the confidence interval is 104.6 ± 6.13, or $98.47 \leq \mu \leq 110.73$ or $[98.47, 110.73]$.

Behaviour as $n \to \infty$: As n increases, confidence intervals behave in a largely predictable fashion. First the estimated standard deviation gets closer to the true standard deviation σ[30]. 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 6.2.1 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. If we have a rough idea what the value of σ is, 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 study. A 95% confidence interval using the normal quantiles takes the form $\bar{y} \pm 1.96\sigma/\sqrt{n}$

[30] this will be justified shortly

so if we have a rough idea of the value of σ and if we wish a confidence interval of the form $\bar{y} \pm d$ (the width of the confidence interval is then $2d$), we should choose

$$\begin{aligned} 1.96\frac{\sigma}{\sqrt{n}} &\approx d \\ \text{or} \quad n &\approx \left(\frac{1.96\sigma}{d}\right)^2 . \end{aligned}$$

We would normally 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.

Hypothesis Tests for μ

For a normally distribued population, we may wish to test a hypothesis $H_0 : \mu = \mu_0$, where μ_0 is some specified value. To do this we can use the test statistic

$$D = \frac{|\bar{Y} - \mu_0|}{S/\sqrt{n}} \tag{6.7}$$

We then obtain a p-value from the t distribution as follows. Let

$$d = \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \tag{6.8}$$

be the value of D observed in a sample with mean $\bar{y}$ and standard deviation s, then

$$\begin{aligned} p - value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(|T| \geq d) \\ &= 1 - P(-d \leq T \leq d) \quad \text{where } T \sim t\,(n-1)\,. \end{aligned} \tag{6.9}$$

Example 6.2.2 For the setting in Example 6.2.1, test $H_0 : \mu = 110$. From (6.8), the observed value of D is

$$d = \frac{|104.6 - 110|}{8.57/\sqrt{10}} = 1.99$$

and by (6.9) the p-value is

$$\begin{aligned} p - value &= P(|T| \geq 1.99) \\ &= 1 - P(-1.99 \leq T \leq 1.99) \quad \text{where } T \sim t\,(9) \\ &= 0.078. \end{aligned}$$

Based on the observed data there is no strong evidence against $H_0 : \mu = 110$. (Such tests are sometimes used to compare IQ test scores for a sub-population (e.g. students in one school district) with a known mean μ for a "reference" population.)

Remark: The likelihood ratio statistic could also be used for testing $H_0 : \mu = \mu_0$ or constructing a confidence interval for μ, but the methods above are a little simpler. In fact, it can be shown that the likelihood ratio statistic for H_0 is a one-to-one function of $|\frac{\bar{Y}-\mu}{S/\sqrt{n}}|$; see Problem 2 at the end of this Chapter.

Remark: The function *t.test* in *R* will give confidence intervals and test hypotheses about μ; for a data set y use *t.test(y).*

Confidence Intervals and Tests for σ

Suppose that we have a sample $Y_1, Y_2, ..., Y_n$ of independent random varaibles each from the same $G(\mu, \sigma)$ distribution. We have seen that there are two closely related estimators for the population variance, $\tilde{\sigma}^2$ and the sample variance S^2. Suppose 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 4 *Suppose $Y_1, Y_2, ..., Y_n$ are independent random variables with common $G(\mu, \sigma)$ distribution and suppose S^2 is the sample variance. Then*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{6.10}$$

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 on the surface seems wrong since $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the sum of n squared normal random variables. Is this in direct contradiction to Theorem xxx? It is in fact true that each $(Y_i - \bar{Y})$ is a normally distributed random variable, but **not** in general standard normally distributed and more importantly **not independent**! It is easy to see that they are not independent since $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ implies that the last value can be determined using the sum of the first $n-1$ terms:

$$Y_n - \bar{Y} = -\sum_{i=1}^{n-1}(Y_i - \bar{Y}).$$

Although there are n terms $(Y_i - \bar{Y})$, $i = 1, 2, ..., n$ in the summand for S^2 there are really only $n-1$ that are *free* (i.e. *linearly independent*); the last is determined by the first $n-1$. 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, i.e. $Y_i - \bar{Y}$, $i = 1, ..., n$.

We will now show how we can use the above theorem to construct a $100p\%$ confidence interval for the parameter σ^2 or σ. First note that $(n-1)\, S^2/\sigma^2$ is a pivotal quantity since its distribution is completely known. Using chi-squared tables or R we can find constants a_1 and a_2 such that

$$P(a_1 \le U \le a_2) = p$$

where $U \sim \chi^2(n-1)$. The confidence interval is then obtained by solving the inequality

$$a_1 \le \frac{(n-1)s^2}{\sigma^2} \le a_2$$

for the parameter σ^2 where s^2 is the observed sample variance. We obtain

$$\frac{(n-1)s^2}{a_2} \le \sigma^2 \le \frac{(n-1)s^2}{a_1}. \tag{6.11}$$

Of course we can also solve (6.11) for the parameter σ to obtain

$$\sqrt{\frac{(n-1)\,s^2}{a_2}} \le \sigma \le \sqrt{\frac{(n-1)\,s^2}{a_1}} \tag{6.12}$$

so a $100p\%$ confidence interval for σ is $\left[\sqrt{\frac{(n-1)s^2}{a_2}},\ \sqrt{\frac{(n-1)s^2}{a_1}}\right]$. For such "two-sided" confidence intervals we usually choose a_1 and a_2 such that

$$P(U \le a_1) = P(U > a_2) = \frac{p}{2}$$

where $U \sim \chi^2(n-1)$. This choice of a_1, a_2 is not unique but traditional for two-sided intervals. Note that these two-sided confidence intervals are **not symmetric** about the estimate of σ.

In some applications we are interested in an upper bound on σ (because small σ is "good" in some sense); then we take $a_2 = \infty$ and find a_1 such that $P(U \le a_1) = p$ so that a "one-sided" $100p\%$ confidence interval for σ is $\left[0,\ s\sqrt{\frac{n-1}{a_1}}\right]$.

Example 6.2.3. 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, $Y \sim G(\mu, \sigma)$. 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

$$\bar{y} = 25.009 \text{ and } \sum_{i=1}^{15} (y_i - \bar{y})^2 = (14)\,s^2 = 0.002347.$$

To obtain a 95% confidence interval for σ, we use the pivotal quantity

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

with $n = 15$. From chi-squared tables or R we obtain

$$P\left(U \leq 5.63\right) = 0.025 = P\left(U > 26.12\right)$$

where $U \sim \chi^2\left(14\right)$ so that

$$\begin{aligned} 0.95 &= P\left(5.63 \leq \frac{(n-1)S^2}{\sigma^2} \leq 26.12\right) \\ &= P\left(\frac{(n-1)S^2}{26.12} \leq \sigma^2 \leq \frac{(n-1)^2}{5.63}\right). \end{aligned}$$

Substituting $a_1 = 5.63$, $a_2 = 26.12$ and $\left(14\right)s^2 = 0.002347$ into (6.12) we obtain

$$\sqrt{\frac{0.002347}{26.12}} \leq \sigma \leq \sqrt{\frac{0.002347}{5.63}}$$

so that a 95% confidence interval for σ is given by $0.0095 \leq \sigma \leq 0.0204$ or $[0.0095,\ 0.0204]$. 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 $\sigma \leq 0.0189$ and in this case 0.02 is not in the interval. Why are they different? Both cover the true value of the parameter σ for 95% of all samples so they have the same confidence coefficient but the one-sided interval, since it allows smaller (as small as zero) values on the left end of the interval, it can achieve the same coverage with a smaller right end-point. If our primary concern was for values of σ being too large, i.e. for an upper bound for the interval, then the one-sided interval is the one that should be used for this purpose.

Hypothesis Tests for σ. We have discussed constructing confidence intervals for the parameter σ for a Gaussian population but we may also wish to test a hypthesis that this parameter takes a specific value; $H_0 : \sigma = \sigma_0$. One approach is to use a likelihood ratio statistic, as described in Chapter 4. It can be shown (see Problem 2) that the likelihood ratio statistic Λ is a function of $U = (n-1)S^2/\sigma_0^2$ and in fact

$$\Lambda = U - n\log\left(\frac{U}{n}\right) - n \tag{6.13}$$

This is not a one-to-one function of U but Λ is zero when $U = n$ and Λ is large when U/n is much bigger than or much less than one (i.e. 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 Λ discussed in Chapter 4.

If we use the test statistic $U = (n-1)S^2/\sigma_0^2$ for testing $H_0 : \sigma = \sigma_0$ then it is clear that large values of U and small values of U provide evidence against H_0. 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 $u > n-1$ compute the p-value as
$$p - value = 2P(U \geq u)$$
where $U \sim \chi^2(n-1)$.
If $u < n-1$ compute the p-value as
$$p - value = 2P(U \leq u)$$
where $U \sim \chi^2(n-1)$.

Example 6.2.4 For the manufacturing process in Example 6.2.3, 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 6.2.3, the p-value for a test of H_0 based on the test statistic Λ (or equivalently, $U = (n-1)S^2/\sigma_0^2$) will be less than 0.05. To find the p-value, we follow the procedure above:

1. $u = (n-1)s^2/\sigma_0^2 = (14)\, s^2/\left(0.008\right)^2 = 0.002347/\left(0.008\right)^2 = 36.67$

2. The p-value is
$$p - value = 2P(U \geq u) = 2P(U \geq 36.67) = 0.0017$$
where $U \sim \chi^2(14)$.

This indicates very strong evidence against H_0 and, since the observed value of $s = \sqrt{0.002347/14} = 0.0129$ is greater than 0.008, the data suggests that σ is bigger than 0.008.

6.3 General Gaussian Response Models

We now consider general models of the form (6.3): $Y_i \sim G(\mu_i, \sigma)$ with $\mu(\mathbf{x}_i) = \sum\limits_{j=1}^{k} \beta_j x_{ij}$ for independent units $i = 1, 2, \ldots, n$. For convenience we define the $n \times k$ (where $n > k$) matrix X of covariate values:

$$X = (x_{ij}) \quad \text{for } i = 1, ..., n \text{ and } j = 1, 2, ...k. \tag{6.3.1}$$

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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ and σ. (Note that to facilitate the matrix proof below we have taken $\beta_0 = 0$

in (6.3). The estimator of β_0 can be obtained from the result below by letting $x_{i1} = 1$ for $i = 1, \ldots, n$ and $\beta_0 = \beta_1$.)

Maximum Likelihood Estimators of $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ and of σ.

Theorem 5 *The maximum likelihood estimators $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ and σ are, with $\mathbf{Y}_{n\times 1} = (Y_1, ..., Y_n)^T$,*

$$\tilde{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} \tag{6.3.2}$$

$$\text{and } \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_i)^2 \text{ where } \tilde{\mu}_i = \sum_{j=1}^{k} \tilde{\beta}_j x_{ij} \tag{6.3.3}$$

Proof. The likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2\right] \quad \text{where } \mu_i = \sum_{j=1}^{k} \beta_j x_{ij}$$

and the log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma) &= \log L(\boldsymbol{\beta}, \sigma) \\ &= -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 . \end{aligned}$$

Note that if we take the derivative with respect to a particular β_j and set this derivative equal to 0, we obtain,

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ell(\boldsymbol{\beta}, \sigma) &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \mu_i}{\partial \beta_j} = 0 \\ \text{or } \quad \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} &= 0 \text{ for each } \; j = 1, 2, ..., k \end{aligned}$$

In terms of the matrix X and the vector $\mathbf{y}$ =$(y_1, ..., y_n)^T$we can rewrite this system of equations more compactly as

$$\begin{aligned} X^T(\mathbf{y} - X\boldsymbol{\beta}) &= \mathbf{0} \text{ or} \\ X^T\mathbf{y} &= X^T X \boldsymbol{\beta} \end{aligned}$$

Assuming that the $k \times k$ matrix $X^T X$ has an inverse we can solve these equations to obtain the maximum likelihood estimate of $\boldsymbol{\beta}$, in matrix notation as

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

with corrresponding 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,

$$\frac{\partial}{\partial \sigma}\ell(\boldsymbol{\beta}, \sigma) = \frac{\partial}{\partial \sigma}\left[-n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\right] = 0$$

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 $\tilde{\beta}$ 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 an estimator of the variance σ^2 similar to the maximum likelihood estimator:

$$S_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_i)^2 = \frac{n}{n-k}\tilde{\sigma}^2$$[31]

Theorem 6 *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 $\sigma^2(X^TX)^{-1}$, $j = 1, 2, ..., k$.*

2. The random variable

$$\frac{n\tilde{\sigma}^2}{\sigma^2} = \frac{(n-k)S_e^2}{\sigma^2} \tag{6.3.4}$$

has a chi-squared distribution with $n - k$ degrees of freedom.

3. W is independent of $(\tilde{\beta}_1, ..., \tilde{\beta}_k)$.

[31] 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.

Proof. The estimator $\tilde{\beta}_j$ can be written using (6.3.2) 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 $B = (b_{ji})_{k\times n} = (X^T X)^{-1} X^T$. Note that $BX = (X^T X)^{-1} X^T X$ 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 $\mu_i = \sum\limits_{l=1}^{k} \beta_l x_{il}$ 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 $BB^T = (X^T X)^{-1}$, that $\sum\limits_{i=1}^{n} b_{ji}^2$ is the j'th diagonal element of the matrix $(X^T X)^{-1}$. 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 a **least squares (LS) 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 μ_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 maximum likelihood. Problem 16 describes the method of least squares method.

Remark 7 [32]*From the above theorem we can obtain confidence intervals and test hypotheses for the regression coefficients using the pivotal*

$$\frac{\tilde{\beta}_j - \beta_j}{S_e\sqrt{c_j}} \tag{6.3.5}$$

[32] Recall: if $Z \sim G(0,1)$ and $W \sim \chi^2_{(r)}$ then the random variable $T = Z/\sqrt{W/r}$ has a $t_{(r)}$ distribution. Put $Z = \frac{\tilde{\beta}_j - \beta_j}{\sigma\sqrt{c_j}}$ and $W = \frac{(n-k)S^2}{\sigma^2}$ and $r = n-k$ to obtain this result.

which has a t distribution with $n-k$ degrees of freedom. Here c_j is the j'th diagonal element of the matrix $\left(X^T X\right)^{-1}$.

Confidence intervals for β_j. Exactly as we constructed confidence intervals for the parameter μ for observations from the $G(\mu, \sigma)$ distribution, we can use this result 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

$$P(-a < T < a) = 0.95 \text{ where } T \sim t\,(n-k)\,.$$

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

$$\hat{\beta}_j - as_e\sqrt{c_j} \leq \beta_j \leq \hat{\beta}_j + as_e\sqrt{c_j}$$

where

$$s_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \text{ and } \hat{\mu}_i = \sum_{j=1}^{k}\hat{\beta}_j x_{ij}.$$

Thus a 95% confidence interval for β_j is

$$\left[\hat{\beta}_j - as\sqrt{c_j},\ \hat{\beta}_j + as\sqrt{c_j}\right]$$

which takes the familiar form

estimate ± a × estimated standard deviation of estimator.

We will now consider a number of special cases of these Gaussian response models that are most comonly applied. The first, we have already seen, but it provides a simple example to validate the more general formulae.

Single Gaussian distribution. Here, $Y_i \sim G(\mu, \sigma)$ $i = 1, ..., n$, i.e., $\mu(\mathbf{x}_i) = \mu$ and $\mathbf{x}_i = x_{1i} = 1$, for all $i = 1, 2, ..., n$, $k = 1$ we use the parameter μ instead of $\boldsymbol{\beta} = (\beta_1)$. Notice that in this case $X_{n\times 1} = (1, 1, ..., 1)^T$. This model was discussed in detail in Section 6.2. The pivotal quantity (6.3.5) 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 (6.3.4) that

$$\frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared $(n-1)$ distribution, as determined in Section 6.2.

Comparing Two Gaussian Distributions $G(\mu_1,\sigma)$ and $G(\mu_2,\sigma)$. Suppose we have two independent samples from Gaussian distributions, of sample size n_1, n_2 respectively,

$$\begin{aligned} &Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ are obtained from } G(\mu_1,\sigma) \\ &\text{and } Y_{21}, yY_{22}, ..., Y_{2n_2} \text{ are obtained from } G(\mu_2,\sigma). \end{aligned}$$

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 (6.3) by stacking these two sets of observations in a vector of $n = n_1 + n_2$ observations:

$$(Y_{11}, Y_{12}, ..., Y_{1n_1}, Y_{21}, Y_{22}, ..., Y_{2n_2})^T$$

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

$$L(\mu_1,\mu_2,\sigma) = \prod_{j=1}^{2}\prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_{ji}-\mu_j}{\sigma}\right)^2\right]$$

Maximization of the likelihood function gives the maximum likelihood estimators:

$$\begin{aligned} \tilde{\mu}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} = \bar{Y}_1, \\ \tilde{\mu}_2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{2i} = \bar{Y}_2, \\ \text{and } \tilde{\sigma}^2 &= \frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji}-\tilde{\mu}_j)^2. \end{aligned}$$

Note that the estimator of the variance σ^2 (sometimes referred to as the pooled estimator of variance) adjusted for the degrees of freedom is

$$\begin{aligned} S_p^2 &= \frac{1}{n_1+n_2-2}\sum_{j=1}^{2}(n_j-1)S_j^2 \\ &= \frac{n_1+n_2}{n_1+n_2-2}\tilde{\sigma}^2 \end{aligned}$$

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. You should observe that the overall estimator of variance S_p^2 can be written as a *weighted average* of the estimators S_j^2. In fact

$$S_p^2 = \frac{w_1 S_1^2 + w_2 S_2^2}{w_1 + w_2} \tag{6.3.6}$$

where the weights are $w_j = n_j - 1$. Although you could substitute weights other than $n_j - 1$ in (6.3.6)[33], 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.

Confidence intervals for the difference between two expected values.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 $\overline{Y}_1 - \overline{Y}_2$ and it 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 σ from the pooled data is

$$S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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:

Proposition 8 *The random variable*

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Similarly the random variable

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2$$

has a chi-squared distribution with $n_1 + n_2 - 2$ degrees of freedom.

Confidence intervals or tests for $\mu_1 - \mu_2$ and σ can be obtained by using these pivotal quantities exactly as in Section 6.2 for a single distribution.

Example 6.3.1. 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

[33] you would most likely be tempted to use $w_1 = w_2 = 1/2$.

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 are 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}, $i = 1, \ldots, 12$ for paint A are independent $G(\mu_1, \sigma_1)$ random variables, and independently the measurements Y_{2i}, $i = 1, \ldots, 12$ 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

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} \tag{6.3.7}$$

which in this case has a t distribution with $12 + 12 - 2 = 22$ degrees of freedom. We have assumed[34] that the two population variances are identical, $\sigma_1 = \sigma_2 = \sigma$, with σ 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

$$D = \frac{|\overline{Y}_1 - \overline{Y}_2 - 0|}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{|\overline{Y}_1 - \overline{Y}_2|}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}}$$

From the data above we find

$$\begin{array}{llll} n_1 = 12 & \bar{y}_1 = 10.4 & \sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 = 14.08 & s_1^2 = 1.2800 \\ n_2 = 12 & \bar{y}_2 = 9.0 & \sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2 = 38.64 & s_2^2 = 3.5127. \end{array}$$

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

$$d = \frac{|\bar{y}_1 - \bar{y}_2|}{s_p\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{1.4}{\sqrt{2.3964\left(\frac{1}{6}\right)}} = 2.22$$

[34] if it were easy to do without this assumption, or the sample variances differed by a lot, we would not make it, but without assuming the variances are the same, the problem is a little more complicated. See below.

with

$$p - value = P(|T| \geq 2.22) = 0.038$$

where $T \sim t\,(22)$. There is fairly strong 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 (6.3.7) is obtained using

$$\begin{aligned} 0.95 &= P(-2.074 \leq T \leq 2.074) \\ &= P\left(-2.074 \leq \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} \leq 2.074)\right) \\ &= P\left(-2.074 S_p\sqrt{\frac{2}{12}} \leq \mu_1 - \mu_2 \leq 2.074 S_p\sqrt{\frac{2}{12}}\right). \end{aligned}$$

This gives the 95% confidence interval for $\mu_1 - \mu_2$ as

$$\hat{\mu}_1 - \hat{\mu}_2 \pm 2.074 s_p\sqrt{\frac{2}{12}} \quad \text{or} \quad 0.09 \leq \mu_1 - \mu_2 \leq 2.71.$$

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%).

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 $Y_i \sim G(\mu_1, \sigma_1)$ and $Y_2 \sim G(\mu_2, \sigma_2)$. In this case there is no exact pivotal quantity with which to get a confidence interval for the difference in means $\mu_1 - \mu_2$. However the random variable

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{6.3.8}$$

has approximately a standard Gaussian distribution, especially if n_1, n_2 are large.

To illustrate its use, consider Example 6.3.1, where we had $s_1^2 = 1.2800$ 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 (6.3.8) to obtain a 95% confidence interval for $\mu_1 - \mu_2$. This resulting approximate 95% confidence interval is

$$\bar{y}_1 - \bar{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.3.9}$$

For the given data this equals 1.4 ± 1.24, or $0.16 \leq \mu_1 - \mu_2 \leq 2.64$ 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:

District 1:	$n_1 = 278$	$\bar{y}_1 = 60.2$	$s_1 = 10.16$
District 2:	$n_2 = 345$	$\bar{y}_2 = 58.1$	$s_2 = 9.02$

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. Assuming that in a given district the scores Y have a $G(\mu, \sigma)$ distribution, we can test that μ_1, the mean score in district 1 is the same as μ_2,*themeanscoreindistrict*2. *Alternativelywcangiveaconfidenceir*
μ_2. (Achievement tests are usually designed so that the scores are approximately Gaussian, so this is a sensible procedure.)

If we use (6.3.9) to construct an approximate 95% confidence interval for $\mu_1 - \mu_2$ we obtain

$$60.2 - 58.1 \pm 1.96\sqrt{\frac{(10.16)^2}{278} + \frac{(9.02)^2}{345}} = 2.1 \pm (1.96)(0.779) \quad \text{or} \quad 0.57 \leq \mu_1 - \mu_2 \leq 1.63.$$

Since $\mu_1 - \mu_2 = 0$ is outside the aproximate 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 that $\mu_1 = \mu_2$, suggesting that $\mu_1 > \mu_2$.

We should not rely only on a comparison of their means. It is always a good idea to look carefully at the data and the distributions suggested for the two groups. Figure 6.5 shows a box plot of the two samples; this type of plot was mentioned in Section 1.3. It shows both the median value and other summary statistics of each sample: the upper and lower quartiles (i.e. 25th and 75th percentiles) and the smallest and largest values. Figure 6.5 was obtained using the R function *boxplot*().

Note that the distributions of marks for districts 1 and 2 are actually quite similar. The median (and 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[35] 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

[35] 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?

being “superior” to another.

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(*y_1*,*y_2*,var.equal=T)*, where y_1 and y_2 are the data vectors from 1 and 2.

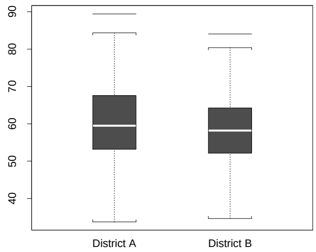


Figure 6.5: **Box Plot of Math Test Scores for Two School Districts.**

6.4 Inference for Paired Data

Although this and the next section are also special cases of the general Gaussian model of Section 6.3, the procedures are sufficiently important that they warrant seperate sections.

Often experimental studies designed to compare means are conducted with **pairs of units,** where the responses within a pair are not independent. The following examples illustrate this.

Example 6.4.1 Heights of Males vs Females
In a study in England, the heights of 1401 (brother, sister) pairs of adults were determined. One objective of the study was to compare the heights of adult males and females; another was to examine the relationship between the heights of male and female siblings.[36]

Let Y_{1i} and Y_{2i} be the heights of the male and female, respectively, in the i’th (brother,

[36] ask yourself "if I had (another?) brother/sister, how tall would they grow to?"

sister) pair ($i = 1, 2, \ldots, 1401$). Assuming that the pairs are sampled randomly from the population, we can use them to estimate

$$\mu_1 = E(Y_{1i}) \text{ and } \mu_2 = E(Y_{2i})$$

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; it requires 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 scatter plot.

Example 6.4.2 Comparing 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 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 ($i = 1, \ldots, 50$) and Y_{2i}'s ($i = 1, \ldots, 50$) 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 Example 6.4.1. This is a continuous bivariate model for which each component has a Normal distributions and the components may be dependent. We will not describe this model here (it is studied in third year courses), except to note one fundamental property: If (Y_{1i}, Y_{2i}) has a Bivariate Normal distribution then the difference between the two is also Normally distributed;

$$Y_{1i} - Y_{2i} \sim N(\mu_1 - \mu_2, \sigma^2) \tag{6.3.10}$$

where $\sigma^2 = Var(Y_{1i}) + Var(Y_{2i}) - 2Cov(Y_{1i}, Y_{1i})$. 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

$$Y_{1i} \sim G(\mu_1 + \alpha_i, \sigma_1^2), \text{ and } Y_{2i} \sim G(\mu_2 + \alpha_i, \sigma_2^2) \text{ are independent,}$$

where the α_i's are unknown constants. Here it is assumed that Y_{1i} and Y_{2i} are independent random variables, and 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 (6.3.10), 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 6.4.2, 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 (6.3.10).

Example 6.4.1 revisited. The data on 1401 (brother, sister) pairs gave differences $Y_i = Y_{1i} - Y_{2i}$, $i = 1, \ldots, 1401$ 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 $\bar{y} \pm 1.96 s/\sqrt{n}$ 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 $4.76 \leq \mu \leq 5.03$ 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 2.4.1 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 as an Experimental Design Choice

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 6.4.1 and 6.4.2, so the pairing in these studies is a good idea.

To illustrate this further, in Example 6.4.1 the height measurement on the 1401 males gave $\bar{y}_1 = 69.720$ and $s_1^2 = 7.3861$ and those 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 (6.3.9) to construct an approximate 95% confidence interval for $\mu_1 - \mu_2$. For the given data we obtain

$$69.720 - 64.825 \pm 1.96\sqrt{\frac{7.3861}{1401} + \frac{6.7832}{1401}} \quad \text{or} \quad 4.70 \leq \mu_1 - \mu_2 \leq 5.09.$$

We note that it is slightly wider than the 95% confidence interval $4.76 \leq \mu \leq 5.03$ obtained using the pairings.

To see why the pairing is helpful in estimating the mean difference $\mu_1 - \mu_2$, suppose that $Y_{1i} \sim G(\mu_1, \sigma_1^2)$ and $Y_{2i} \sim G(\mu_2, \sigma_2^2)$, 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

$$\begin{aligned} Var(\bar{Y}_1 - \bar{Y}_2) &= Var(\bar{Y}_1) + Var(\bar{Y}_2) - 2Cov(\bar{Y}_1, \bar{Y}_2) \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2\sigma_{12}}{n}, \end{aligned}$$

where $\sigma_{12} = Cov(Y_{1i}, Y_{2i})$. If $\sigma_{12} > 0$, then $Var(\bar{Y}_1 - \bar{Y}_2)$ is **smaller** than when $\sigma_{12} = 0$ (i.e. 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 6.4.3. Fibre in Diet and Cholesterol Level[37]

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

[37] from the old Stat 231 notes of MacKay and Oldford

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.

Table 6.4.1 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

$$Y_i = Y_{1i} - Y_{2i} \sim G(\mu_1 - \mu_2, \sigma) \quad \text{for } i = 1, \ldots, 20.$$

The differences y_i are also shown in Table 6.4.1, and from them we calculate the sample mean and standard deviation

$$\bar{y} = -0.020 \;\text{ and }\; s = 0.411$$

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 (6.6) is

$$\bar{y} \pm 2.093 \left(\frac{s}{\sqrt{n}}\right) = -0.020 \pm 2.093 \left(\frac{0.411}{\sqrt{20}}\right) = -0.020 \pm 0.192$$

or

$$-0.212 \leq \mu_1 - \mu_2 \leq 0.172$$

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 (6.7).

Table 6.4.1. 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

Final Remarks: When you see data from a **comparative study** (i.e. one whose objective is to compare two distributions, often through their means), you have to determine whether it involves paired data or not. Of course, a sample of Y_{1i}'s and Y_{2i}'s cannot be from a paired study unless there are equal numbers of each, but if there are equal numbers the study might be either "paired" or "unpaired". Note also that there is a subtle difference in the study populations in paired and unpaired studies. In the former it is pairs of individual units that form the population where as in the latter there are (conceptually at least) separate individual units for Y_1 and Y_2 measurements.

6.5 Linear Regression Models

Many studies involve covariates $\mathbf{x}$, as described in Sections 6.1 and 6.3. In this section we consider simpler settings where there is a single covariate or x-variable. We start by summarizing results from Sections 6.1 and 6.3. Consider the model with independent Y_i's such that

$$Y_i \sim G(\mu\,(x_i)\,,\sigma) \quad \text{where} \quad \mu\,(x_i) = \alpha + \beta x_i \tag{6.3.11}$$

This is of the form (6.3) with (β_0, β_1) replaced by (α, β).

We can use the general results of Section 6.3 or just maximize the likelihood

$$L(\alpha,\beta,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\,(y_i-\alpha-\beta x_i)^2\right]$$

directly to get the maximum likelihood estimators. The maximum likelihood estimators are:

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}}, \tag{6.3.12}$$

$$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}, \tag{6.3.13}$$

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2 = \frac{1}{n}(S_{yy} - \tilde{\beta}S_{xy})$$

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 alternate expressions for S_{xy}[38] and S_{yy} [39] are easy to obtain. As usual we will use, instead of $\tilde{\sigma}^2$, the version of the variance estimator

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \tilde{\beta}S_{xy})$$

Notice that we can rewrite the expression for $\tilde{\beta}$ as

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}Y_i$$

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

$$\begin{aligned} E(\tilde{\beta}) &= \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}E(Y_i) \\ &= \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}(\alpha + \beta x_i) \\ &= \beta\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}x_i \text{ since } \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}\alpha = 0 \\ &= \beta\frac{S_{xx}}{S_{xx}} = \beta \end{aligned}$$

Similarly

$$\begin{aligned} Var(\tilde{\beta}) &= \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{S_{xx}^2}Var(Y_i) \\ &= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2 \\ &= \frac{S_{xx}}{S_{xx}^2}\sigma^2 \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

[38] since $\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})x_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{x}$ and $\sum_{i=1}^{n}(x_i - \bar{x})\bar{x} = 0$

[39] since $\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(x_i - \bar{x})Y_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{Y}$ and $\sum_{i=1}^{n}(x_i - \bar{x})\bar{Y} = 0$

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 $\sum\limits_{i=1}^{n} x_i'$ close to zero. (Often we take $c = \bar{x}$, which makes $\sum\limits_{i=1}^{n} x_i'$ 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 here we have kept the given definition of x_i, for simplicity.

Confidence Intervals for β. These are important because β represents the increase in the expected value of Y, i.e. in

$$E(Y|x) = \alpha + \beta x$$

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). We have seen that $\tilde{\beta}$ is Gaussian with expected value β and with variance σ^2/S_{xx}, i.e.

$$\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right) \tag{6.3.14}$$

and combining this with the fact that

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2) \tag{6.3.15}$$

and that $\tilde{\beta}$ and S^2 are independent, we can argue as before that the random variable

$$\frac{\tilde{\beta} - \beta}{S_e/\sqrt{S_{xx}}} \tag{6.3.16}$$

has a t distribution with $n-2$ degrees of freedom. This can be used as a pivotal quantity to get confidence intervals for β, or to test hypotheses about β.

Note also that (6.3.15) can be used to get confidence intervals or tests for σ, but these are usually of less interest than inference about β or the other quantities below.

Confidence Intervals for $\mu(x) = \alpha + \beta x$.

We are often interested in estimating the quantity $\mu(x) = \alpha + \beta x$ for a specified value of 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 $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$. Thus $\mu(x)$ is a linear function of Gaussian random variables (because $\bar{Y}$ and $\tilde{\beta}$ are Gaussian random variables) and so must also have a Gaussian distribution. Its mean and variance are

$$\begin{aligned}
E[\tilde{\mu}(x)] &= E(\bar{Y}) + (x - \bar{x})E(\tilde{\beta}) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) + (x - \bar{x})\beta \\
&= \frac{1}{n}\sum_{i=1}^{n} (\alpha + \beta x_i) + (x - \bar{x})\beta \\
&= \alpha + \beta\bar{x} + (x - \bar{x})\beta \\
&= \alpha + \beta x \\
&= \mu(x)
\end{aligned}$$

and

$$\begin{aligned}
Var\left[\tilde{\mu}(x)\right] &= \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x-\bar{x})(x_i-\bar{x})}{S_{xx}}\right]^2 Var(Y_i) \\
&= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} + \frac{2}{n}\frac{(x-\bar{x})(x_i-\bar{x})}{S_{xx}} + \frac{(x-\bar{x})^2(x_i-\bar{x})^2}{S_{xx}^2}\right] \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right].
\end{aligned}$$

Thus

$$\tilde{\mu}(x) \sim G\left(\mu(x),\ \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

and it then follows that

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \tag{6.3.17}$$

is a pivotal quantity with a t distribution on $n-2$ degrees of freedom. This can be used as a pivotal quantity to get confidence intervals for $\mu(x)$. The corresponding 95% confidence interval is

$$\hat{\mu}(x) \pm a s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \tag{6.3.18}$$

where $P(-a < T < a) = 0.95$ and s_e^2 is our estimate of σ^2 given by

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}S_{xy})$$

where 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 (6.3.18) with $x = 0$ which gives

$$\hat{\alpha} \pm a s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}} \tag{6.3.19}$$

In fact one can see from (6.3.19) 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 6.5.1 Price vs 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 6.1.1 suggested that a linear regression model of the form $E(Y|x) = \alpha + \beta x$ would be reasonable. For the given data $n = 30$, $\bar{x} = 0.954$, $\bar{y} = 549.0$ and $S_{xx} = 22.945$, $S_{xy} = -3316.68$, $S_{yy} = 489,462.62$ so we find

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-3316.68}{22.945} = -144.5,$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 549.0 - (-144.5)(0.954) = 686.9,$$
$$s_e^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta} S_{xy}) = \frac{1}{28}\left[489462.62 - (-144.5)(-3316.68)]\right) = 364.37,$$

and $s_e = 19.09$.

Note that $\hat{\beta}$ is negative: the larger size buildings tend to sell for less per square meter. (The estimate $\hat{\beta} = -144.5$ indicates a drop in average price of \$144.50 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 scatter plot of points (x_i, y_i) as in Figure (6.6), we see the fitted line passes close to the points.

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

$$\hat{\mu}(4.47) = \hat{\alpha} + \hat{\beta}(4.47) = \$40.94$$

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 (6.3.18) and the fact that $P(-2.048 \leq T \leq 2.048) =$

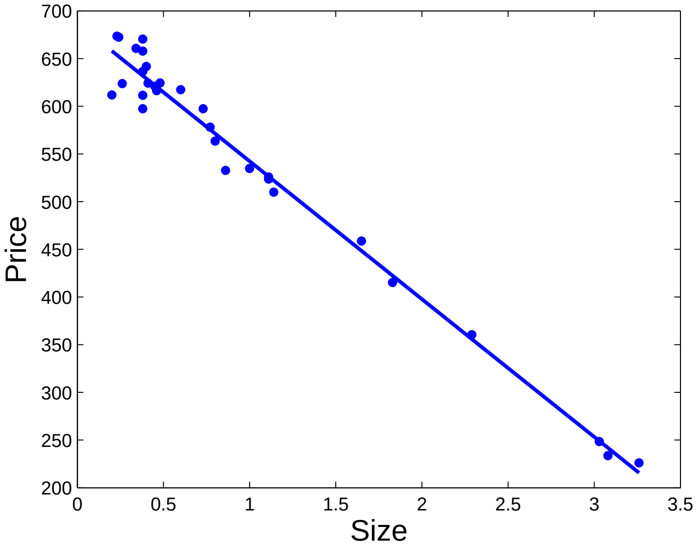


Figure 6.6:

0.95 for $T \sim t\,(28)$ we get a 95% confidence interval for $\mu(4.47)$ as

$$\hat{\mu}(4.47) \pm 2.048 s_e \sqrt{\frac{1}{30} + \frac{(4.47 - \bar{x})^2}{S_{xx}}}$$

or \$40.94 ± \$26.54, or $\$14.40 \leq \mu(4.47) \leq \$67.50$. Thus the assessed value of \$75 is outside this range.

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-neglible 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 to do this.

Prediction Intervals for Y

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 get a pivotal quantity that can be used to give a prediction interval (or interval "estimate") for Y, as follows.

Note that $Y \sim G(\mu(x), \sigma)$ from (6.3.11) or alternatively we can write

$$Y = \mu(x) + R, \quad \text{where } R \sim G(0, \sigma)$$

is independent of $Y_1, \ldots, Y_n$. 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

$$Y - \tilde{\mu}(x) = Y - \mu(x) + \mu(x) - \tilde{\mu}(x) = R + [\mu(x) - \tilde{\mu}(x)].$$

Since, R is independent of $\tilde{\mu}(x)$ (it is not connected to the existing sample), this is the sum of independent Normally distributed random variables and is consequently Nomally distributed. Moreover,

$$E\left(Y - \tilde{\mu}(x)\right) = E\left(R + (\mu(x) - \tilde{\mu}(x))\right) = E(R) + E\mu(x) - E(\tilde{\mu}(x)) = 0.$$

For the variance, since Y and $\tilde{\mu}(x)$ are independent,

$$\begin{aligned} Var\left[Y - \tilde{\mu}(x)\right] &= Var(Y) + Var\left[\tilde{\mu}(x)\right] \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]. \end{aligned}$$

Thus

$$Y - \tilde{\mu}(x) \sim G\left(0, \sigma\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}\right).$$

To generate a prediction interval for Y, we will use the corresponding pivotal quantity

$$\frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \tag{6.3.20}$$

which has a t distribution with $n-2$ degrees of freedom. To get interval estimates for Y, say having confidence coefficient 0.95, we choose a such that since

$$\begin{aligned} 0.95 &= P(-a \leq T \leq a) \quad \text{where } T \sim t\,(n-2) \\ &= P\left(\tilde{\mu}(x) - aS_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \leq Y \leq \tilde{\mu}(x) + aS_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right) \end{aligned}$$

The interval

$$\hat{\mu}(x) - as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \leq Y \leq \hat{\mu}(x) + as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \tag{6.3.21}$$

is usually called $100p\%$ **prediction interval** instead of a confidence interval, since Y is not a parameter but a "future" observation.

Example 6.5.1 Revisited
Let us obtain a 95% prediction interval for Y when $x = 4.47$. Using (6.3.21) and the fact that $P(-2.048 \leq T \leq 2.048) = 0.95$ when $T \sim t\,(28)$ we obtain

$$\tilde{\mu}(4.47) \pm 2.048s\sqrt{1 + \frac{1}{30} + \frac{(4.47 - \bar{x})^2}{22.945}}$$

or $-6.30 \leq Y \leq 88.20$ (dollars per square meter). 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 $0 \leq Y \leq \$88.20$.

Now we find that the assessed value of \$75 is inside this interval! On this basis its hard 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; look again at Figure (6.6). 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. Note also that is a slight suggestion in Figure (6.6) 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 (6.3.18) and (6.3.21) that the confidence intervals for $\mu(x)$ and 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 6.5.2 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 6.7 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. We find (you should check these for yourself)

$$\hat{\alpha} = 1.667, \quad \hat{\beta} = 2.838, \quad s = 0.0515, \quad S_{xx} = 0.2244$$

The fitted regression line $y = \hat{\alpha} + \hat{\beta}x_1$ is shown on the scatter plot in Figure 6.7; the model appears to fit well.

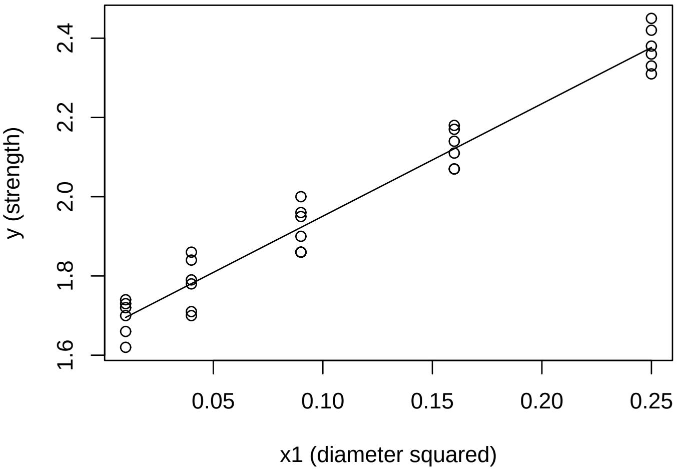


Figure 6.7: **Scatter Plot of Bolt Diameter Squared vs. Strength**

More as a numerical illustration, let us get a confidence interval for β, which represents the increase in average strength $\mu(x_1)$ from increasing $x_1 = x^2$ by one unit. Using the pivotal quantity (6.3.16) and the fact that $P(-2.048 \leq T \leq 2.048) = 0.95$ for $T \sim t\,(28)$, we obtain the 95% confidence interval for β as

$$\hat{\beta} \pm 2.048\frac{s}{\sqrt{S_{xx}}}, \quad \text{or} \quad 2.838 \pm 0.223.$$

A 95% confidence interval for the value of β is therefore[2.605, 3.051].

Exercise: This model could be used to predict the breaking strength of a new bolt of given diameter x. Find a 95% prediction interval for a new bolt of diameter $x = 0.35$.

Summary of distributions for Simple Linear Regression model

Random variable	Distribution	Mean (or parameter)	Standard Deviation
$\tilde{\beta} = \frac{S_{xy}}{S_{xx}}$	Gaussian	β	$\sigma \left[\frac{1}{S_{xx}} \right]^{1/2}$
$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{s_{xx}}}$	t	df $= n-2$	-
$\tilde{\alpha} = y - \tilde{\beta}\ \bar{x}$	Gaussian	α	$\sigma \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]^{1/2}$
$\tilde{\mu}(x)$	Gaussian	$\mu(x) = \alpha + \beta x$	$\sigma \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right]^{1/2}$
$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$	t	df $= n-2$	-
$Y - \tilde{\mu}(x)$	Gaussian	0	$\sigma \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right]^{1/2}$
$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$	t	df $= n-2$	-
$\frac{(n-2)S_e^2}{\sigma^2}$	Chi squared	df $= n-2$	-

6.6 Model Checking

There are two main components in Gaussian linear response models:

(i) the assumption that Y_i (given any covariates x_i) is Gaussian with constant standard deviation σ.

(ii) the assumption that $E(Y_i) = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.

Models should always be checked, and in this case there are several ways to do this. Some of these are based on what we term "residuals" of the fitted model: the **residuals** are the values

$$\hat{r}_i = y_i - \hat{\mu}_i \text{ for } i = 1, \ldots, n$$

For example, if $Y_i \sim G(\alpha + \beta x_i; \sigma)$ then the residuals are $\hat{r}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. The *R* function *lm* produces these values as part of its output.

If $Y_i \sim G(\mu_i, \sigma)$ then $R_i = Y_i - \mu_i \sim G(0, \sigma)$. 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 distribution $G(0, \sigma)$. They do have some features that make it easy to identify if your parameter estimates are incorrect. 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 alway zero**. Indeed the equation defining the maximum likelihood estimate $\hat{\beta}$ imposes another linear constraint that the residuals must always satisfy. You can think of this as forcing the last two residuals to be functions of the first $n-2$ residuals which explains why in simple linear regression S_e^2 which obtains from the residual sum of squares $\sum\limits_{i=1}^{n} \hat{r}_i^2$ has $n-2$ degrees of freedom.

Plots of residuals can be used as a model check. For example, we can

(1) Plot points $(x_i, \hat{r}_i)$, $i = 1, \ldots, n$. If the model is satisfactory these should lie more or less horizontally within a band around the line $\hat{r}_i = 0$.

(2) Plot points $(\hat{\mu}_i, \hat{r}_i)$, $i = 1, \ldots, n$. If the model is satisfactory we should get the same type of pattern as for (1).

Departures from the "expected" pattern in (1) and (2) may suggest problems with the model. For example, if in (2) we see that the variability in the $\hat{r}_i$'s is bigger for larger values of $\hat{\mu}_i$, this suggests that $Var(Y_i) = Var(R_i)$ is not constant, but may be larger when $\mu(x)$ is larger.

Figure 6.8 shows a couple of such patterns; the left hand plot suggests non-constant variance whereas the right hand plot suggests that the function $\mu_i = g(x_i)$ is not correctly specified. Reading these plots is something of an art and we should try not to read too much into plots based on a small number of points.

In problems with only one x covariate, a plot of $\hat{\mu}(x)$ superimposed on the scatterplot of the data (as in Figure 6.7) shows pretty clearly how well the model fits. The residual plots described are however, very useful when there are two or more covariates in the model.

When there are no covariates in the model, as in Section 6.2, plots (1) and (2) are undefined. In this case the only assumption is that $Y_i \sim G(\mu, \sigma)$. We can still define residuals, either as

$$\hat{r}_i^* = y_i - \hat{\mu} \quad \text{or} \quad \hat{r}_i^* = \frac{y_i - \hat{\mu}}{\hat{\sigma}},$$

where $\hat{\mu} = \bar{y}$ and $\hat{\sigma}$ (we could alternatively use s) is the maximum likelihood estimate of σ. One way to check the model is to treat the $\hat{r}_i^*$'s (which are called **standardized residuals**) as a random sample of values $(Y - \mu)/\sigma$. Since $Y - \mu)/\sigma \sim G(0, 1)$ under our assumed model, we could plot the empirical cumulative distribution function from $\hat{r}_i^*$, $i = 1, \ldots, n$ and superimpose on it the $G(0, 1)$ cumulative distribution function. The two curves should agree well if the Gaussian model is satisfactory. This plot can also be used when there are covariates, by defining the standardized residuals

$$\hat{r}_i^* = \frac{\hat{r}_i}{\hat{\sigma}} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}} \quad \text{for } i = 1, \ldots, n.$$

We can also use the $\hat{r}_i^*$'s in place of the $\hat{r}_i$'s in plots (1) and (2) above; in fact that is what we did in Figure 6.8. When the $\hat{r}_i^*$'s are used the patterns in the plot are unchanged but

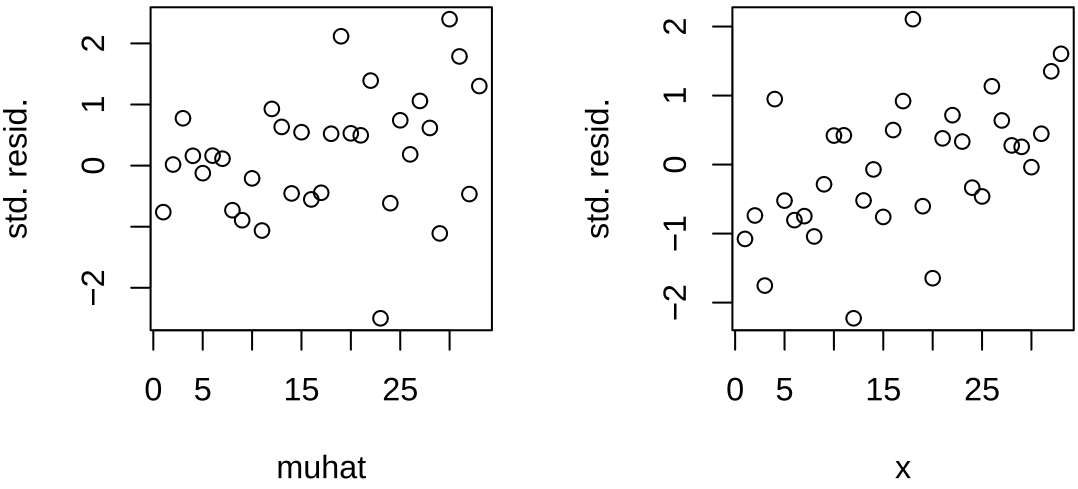


Figure 6.8: **Examples of Patterns in Residual Plots**

the $\hat{r}_i^*$ values tend to lie in the range $(-3, 3)$. (Why is this?)

Example 6.6.1 Residuals for the Steel Bolts Example. Let us define residuals

$$\hat{r}_i = y_i - \hat{\alpha} - \hat{\beta} x_i \quad \text{for } i = 1, \ldots, 30$$

for the model fitted in Example 6.5.3. Figure 6.9 shows a plot of the points $(x_{1i}, \hat{r}_i)$; no deviation from the expected pattern is observed. This is of course also evident from Figure 6.7.

A further check on the Gaussian distribution is shown in Figure 6.10. Here we have plotted the empirical distribution function based on the standardized residuals

$$\hat{r}_i^* = \frac{y_i - \hat{\alpha} - \hat{\beta} x_{1i}}{\hat{\sigma}} \quad \text{for } i = 1, \ldots, 30.$$

On the same graph is the $G(0, 1)$ cumulative distribution function. There is reasonably good agreement between the two curves.

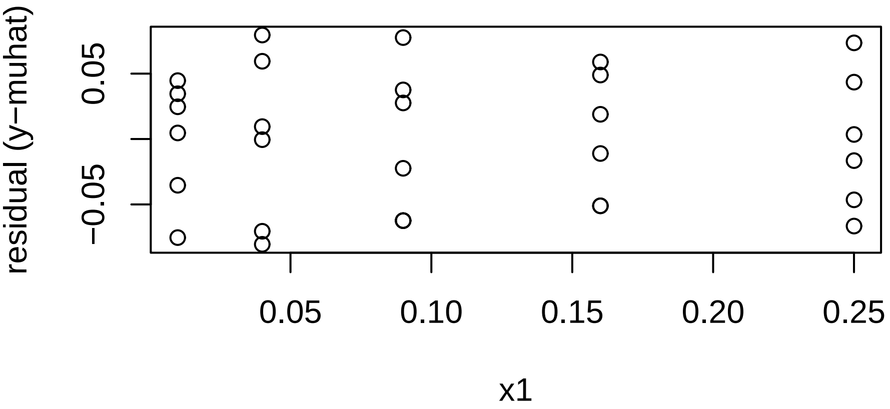


Figure 6.9: **Residual Plot for Bolt Strength Mode**l

6.7 Problems

1. Student's t Distribution

Suppose that Z and U are independent variates with

$$Z \sim N(0,1) \text{ and } U \sim \chi^2(k).$$

Consider the random variable

$$X \equiv \frac{Z}{\sqrt{U/k}}.$$

Its distribution is called the t (Student's) distribution with k degrees of freedom, and we write $X \sim t(k)$. It can be shown by change of variables that X has probability density function

$$f(x;k) = c_k \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} \quad \text{for } -\infty < x < \infty \text{ and } k = 1, 2, \ldots$$

where c_k is a normalizing constant such that the total area under the probability density function is one:

$$c_k = \Gamma\left(\frac{k+1}{2}\right) / \sqrt{k\pi}\Gamma\left(\frac{k}{2}\right).$$

The probability density function is symmetric about the origin, and is similar in shape to the probability density function of $N(0,1)$ random variable but has more probability in the tails. It can be shown that $f(x;k)$ tends to the $N(0,1)$ probability density function as $k \to \infty$.

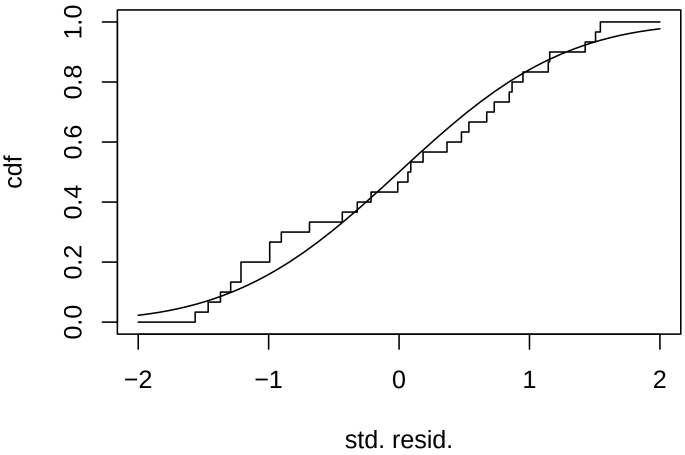


Figure 6.10: **Empirical Distribution Function of Standard Residuals and $G(0,1)$ c.d.f.**

(a) Plot the probability density function for $k = 1$ and $k = 5$.

(b) Find values a, b such that

$$P(-a \leq X \leq a) = 0.98 \text{ and } P(X \geq b) = 0.95 \text{ where } X \sim t(15)$$

(c) Show that $f(x;k)$ is unimodal for all x.

(d) Show that as $k \to \infty$, $f(x;k) \to \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2,\right)$ probability density function of the $G(0,1)$ distribution.
(Note: To do this you will need to use the fact that $c_k \to 1/\sqrt{2\pi}$ as $k \to \infty$; this is from a property of gamma functions.)

2. Suppose that $Y_1, \ldots, Y_n$ are independent $G(\mu, \sigma)$ observations.

(a) Show that the likelihood ratio statistic for testing a value of μ is given by (assume σ is unknown)
$\Lambda(\mu) = n \log\left(1 + \frac{T^2}{n-1}\right)$
where $T = \sqrt{n}(\overline{Y} - \mu)/S$ and S is the sample standard deviation.

(b) Show that the likelihood ratio statistic for testing a value of σ is a function of

$$W = \frac{(n-1)S^2}{\sigma^2}.$$

3. The following data are instrumental measurements of level of dioxin (in parts per billion) in 20 samples of a “standard” water solution known to contain 45 ppb dioxin.

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

   (a) Assuming that the measurements are independent and $G(\mu, \sigma)$, obtain a 95% confidence interval for μ and test the hypothesis that $\mu = 45$.

   (b) Obtain a 95% confidence interval for σ. Of what interest is this scientifically?

4. A new method gave the following ten measurements of the specific gravity of mercury:

13.696	13.699	13.683	13.692	13.705
13.695	13.697	13.688	13.690	13.707

   Assume these to be independent observations from $G(\mu, \sigma)$.

   (a) An old method produced measurements with standard deviation $\sigma = 0.02$. Test the hypothesis that the new method has the same standard deviation as the old.

   (b) A physical chemistry handbook lists the specific gravity of mercury as 13.75. Are the data consistent with this value?

   (c) Obtain 95% confidence intervals for μ and σ.

5. Sixteen packages are randomly selected from the production of a detergent packaging machine. Their weights (in grams) are as follows:

287	293	295	295	297	298	299	300
300	302	302	303	306	307	308	311

   (a) Assuming that the weights are independent $G(\mu, \sigma)$ random variables, obtain 95% confidence intervals for μ and σ.

   (b) Let $\bar{Y} = \frac{1}{n} \sum\limits_{i=1}^{n} Y_i$ and $S^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ be the mean and variance in a sample of size n, and let Y represent the weight of a future, independent, randomly selected package. Show that $Y - \bar{Y} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$ and thus

$$Z = \frac{Y - \bar{Y}}{S\sqrt{1 + \frac{1}{n}}} \sim t\left(n - 1\right).$$

   For the data above, use this as a pivotal to obtain a 95% prediction interval for Y.

6. A manufacturer wishes to determine the mean breaking strength (force) μ of a type of string to "within a pound", which we interpret as requiring that the 95% confidence interval for a μ should have length at most 2 pounds. If breaking strength Y of strings tested are $G(\mu, \sigma)$ and if 10 preliminary tests gave $\sum_{i=1}^{10}(y_i - \bar{y})^2 = 80$, how many additional measurements would you advise the manufacturer to take?

7. 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), $i = 1, \ldots, 30$, is as follows:

Math students:	$n = 30$	$\bar{x} = 120$	$\sum_{i=1}^{30}(x_i - \bar{x})^2 = 3050$
Engineering students:	$n = 30$	$\bar{y} = 114$	$\sum_{i=1}^{30}(y_i - \bar{y})^2 = 2937$

Obtain a 95% confidence interval for the difference in mean scores for first year Math and Engineering students, and test the hypothesis that the difference is zero.

8. 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	*Type II*
3.03	3.19
5.53	4.26
5.60	4.47
9.30	4.53
9.92	4.67
12.51	4.69
12.95	12.78
15.21	6.79
16.04	9.37
16.84	12.75

(a) Assuming that Y, the number of million cycles to failure, has a normal distribution with the same variance for each type of bearing, obtain a 90% confidence interval for the difference in the means μ_1 and μ_2 of the two distributions.

(b) Test the hypothesis that $\mu_1 = \mu_2$.

(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.

9. 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 $\bar{y} = 14.564$ and $s^2 = 0.0914$. Similar tests on four additional girders with repaired welds gave $\bar{y} = 14.291$ and $s^2 = 0.0422$. Log failure times are assumed to be independent with a $G(\mu, \sigma)$ distribution.

(a) Test the hypothesis that the variance of Y is the same for repaired welds as for the normal welds.

(b) Assuming equal variances, obtain a 90% confidence interval for the difference in mean log failure time.

(c) Note that $\mu_1 - \mu_2$ in part (b) is also the difference in median log failure times. Obtain a 90% confidence interval for the ratio

$$\frac{\text{median lifetime (cycles) for repaired welds}}{\text{median lifetime (cycles) for normal welds}}$$

10. Let $Y_1, \ldots, Y_n$ be a random sample from $G(\mu_1, \sigma_1)$ and $X_1, \ldots, X_n$ be a random sample from $G(\mu_2, \sigma_2)$. Obtain the likelihood ratio statistic for testing the hypothesis $\sigma_1 = \sigma_2$ 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.

11. 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 kg. weight yielded $\bar{y} = 10.004$ and $s = 0.013$ as the sample mean and standard deviation. Ten weighings of a 40 kg. 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 kg. item. How could you study the question of interest further?

12. 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 average gas mileages were as follows:

Car	1	2	3	4	5	6	7	8
Synthetic oil	21.2	21.4	15.9	37.0	12.1	21.1	24.5	35.7
Conventional oil	18.0	20.6	14.2	37.8	10.6	18.5	25.9	34.7

(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?

13. Consider the data in Problem 8 of Chapter 1 on the lengths of male and female coyotes.

(a) Fit separate Gaussian models for the lengths of males and females. Estimate the difference in mean lengths for the two sexes.

(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.

14. **Comparing sorting algorithms**. Suppose you want to compare two algorithms A and B that will sort a set of number 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.

15. **Sorting algorithms continued**. Two sort algorithms as in the preceding question 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

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

(a) Plot the data so as to illustrate its main features.

(b) Estimate (give a confidence interval) for the difference in the average time to sort with algorithms A and B, assuming a Gaussian model applies.

(c) Suppose you are asked to estimate the probability that A will sort a randomly selected list fast than B. Give a point estimate of this probability.

(d) Another way to estimate the probability p in part (b) is just to notice that of the 20 sets of numbers in the study, A sorted faster on 15. Indicate how you could also get a confidence interval for p using this approach. (It is also possible to get a confidence interval using the Gaussian model.)

16. **Least squares estimation**. Suppose you have a model where the mean of the response variable Y_i given the covariates $\mathbf{x}_i$ has the form

$$\mu_i = E(Y_i|\mathbf{x}_i) = g(\mathbf{x}_i; \boldsymbol{\beta}) \tag{6.3.22}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. Then the **least squares (LS) estimate** of $\boldsymbol{\beta}$ based on data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ is the value that minimizes the objective function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - g(\mathbf{x}_i; \boldsymbol{\beta})]^2$$

Show that the LS 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 (6.3.22).

17. 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 on the times are shown below for the 24 students.

"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		$\bar{y} = 1.370$, $s = 0.235$				
"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		$\bar{y} = 1.599$, $s = 0.180$				

Analyze the data with the objective of seeing when there is any evidence that the dose of alcohol increases reaction time. Justify any models that you use.

18. 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 10 units.

x :	4.01	8.12	12.53	15.90	20.24	24.81	30.92	37.26	38.94	40.15
y :	3.70	7.80	12.40	16.00	19.90	24.90	30.80	37.20	38.40	39.40

(a) Fit a Gaussian linear regression model for Y given x to the data and obtain 95% confidence intervals for the slope β and standard deviation σ. Use a plot to check the adequacy of the model.

(b) Describe briefly how you would characterize the cheap measurement process's accuracy to a lay person.

(c) Assuming that the units being 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 with an argument based on the data.

19. The following data, collected by Dr. Joseph Hooker in the Himalaya mountains, relates atmospheric pressure to the boiling point of water. Theory suggests that a

graph of log pressure versus boiling point should give a straight line.

Temp (°F)	Pres (in. Hg)	Temp (°F)	Pres (in. 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		

(a) Prepare a scatterplot of $y = \log(\text{Pressure})$ versus $x = $ Temperature. Do the same for $y = $ Pressure versus x. Which is better described by a linear model? Does this confirm the theory's model?

(b) Fit a normal linear regression model for $y = \log(\text{Pressure})$ versus x. Are there any obvious difficulties with the model?

(c) Obtain a 95% confidence interval for the atmospheric pressure if the boiling point of water is $195°F$.

(a) For the steel bolt experiment in Examples 6.1.3 and 6.5.2, use a Gaussian model to

(i) estimate the average breaking strength of bolts of diameter 0.35

(ii) estimate (predict) the breaking strength of a single bolt of diameter 0.35

Give interval estimates in each case.

(b) 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.

20. **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.)

21. Sometimes we want one-sided confidence intervals of the form $L(y_1, \ldots, y_n) \leq \mu$ or $U(y_1, \ldots, y_n) \geq \mu$ which are obtained by taking $a_1 = -\infty$ and $a_2 = \infty$, respectively, in (6.5). For "two-sided" intervals based on the normal or the t distribution, we usually pick $a_1 = -a_2$ so that the interval is symmetrical about $\bar{y}$. Show that for the t distribution, the choice $a_1 = -a_2$ provides the *shortest* $100p\%$ confidence interval.

TESTS AND INFERENCE PROBLEMS BASED ON MULTINOMIAL MODELS

7.1 Introduction

Many important hypothesis testing problems can be addressed using multinomial models. An example was given in Chapter 5, whose general ideas we will use here. To start, recall the setting in Chapter 5, Section 2, where data were assumed to arise from a multinomial distribution with probability function

$$f(y_1, ..., y_m; \theta_1, ..., \theta_m) = \frac{n!}{y_1! \cdots y_m!} \theta_1^{y_1} \cdots \theta_m^{y_m} \tag{6.3.2}$$

where $0 \le y_j \le n$ and $\sum\limits_{j=1}^{m} y_j = n$. The multinomial probabilities θ_j satisfy $0 \le \theta_j \le 1$ and $\sum\limits_{j=1}^{m} \theta_j = 1$, and we define $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$. Suppose now that we wish to test the hypothesis that the probabilities are related in some way, for example that they are all functions of a lower dimensional parameter $\boldsymbol{\alpha}$

$$H_0 : \theta_j = \theta_j(\boldsymbol{\alpha}) \quad \text{for } j = 1, ..., m \tag{6.3.3}$$

where $\dim(\boldsymbol{\alpha}) = p < m - 1$.

The likelihood function based on (6.3.2) is proportional to

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} \theta_j^{y_j}. \tag{6.3.4}$$

Let Ω be the parameter space for $\boldsymbol{\theta}$. It was shown earlier that $L(\boldsymbol{\theta})$ is maximized over Ω (of dimension $m-1$) by the vector $\hat{\boldsymbol{\theta}}$ with $\hat{\theta}_j = y_j/n$, $j = 1, ..., m$. A likelihood ratio test of the hypothesis (6.3.3) is based on the likelihood ratio statistic

$$\Lambda = 2\ell(\tilde{\boldsymbol{\theta}}) - 2\ell(\tilde{\boldsymbol{\theta}}_0) = -2 \log \left\{ \frac{L(\tilde{\boldsymbol{\theta}}_0)}{L(\tilde{\boldsymbol{\theta}})} \right\}, \tag{6.3.5}$$

where $\tilde{\boldsymbol{\theta}}_0$ maximizes $L(\boldsymbol{\theta})$ under the hypothesis (6.3.3), which restricts $\boldsymbol{\theta}$ to lie in a space $\Omega_0 \subset \Omega$ of dimension p. (note that Ω_0 is the space of all $(\theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), ..., \theta_m(\boldsymbol{\alpha}))$ as $\boldsymbol{\alpha}$ varies over its possible values.) If H_0 is true (that is, if $\boldsymbol{\theta}$ really lies in Ω_0) and n is large the distribution of Λ is approximately $\chi^2\,(m-1-p)$. This enables us to compute p-values from observed data by using the approximation

$$P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda) \quad \text{where } W \sim \chi^2\,(m-1-p) \tag{6.3.6}$$

and

$$\lambda = 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\hat{\boldsymbol{\theta}}_0)$$

is the observed value of Λ. This approximation is very accurate when n is large and none of the θ_j's is too small; when the observed expected frequencies under H_0 all exceed five it is accurate enough for testing purposes.

The test statistic (6.3.5) can be written in a simple form. Let $\tilde{\boldsymbol{\theta}}_0 = (\theta_1(\tilde{\alpha}), ..., \theta_m(\tilde{\alpha}))$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$ under the hypothesis (6.3.3). Then, by (6.3.5), we get

$$\begin{aligned}\Lambda &= 2\ell(\tilde{\boldsymbol{\theta}}) - 2\ell(\tilde{\boldsymbol{\theta}}_0)\\ &= 2\sum_{j=1}^{m} Y_j \log\left[\tilde{\theta}_j/\theta_j(\tilde{\alpha})\right].\end{aligned}$$

Noting that $\tilde{\theta}_j = Y_j/n$ and defining "expected frequencies" under H_0 as

$$E_j = n\theta_j(\tilde{\alpha}) \quad \text{for } j = 1, ..., m$$

we can rewrite Λ as

$$\Lambda = 2\sum_{j=1}^{m} Y_j \log(Y_j/E_j). \tag{6.3.7}$$

An alternative test statistic that was developed historically before Λ is the "Pearson" statistic

$$D = \sum_{j=1}^{m} \frac{(Y_j - E_j)^2}{E_j}. \tag{6.3.8}$$

This has similar properties to Λ; for example, their observed values both equal zero when $y_j = e_j = n\theta_j(\hat{\alpha})$ for all $j = 1, ..., m$ and are larger when y_j's and e_j's differ greatly. It turns out that, like Λ, the statistic D also has a limiting $\chi^2\,(m-1-p)$ 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 relative frequencies f_j/n with the estimates $\hat{p}_j$ from the distributional model. This is equivalent to comparing the observed frequencies f_j and the expected frequencies $e_j = n\hat{p}_j$. In Section 2.4 this comparison was informal, with only a rough guideline for how closely the f_j's and e_j's should agree.

It is possible to test the correctness of a parametric model by using an implied multinomial model. We illustrate this through two examples.

Example 7.2.1. 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

$$\theta_1 = \alpha^2 \quad \theta_2 = 2\alpha(1-\alpha) \quad \theta_3 = (1-\alpha)^2. \tag{7.2.1}$$

Data collected on 100 persons gave $y_1 = 17$, $y_2 = 46$, $y_3 = 37$, and we can use this to test the hypothesis H_0 that (7.2.1) is correct. (Note that $(Y_1, Y_2, Y_3) \sim \text{Multinomial}(n; \theta_1, \theta_2, \theta_3)$ with $n = 100$.) The likelihood ratio test statistic is given by (6.3.7), but we have to find $\tilde{\alpha}$ and then the E_j's. The likelihood function under (7.2.1) 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 (7.2.1) 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, y_2, y_3. The observed value of the likelihood ratio statistic (6.3.7) is

$$2\sum_{j=1}^{3} y_j \log(y_j/e_j) = 2\left[17\log(17/16) + 46\log(46/48) + 37\log(37/36)\right] = 0.17$$

and the p-value is

$$p - value = P(\Lambda \geq 0.17; H_0) \approx P(W \geq 0.17) = 0.68 \quad \text{where } W \sim \chi^2(1)$$

so there is no evidence against the model (7.2.1).

The observed values of the Pearson statistic (6.3.8) and the likelihood ratio statistic Λ are usually close when n is large. In this case we find that the observed value of (6.3.8) for these data is also 0.17.

Example 7.2.2. Continuous distributions can also be tested by grouping the data into intervals and then using the multinomial model. Example 2.4.1 previously did this in an informal way for an Exponential distribution. For example, suppose that T is thought to have an Exponential distribution with probability density function

$$f(t;\alpha) = \frac{1}{\alpha} e^{-t/\alpha} \quad \text{for } t > 0. \tag{7.2.2}$$

Suppose a random sample $t_1, ..., t_{100}$ is collected and the objective is to test the hypothesis H_0 that (7.2.2) is correct. To do this we partition the range of T into intervals $j = 1, ..., m$, and count the number of observations y_j that fall into each interval. Under (7.2.2), the probability that an observation lies in the j'th interval $I_j = (a_j, b_j)$ is

$$p_j(\alpha) = \int_{a_j}^{b_j} f(t;\alpha)dt \quad \text{for } j = 1, ..., m \tag{7.2.3}$$

and if y_j is the number of observations (t's) that lie in I_j, then $Y_1, ..., Y_m$ follow a Multinomial $(n; p_1(\alpha), \ldots, p_m(\alpha))$ distribution with $n = 100$. Thus we can test (7.2.2) by testing that (7.2.3) is true.

Consider the following data, which have been divided into $m = 7$ intervals:

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

We have also shown expected frequencies e_j, calculated as follows. The distribution of $(Y_1, ..., Y_7)$ is multinomial with probabilities given by (7.2.3) when the model (7.2.2) is correct. In particular,

$$p_1(\alpha) = \int_0^{100} \frac{1}{\alpha} e^{-t/\alpha} dt = 1 - e^{-100/\alpha},$$

and so on. Expressions for $p_2, ..., p_7$ are $p_2(\alpha) = e^{-100/\alpha} - e^{-200/\alpha}$, $p_3(\alpha) = e^{-200/\alpha} - e^{-300/\alpha}$, $p_4(\alpha) = e^{-300/\alpha} - e^{-400/\alpha}$, $p_5(\alpha) = e^{-400/\alpha} - e^{-600/\alpha}$, $p_6(\alpha) = e^{-600/\alpha} - e^{-800/\alpha}$, $p_7(\alpha) = e^{-800/\alpha}$. The likelihood function from $y_1, ..., y_7$ based on model (7.2.2) is then

$$L_1(\alpha) = \prod_{j=1}^{7} [p_j(\alpha)]^{y_j}.$$

It is possible to maximize $L_1(\alpha)$ mathematically. (Hint: rewrite $L_1(\alpha)$ in terms of the parameter $\beta = e^{-100/\alpha}$ and find $\tilde{\beta}$ first; then $\tilde{\alpha} = -100/\ln\tilde{\beta}$.) This gives $\hat{\alpha} = 310.3$ and the expected frequencies $e_j = 100p_j(\hat{\alpha})$ given in the table are then obtained.

The observed value of the likelihood ratio statistic (6.3.7) is

$$2\sum_{j=1}^{7} y_j \log(y_j/e_j) = 2\left[29\log(19/27.6) + \cdots + 8\log(8/7.6)\right] = 1.91$$

and the 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 (7.2.2). Note that the reason the χ^2 degrees of freedom are 5 is because $m - 1 = 6$ and $p = \dim(\alpha) = 1$.

The goodness of fit test just given has some arbitrary elements, since we could have used different intervals and a different number of intervals. Theory and guidelines as to how best to choose the intervals can be developed, but we won't consider this here. Rough guidelines for our purposes are to chose $4 - 10$ intervals, so that the observed expected frequencies under H_0 are at least 5.

7.3 Two-Way Tables and Testing for Independence of Two Variables

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 sometimes called "contingency" tables in the statistics literature. We will consider the two types of studies in turn.

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 $A_1, A_2, ..., A_a$ and for B an individual can be any of b mutually exclusive types $B_1, B_2, ..., B_b$, 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. Let θ_{ij} be the probability a randomly selected individual is combined type (A_i, B_j). Note that

$$\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} = n \quad \text{and} \quad \sum_{i=1}^{a}\sum_{j=1}^{b} \theta_{ij} = 1$$

and that the $a \times b$ frequencies $(Y_{11}, Y_{12}, ..., Y_{ab})$ follow a Multinomial distribution with $m = ab$ classes.

To test independence of the A and B classifications, we consider the hypothesis

$$H_0 : \theta_{ij} = \alpha_i \beta_j \quad \text{for } i = 1, ..., a;\ j = 1, ..., b \qquad (7.2.4)$$

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 $\alpha_i = P$(an individual is A-type A_i) and $\beta_j = P$(an individual is B-type B_j), and that (7.2.4) is the standard definition for independent events: $P(A_i \cap B_j) = P(A_i)P(B_j)$.

We recognize that testing (7.3.1) falls into the general framework of Section 7.1, where $m = ab$, $k = m - 1$, and the dimension of the parameter space under (7.2.4) is $p = (a-1) + (b-1) = a + b - 2$. All that needs to be done in order to use the statistics (6.3.7) or (6.3.8) to test H_0 given by (7.3.1) is to obtain the m.l.e.'s $\tilde{\alpha}_i$, $\tilde{\beta}_j$ under model (7.3.1), and then the expected frequencies e_{ij}. Under (7.2.4), the likelihood function for the y_{ij}'s is proportional to 2

$$\begin{aligned} L_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{i=1}^{a}\prod_{j=1}^{b}[\theta_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{y_{ij}} \\ &= \prod_{i=1}^{a}\prod_{j=1}^{b}(\alpha_i\beta_j)^{y_{ij}}. \end{aligned}$$

It is easy to maximize $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ subject to the linear constraints $\sum_{i=1}^{a}\alpha_i = 1$, $\sum_{j=1}^{b}\beta_j = 1$. This gives the maximum likelihood estimates

$$\hat{\alpha}_i = \frac{y_{i+}}{n},\ \hat{\beta}_j = \frac{y_{+j}}{n} \quad \text{and} \quad e_{ij} = n\hat{\alpha}_i\hat{\beta}_j = \frac{y_{i+}y_{+j}}{n}, \tag{7.2.5}$$

where $y_{i+} = \sum_{j=1}^{b} y_{ij}$ and $y_{+j} = \sum_{i=1}^{a} y_{ij}$. The observed value of the likelihood ratio statistic (7.1.6) for testing the hypothesis (7.2.4) is then

$$\lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\log(y_{ij}/e_{ij}).$$

The p-value is computed as

$$P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda) \quad \text{where } W \sim \chi^2((a-1)(b-1))$$

The χ^2 degrees of freedom $(a-1)(b-1)$ come from $m - 1 - p = (ab - 1) - (a + b - 2) = (a-1)(b-1)$.

Example 7.3.1. Human blood is classified according to several systems. Two 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 a person is Rh+ or Rh−. A random sample of 300 persons produced the observed frequencies in the following table. Expected frequencies, computed below, are in brackets after each observed frequency.

	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

It is of interest to see whether these two classification systems are genetically independent. The row and column totals in the table are also shown, since they are the values y_{i+} and y_{+j} needed to compute the e_{ij}'s in (7.3.3). In this case 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. Thus $a = 2$, $b = 4$ and the χ^2 degrees of freedom are $(a-1)(b-1) = 3$.

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 (7.2.5). This gives, for example,

$$e_{11} = \frac{(244)(95)}{300} = 77.3, \quad e_{12} = \frac{244(116)}{300} = 94.4$$

and, similarly, $e_{13} = 49.6$, $e_{14} = 22.8$, $e_{21} = 17.7$, $e_{22} = 21.6$, $e_{23} = 11.4$, $e_{24} = 5.2$.

It may be noted that $e_{i+} = y_{i+}$ and $e_{+j} = y_{+j}$, so it is necessary to compute only $(a-1)(b-1)$ e_{ij}'s using (7.2.5); the remainder can be obtained by subtraction from row and column totals. For example, if we compute e_{11}, e_{12}, e_{13} here then $e_{21} = 95 - e_{11}$, $e_{22} = 116 - e_{12}$, and so on. (This is not an advantage with a computer to calculate the numbers; however, it suggests where the term "degrees of freedom" comes from.)

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 some degree of evidence against the hypothesis of independence. 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 $B_1, ..., B_b$, but where the population is sub-divided into a groups $A_1, ..., A_a$. In this case, we might be interested in whether the proportions of individuals of types $B_1, ..., B_b$ 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 $i = 1, ..., a$. That is, $\theta_{ij} = P(B_j|A_i)$ and we want to know if this deends 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

$$\theta_{i1} + \theta_{i2} + \cdots + \theta_{ib} = 1 \quad \text{for each} \quad i = 1, ..., a \tag{7.2.6}$$

and the hypothesis we are interested in testing is

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_a, \tag{7.2.7}$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{ib})$. Furthermore, the data in this case arise by selecting specified numbers of individuals n_i from groups $i = 1, ..., a$ and so there are actually a multinomial distributions, Multinomial$(n_i; \theta_{i1}, ..., \theta_{ib})$.

If we denote the observed frequency of B_j-type individuals in the sample from the i'th group as y_{ij} (where $y_{i1} + \cdots + y_{ib} = n_i$), then it can be shown that the likelihood ratio statistic for testing (**??**) is exactly the same as (7.2.5), where now the expected frequencies e_{ij} are given by

$$e_{ij} = n_i \left(\frac{y_{+j}}{n} \right) \quad \text{for } i = 1, ..., a; \; j = 1, ..., b \tag{7.2.8}$$

where $n = n_1 + \cdots + n_a$. Since $n_i = y_{i+}$ 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. 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. 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. They were then 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 also 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 values shown in the table. The observed value of the likelihood ratio statistic is

$$2\sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij}\log(y_{ij}/e_{ij}) = 5.25$$

and the p-value is

$$p - value \approx P(W \geq 5.25) = 0.022 \quad \text{where } W \sim \chi^2(1)$$

so there is fairly strong evidence **against** H_0. 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 p_{11} and p_{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 confidence intervals for θ_{11} and θ_{21}; approximate 95% confidence intervals based on earlier methods are $0.211 \leq p_{11} \leq 0.323$ and $0.303 \leq p_{21} \leq 0.425$. Confidence intervals for the difference in proportions $\theta_{11}-\theta_{21}$ can also be obtained from the approximate $G(0,1)$ pivotal quantity

$$\frac{(\hat{\theta}_{11}-\hat{\theta}_{21})-(\theta_{11}-\theta_{21})}{\sqrt{\hat{\theta}_{11}(1-\hat{\theta}_{11})/n_1+\hat{\theta}_{21}(1-\hat{\theta}_{21})/n_2}}.$$

Remark: This and other tests involving binomial probabilities and contingency tables can be carried out using the *R* function *prop.test*.

7.4 Problems

1. 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:

	Cars Rust-Proofed	Cars Not Rust Proofed
Rust present	14	28
Rust absent	36	22
	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?

2. 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.

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
Frequency observed:	103	80	31	19	11	5	1

Test the hypothesis that the number of defective items Y in a single carton has a Binomial$(12, p)$ distribution. Why might the binomial not be a suitable model?

4. The numbers of service interruptions in a communications system over 200 separate weekdays is summarized in the following frequency table:

Number of interruptions:	0	1	2	3	4	≥ 5
Frequency observed:	64	71	42	18	4	1

Test whether a Poisson model for the number of interruptions Y on a single day is consistent with these data.

5. The table below records data on 292 litters of mice classified according to litter size and number of females in the litter.

		Number of females					
		0	1	2	3	4	Total number of litters
	1	8	12				20
Litter	2	23	44	13			80
Size	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 follows 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. 1398 school children with tonsils present 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
Carrier present	19	29	24
Carrier absent	497	560	269

Is there evidence of an association between the two classifications?

8. The following data on heights of 210 married couples were presented by Yule in 1900.

	Tall wife	Medium wife	Short wife
Tall husband	18	28	19
Medium husband	20	51	28
Short husband	12	25	9

Test the hypothesis that the heights of husbands and wives are independent.

9. 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 x boys in a set of triples ($x = 0, 1, 2, 3$), and test whether the column totals are consistent with this distribution.

10. A study was undertaken to determine whether there is an association between the birth weights of infants and the smoking habits of their parents. Out of 50 infants of above average weight, 9 had parents who both smoked, 6 had mothers who smoked but fathers who did not, 12 had fathers who smoked but mothers who did not, and 23 had parents of whom neither smoked. The corresponding results for 50 infants of below average weight were 21, 10, 6, and 13, respectively.

    (a) Test whether these results are consistent with the hypothesis that birth weight is independent of parental smoking habits.

    (b) Are these data consistent with the hypothesis that, given the smoking habits of the mother, the smoking habits of the father are not related to birth weight?

CAUSE AND EFFECT

8.1 Introduction

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 it does 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 it does we say** x **has a causal effect on** Y**.**

These definitions are unfortunately unusable in most settings since we cannot hold all other factors that affect y constant; often we don't even know what all the factors are. However, the definition serves as a useful ideal for how we should carry out studies in order to show that a causal relationship exists. What we do is try to design our 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. 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. Suppose that data had been collected on $10,000$ persons ages 40-80 who had smoked for at least 20 years, and $10,000$ persons in the same age range who had 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 exemplifies 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 variables fixed.

(ii) Use randomization to control for other variables.

These devices are mostly simply explained via examples.

Example 8.2.1 Blood thinning 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 variables 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 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 opts 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 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

$$P(A|B_1C_i) > P(A|B_2C_i) \text{ for each } i = 1, \ldots, k$$

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). These include

- the need to account for other possible risk factors and to demonstrate that x and Y are consistently related when these factors vary.
- the demonstration that association between x and Y holds in different types of settings
- the existence of a plausible scientific explanation

Similar criteria apply to other areas.

8.4 Example

In the early seventies, the Coronary Drug Research Group implemented a large medical trial[40] 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:

[40] *The Coronary Drug Research Group, New England Journal of Medicine (1980), pg. 1038.*

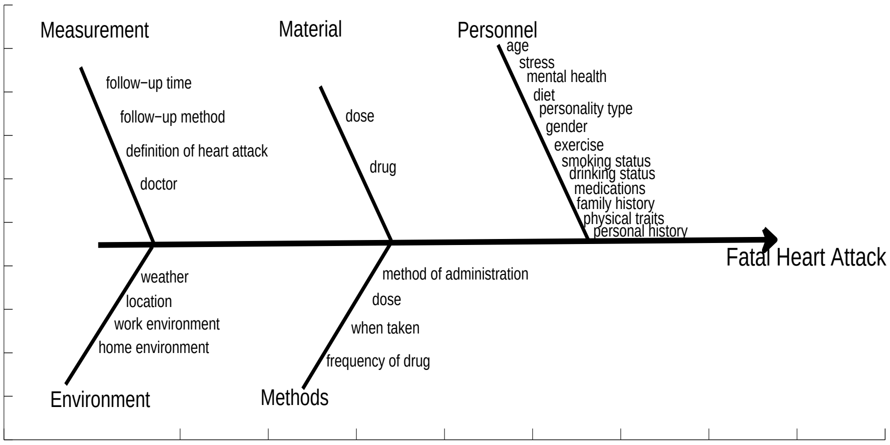


Figure 8.2: Fishbone diagram for Chlofibrate example

- 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 occurence/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 8.2) 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.)

- 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 electro-cardiograms 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 occurence 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 sufferring 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 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 are assumed roughly Gaussian, as follows:

1. Rust-proofed cars: $Y \sim G(\mu_1, \sigma)$
2. Non-rust-proofed cars: $Y \sim G(\mu_2, \sigma)$
   Sample means and variances from the two sets of cars were found to be (higher y means more rust)

$$\begin{array}{lll} 1. & \bar{y}_1 = 11.7 & s_1 = 2.1 \\ 2. & \bar{y}_2 = 12.0 & s_2 = 2.4 \end{array}$$

(a) Test the hypothesis that there is no difference in μ_1 and μ_2.

(b) The manufacturer was surprised to find that the data did not show a beneficial effect of rust-proofing. Describe problems with their study and outline how you might carry out a study designed to demonstrate a causal effect of rust-proofing.

5. In 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 **not** doing this might not be a good idea in a causative study (i.e. a study where you want to assess the causative effect of one or more treatments).

6. 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 followup in grade 11 to determine each student' s smoking “history". Briefly discuss some factors you'd want to consider in designing such a study, and how you might address them.

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

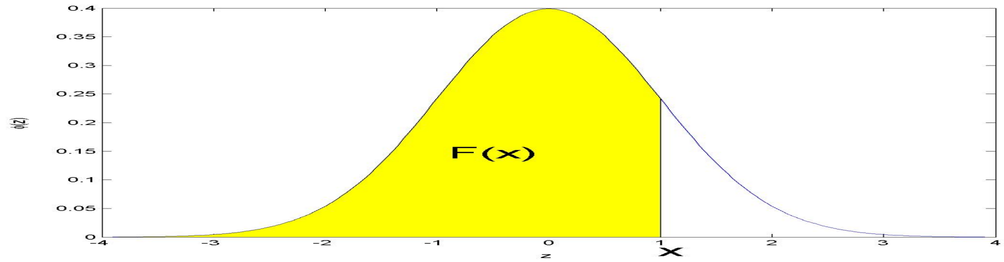
9.2 Departmental Web Resources

Videos on sections: see **www.watstat.ca**

Summary of Distributions

Discrete				
Notation and Parameters	Probability function $f(y)$	Mean	Variance	Moment generating function $M_Y(t)$
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) $0 < p < 1, q = 1 - p$	$\binom{y+k-1}{y} p^k q^y$ $y = 0,1,2,\ldots$	$\frac{kq}{p}$	$\frac{kq}{p^2}$	$(\frac{p}{1-qe^t})^k$
Geometric(p) $0 < p < 1, q = 1 - p$	pq^y $y = 0,1,2,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$	$(\frac{p}{1-qe^t})$
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	$e^{\theta(e^t-1)}$
Continuous	p.d.f. $f(y)$	E(Y)	Var(Y)	Moment generating function $M_Y(t)$
Uniform(a,b)	$f(y) = \frac{1}{b-a},\ a < y < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Exponential(θ) $0 < \theta$	$f(y) = \frac{1}{\theta}e^{-y/\theta},\ 0 < y$	θ	θ^2	$\frac{1}{1-\theta t},\ t < 1/\theta$
Normal(μ,σ^2) or Gaussian(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	$e^{\mu t+\sigma^2 t^2/2}$
Chisquared(r) d.f. $r > 0$	$f(y) = \frac{1}{2^{r/2}\Gamma(r/2)}y^{(r/2)-1}e^{-y/2}$ $0 < y < \infty$, where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$	r	$2r$	$(1-2t)^{-r/2}$
student -t d.f. $\nu > 0$	$f(y) = k_\nu(1+\frac{y^2}{\nu})^{-(\nu+1)/2}$ $-\infty < y < \infty$ where $k_\nu = \Gamma(\frac{\nu+1}{2})/\sqrt{\nu\pi}\,\Gamma(\nu/2)$	0 if $\nu > 1$	$\frac{\nu}{\nu-2}$ if $\nu > 2$	undefined

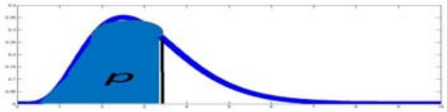
Probabilities for Standard Normal N(0,1) Distribution



The 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	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
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

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

Quantiles for a t_n distribution with n degrees of freedom

df/p	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.9995
1	0.325	0.727	1.38	3.08	6.31	12.7	31.8	63.7	637.
2	0.289	0.617	1.06	1.89	2.92	4.30	6.96	9.92	31.6
3	0.277	0.584	0.978	1.64	2.35	3.18	4.54	5.84	12.9
4	0.271	0.569	0.941	1.53	2.13	2.78	3.75	4.60	8.61
5	0.267	0.559	0.920	1.48	2.02	2.57	3.36	4.03	6.87
6	0.265	0.553	0.906	1.44	1.94	2.45	3.14	3.71	5.96
7	0.263	0.549	0.896	1.41	1.89	2.36	3.00	3.50	5.41
8	0.262	0.546	0.889	1.40	1.86	2.31	2.90	3.36	5.04
9	0.261	0.543	0.883	1.38	1.83	2.26	2.82	3.25	4.78
10	0.260	0.542	0.879	1.37	1.81	2.23	2.76	3.17	4.59
11	0.260	0.540	0.876	1.36	1.80	2.20	2.72	3.11	4.44
12	0.259	0.539	0.873	1.36	1.78	2.18	2.68	3.05	4.32
13	0.259	0.538	0.870	1.35	1.77	2.16	2.65	3.01	4.22
14	0.258	0.537	0.868	1.35	1.76	2.14	2.62	2.98	4.14
15	0.258	0.536	0.866	1.34	1.75	2.13	2.60	2.95	4.07
16	0.258	0.535	0.865	1.34	1.75	2.12	2.58	2.92	4.01
17	0.257	0.534	0.863	1.33	1.74	2.11	2.57	2.90	3.97
18	0.257	0.534	0.862	1.33	1.73	2.10	2.55	2.88	3.92
19	0.257	0.533	0.861	1.33	1.73	2.09	2.54	2.86	3.88
20	0.257	0.533	0.860	1.33	1.72	2.09	2.53	2.85	3.85
21	0.257	0.532	0.859	1.32	1.72	2.08	2.52	2.83	3.82
22	0.256	0.532	0.858	1.32	1.72	2.07	2.51	2.82	3.79
23	0.256	0.532	0.858	1.32	1.71	2.07	2.50	2.81	3.77
24	0.256	0.531	0.857	1.32	1.71	2.06	2.49	2.80	3.75
25	0.256	0.531	0.856	1.32	1.71	2.06	2.49	2.79	3.73
26	0.256	0.531	0.856	1.31	1.71	2.06	2.48	2.78	3.71
27	0.256	0.531	0.855	1.31	1.70	2.05	2.47	2.77	3.69
28	0.256	0.530	0.855	1.31	1.70	2.05	2.47	2.76	3.67
29	0.256	0.530	0.854	1.31	1.70	2.05	2.46	2.76	3.66
30	0.256	0.530	0.854	1.31	1.70	2.04	2.46	2.75	3.65
40	0.255	0.529	0.851	1.30	1.68	2.02	2.42	2.70	3.55
50	0.255	0.528	0.849	1.30	1.68	2.01	2.40	2.68	3.50
100	0.254	0.526	0.845	1.29	1.66	1.98	2.36	2.63	3.39
> 100	0.253	0.525	0.842	1.28	1.65	1.96	2.33	2.58	3.30

Table 8.2: The quantiles of the t-distribution

APPENDIX. ANSWERS TO SELECTED PROBLEMS

Chapter 1

1. (b) .032 (c) .003 (.002 using Gaussian approx.)
2. (c) $p_1 = .489$, $p_2 = .325$, $p_3 = 0.151$, $p_4 = 0.035$
3. (b) .003 and .133 (d) $y_0 = 124.3$
4. (a) .933 (b) .020 (c).949 and .117 (d) 4.56
5. (a) .9745

7. (a) $E(R) = 1 + 2(n-1)p(1-p)$
   (b) $Var(R) = 2(n-1)p(1-p)[1-2p(1-p)] + 2(n-2)p(1-p)(1-2p)^2$
   (c) $E(R) = 50.5$, $Var(R) = 24.75$ and $P(R \leq 20) < 10^{-6}$

Chapter 2

1. (a) 4.1 (b) .000275
2. (a) .10 (b) $n = 140$
3. $(2x_1 + x_2)/n$
4. (b) .28

6. (a) $\frac{(f_0+3T)-[(f_0+3T)^2-8T^2]^{1/2}}{4T}$ where $T = \sum k f_k$
   (b) $c = (1-\alpha)^2/\alpha$
   (c) $\hat{\alpha} = 0.195$; $\hat{Pr}(X = 0) = .758$
   (d) $\hat{\alpha} = .5$
7. $\hat{\lambda} = \sum y_i / \sum t_i$

9. (a) $\hat{\alpha} = .35$ $\hat{\beta} = .42$
   (b) 14.7, 20.3, 27.3 and 37.7

Chapter 4

1. (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) .098 (M) and .0004 (F)
   (d) 11/50=.073 (M) 0(F)

2. (c) 0.1414 and 0.1768, respectively

3. (b) $n = 1024$

7. (b) $\hat{\theta} = 1 - (x/n)^{1/k}$ (c) $\hat{\theta} = 0.0116$; interval approximately (.0056,.0207)

8. (a) $0 \le \alpha \le .548$ (b) .10 likelihood interval is now $0.209 \le \alpha \le .490$

10. (a) $\hat{\lambda} = 3n/\sum t_i$ (b) $\hat{\lambda} = 0.06024$; $.0450 \le \lambda \le 0.0785$
    (c) .95 CI for λ is (.0463,.0768) and for μ is $39.1 \le \mu \le 64.8$
    (d) CI's are $.408 \le p \le .738$ (using model) and $0.287 \le p \le .794$ (using binomial). The binomial model involves fewer assumptions but gives a less precise (wider) interval.
    (Note: the 1st CI can be obtained directly from the CI for λ in part (c).)

12. (a) $\hat{\theta} = 380$ days; CI is $285.5 \le \theta \le 521.3$
    (b) $197.9 \le m \le 361.3$

13. (b) $288.3 \le \theta \le 527.9$

14. (a) $.637 \le p \le .764$

Chapter 5

1. $SL = P(D \ge 15) = P(Y \ge 25; \lambda = 10) = 0.000047$

4. (a) LR statistic gives $\Lambda_{0bs} = 0.0885$ and $SL = .76$.

5. (a) $\Lambda_{0bs} = 23.605$ and $SL = 0.005$
   (b) $SL = 1 - .995^6 = 0.03$ now

6. $\Lambda_{0bs} = 0.042$ and $SL = .84$. There is no evidence against the model.

9. (c) LR statistic gives $\Lambda_{0bs} = 3.73$ and $SL = P(\chi^2_{(4)} \geq 3.73) = .44$. There is no evidence that the rates are not equal.

Chapter 6

3. (a) $43.28 \leq \mu \leq 45.53$ (b) $1.82 \leq \sigma \leq 3.50$

4. (a) $D_{0bs} = 9s^2/.02^2 = 1.2$ and $SL = 2P(\chi^2_{(9)} \leq 1.2) = 0.0024$ so there is strong evidence against $H : \sigma = 0.02$

   (b) No: testing $H : \mu = 13.75$ gives $SL < .001$

   (c) $13.690 \leq \mu \leq 13.700$ and $.0050 \leq \sigma \leq 0.0132$

5. (a) $296.91 \leq \mu \leq 303.47$; $4.55 \leq \sigma \leq 9.53$

   (b) $286.7 \leq X \leq 313.7$

7. $.75 \leq \beta \leq 11.25$ where $\beta = \mu_1 - \mu_2$

8. (a) $0.64 \leq \mu_1 - \mu_2 \leq 7.24$

   (b) $SL = 0.05$ (c) $SL = 0.07$

9. (a) LR test gives $SL = .4$

   (b) $-.011 \leq \mu_1 - \mu_2 \leq .557$

12. (a) $-0.23 \leq \beta \leq 2.38$ (b) $-8.77 \leq \beta \leq 10.92$

18. (a) $\hat{\beta} = 0.9935$, $\hat{\alpha} = -0.0866$, $s = 0.2694$. Confidence intervals are $0.978 \leq \beta \leq 1.009$ and $0.182 \leq \sigma \leq 0.516$

19. (b) $\hat{\beta} = 0.02087$, $\hat{\alpha} = -1.022$, $s = 0.008389$

   (c) .95 prediction interval for $Y(\log P)$ is $3.030 \leq Y \leq 3.065$ so PI for P is $20.70 \leq P \leq 21.43$.

Chapter 7

1. (a) LR statistic gives $\Lambda_{0bs} = 8.17$ and Pearson statistic $D_{0bs} = 8.05$. The SL is about .004 in each case so there is strong evidence against H.

2. LR statistic gives $\Lambda_{0bs} = 5.70$ and Pearson statistic $D_{0bS} = 5.64$. The SL is about .017 in each case.

5. (a) LR statistics for $n = 2, 3, 4$ are 1.11, 4.22, 1.36. The SL's are $P(\chi^2_{(1)} \geq 1.11) = 0.29$, $P(\chi^2_{(2)} \geq 4.22) = 0.12$, and $P(\chi^2_{(3)} \geq 1.36) = .71$, respectively.

   (b) LR statistic is 7.54 and $SL = P(\chi^2_{(3)} \geq 7.54) = 0.057$.

7. The LR statistic is 7.32 and $SL = P(\chi^2_{(2)} \geq 7.32) = 0.026$ so there is evidence against independence and in favour of an association.

8. LR statistic is 3.13 and $SL = P(\chi^2_{(4)} \geq 3.13) = .54$. There is no evidence against independence.

9. (a) LR statistic gives $\Lambda_{0bs} = 0.57$ and $SL = P(\chi^2_{(3)} \geq .57) = 0.90$ so there is no evidence of association.

   (b) LR statistic gives $\Lambda_{0bs} = 5.44$ and $SL = P(\chi^2_{(3)} \geq 5.44) = 0.14$. There is no evidence against the binomial model.

10. (a) $\Lambda_{0bs} = 10.8$ and $SL = P(\chi^2_{(3)} \geq 10.8) = 0.013$.

Chapter 8

1. (a) LR statistic is 480.65 so SL is almost zero; there is very strong evidence against independence.

3. (a) LR statistic gives $\Lambda_{0bs} = 112$ and $SL = 0$

   (b) Only Program B shows any evidence of non-independence, and that is in the direction of a lower admission rate for males.

APPENDIX: DATA

Here we list the data for Example 1.5.2. In the file *ch1example152.txt*, there are three columns labelled hour, machine and volume. The data are

hour	machine	volume	hour	machine	volume
1	1	357.8	11	1	357
1	2	358.7	11	2	359.6
2	1	356.6	12	1	357.1
2	2	358.5	12	2	357.6
3	1	357.1	13	1	356.3
3	2	357.9	13	2	358.1
4	1	357.3	14	1	356.3
4	2	358.2	14	2	356.9
5	1	356.7	15	1	356
5	2	358	15	2	356.4
6	1	356.8	16	1	357
6	2	359.1	16	2	357.5
7	1	357	17	1	357.5
7	2	357.5	17	2	357.2
8	1	356	18	1	355.9
8	2	356.4	18	2	357.1
9	1	355.9	19	1	356.5
9	2	357.9	19	2	358.2
10	1	357.8	20	1	355.8
10	2	358.5	20	2	359

21	1	356.5	31	1	357.7
21	2	357.3	31	2	357
22	1	356.9	32	1	356.3
22	2	356.7	32	2	357.8
23	1	357.5	33	1	356.6
23	2	356.9	33	2	357.5
24	1	356.9	34	1	356.7
24	2	357.1	34	2	356.5
25	1	356.9	35	1	356.8
25	2	356.4	35	2	357.6
26	1	356.4	36	1	356.6
26	2	357.5	36	2	357.2
27	1	356.5	37	1	356.6
27	2	357	37	2	357.6
28	1	356.5	38	1	356.7
28	2	358.1	38	2	356.9
29	1	357.6	39	1	356.8
29	2	357.6	39	2	357.2
30	1	357.5	40	1	356.1
30	2	356.4	40	2	356.4

Example 1.3.1 New Zealand BMI Data

subject	gender	height	weight	BMI
1	M	1.76	63.81	20.6
2	M	1.77	89.6	28.6
3	M	1.91	88.65	24.3
4	M	1.8	74.84	23.1
5	M	1.81	97.3	29.7
6	M	1.93	106.9	28.7
7	M	1.79	108.94	34
8	M	1.66	74.68	27.1
9	M	1.66	92.31	33.5
10	M	1.82	92.08	27.8
11	M	1.76	93.86	30.3
12	M	1.79	88.11	27.5
13	M	1.77	80.52	25.7
14	M	1.72	75.14	25.4
15	M	1.73	64.95	21.7
16	M	1.81	89.11	27.2
17	M	1.77	96.49	30.8
18	M	1.56	53.78	22.1

19	M	1.71	76.61	26.2
20	M	1.8	82.62	25.5
21	M	1.68	80.44	28.5
22	M	1.75	93.1	30.4
23	M	1.81	71.09	21.7
24	M	1.69	71.12	24.9
25	M	1.74	80.84	26.7
26	M	1.73	75.12	25.1
27	M	1.74	96.88	32
28	M	1.8	73.22	22.6
29	M	1.75	81.77	26.7
30	M	1.81	83.87	25.6
31	M	1.72	55.91	18.9
32	M	1.74	68.73	22.7
33	M	1.74	75.39	24.9
34	M	1.78	94.1	29.7
35	M	1.75	80.54	26.3
36	M	1.68	70.84	25.1
37	M	1.78	100.76	31.8
38	M	1.68	51.65	18.3
39	M	1.75	84.83	27.7
40	M	1.71	70.47	24.1
41	M	1.73	112.23	37.5
42	M	1.71	72.23	24.7
43	M	1.87	105.26	30.1
44	M	1.69	69.97	24.5
45	M	1.73	102.36	34.2
46	M	1.71	81.58	27.9
47	M	1.86	80.61	23.3
48	M	1.73	76.62	25.6
49	M	1.64	71.27	26.5
50	M	1.59	60.17	23.8
51	M	1.78	92.2	29.1
52	M	1.73	78.41	26.2
53	M	1.76	90.76	29.3
54	M	1.8	92.34	28.5
55	M	1.71	68.72	23.5
56	M	1.69	76.54	26.8
57	M	1.8	90.72	28
58	M	1.78	70.66	22.3
59	M	1.73	76.32	25.5

60	M	1.71	88.02	30.1
61	M	1.78	87.76	27.7
62	M	1.74	84.77	28
63	M	1.69	67.4	23.6
64	M	1.82	83.14	25.1
65	M	1.63	69.08	26
66	M	1.74	72.36	23.9
67	M	1.74	69.03	22.8
68	M	1.69	81.68	28.6
69	M	1.79	89.39	27.9
70	M	1.79	75.3	23.5
71	M	1.86	90.3	26.1
72	M	1.7	102.59	35.5
73	M	1.87	94.42	27
74	M	1.65	89.03	32.7
75	M	1.72	78.4	26.5
76	M	1.74	93.55	30.9
77	M	1.69	68.26	23.9
78	M	1.57	53.73	21.8
79	M	1.74	91.13	30.1
80	M	1.8	89.1	27.5
81	M	1.77	87.41	27.9
82	M	1.71	66.38	22.7
83	M	1.78	106.46	33.6
84	M	1.56	66.92	27.5
85	M	1.74	79.93	26.4
86	M	1.79	92.28	28.8
87	M	1.85	79.4	23.2
88	M	1.64	70.2	26.1
89	M	1.83	116.88	34.9
90	M	1.7	78.32	27.1
91	M	1.72	102.66	34.7
92	M	1.72	78.4	26.5
93	M	1.7	83.81	29
94	M	1.64	67.51	25.1
95	M	1.75	69.83	22.8
96	M	1.68	77.62	27.5
97	M	1.71	95.03	32.5
98	M	1.67	74.18	26.6
99	M	1.8	92.99	28.7
100	M	1.77	78.64	25.1

101	M	1.72	79.29	26.8
102	M	1.66	72.75	26.4
103	M	1.78	83.65	26.4
104	M	1.6	61.44	24
105	M	1.72	65.97	22.3
106	M	1.71	78.37	26.8
107	M	1.79	74.01	23.1
108	M	1.74	69.33	22.9
109	M	1.74	88.1	29.1
110	M	1.78	89.35	28.2
111	M	1.77	90.54	28.9
112	M	1.74	91.43	30.2
113	M	1.84	94.8	28
114	M	1.82	86.12	26
115	M	1.83	75.35	22.5
116	M	1.74	70.85	23.4
117	M	1.74	98.7	32.6
118	M	1.89	104.66	29.3
119	M	1.81	91.08	27.8
120	M	1.64	94.67	35.2
121	M	1.77	80.2	25.6
122	M	1.73	73.92	24.7
123	M	1.82	84.8	25.6
124	M	1.73	90.39	30.2
125	M	1.77	74.25	23.7
126	M	1.82	107.32	32.4
127	M	1.8	80.03	24.7
128	M	1.77	105.58	33.7
129	M	1.8	110.48	34.1
130	M	1.7	93.64	32.4
131	M	1.7	68.49	23.7
132	M	1.77	77.7	24.8
133	M	1.77	97.12	31
134	M	1.62	70.86	27
135	M	1.74	82.96	27.4
136	M	1.68	72.25	25.6
137	M	1.64	73.16	27.2
138	M	1.75	92.49	30.2
139	M	1.66	66.69	24.2
140	M	1.86	106.21	30.7
141	M	1.72	88.75	30

142	M	1.69	73.97	25.9
143	M	1.72	81.95	27.7
144	M	1.77	82.4	26.3
145	M	1.66	85.42	31
146	M	1.78	76.04	24
147	M	1.82	78.5	23.7
148	M	1.84	98.86	29.2
149	M	1.75	85.44	27.9
150	M	1.75	65.23	21.3
151	F	1.6	59.9	23.4
152	F	1.6	48.38	18.9
153	F	1.51	77.98	34.2
154	F	1.6	54.53	21.3
155	F	1.67	79.2	28.4
156	F	1.55	87.45	36.4
157	F	1.61	53.66	20.7
158	F	1.56	64	26.3
159	F	1.6	67.58	26.4
160	F	1.58	70.65	28.3
161	F	1.56	51.59	21.2
162	F	1.67	56.89	20.4
163	F	1.64	54.6	20.3
164	F	1.67	63.31	22.7
165	F	1.53	52.67	22.5
166	F	1.6	48.64	19
167	F	1.67	69.72	25
168	F	1.79	65.04	20.3
169	F	1.54	67.35	28.4
170	F	1.65	65.34	24
171	F	1.61	80.87	31.2
172	F	1.76	85.8	27.7
173	F	1.52	87.56	37.9
174	F	1.58	59.16	23.7
175	F	1.69	94.82	33.2
176	F	1.57	60.39	24.5
177	F	1.64	63.47	23.6
178	F	1.7	62.13	21.5
179	F	1.6	63.49	24.8
180	F	1.59	64.21	25.4
181	F	1.64	72.89	27.1
182	F	1.57	74.19	30.1

183	F	1.59	82.67	32.7
184	F	1.53	59.93	25.6
185	F	1.64	79.61	29.6
186	F	1.73	69.14	23.1
187	F	1.57	81.59	33.1
188	F	1.61	63.51	24.5
189	F	1.68	82.13	29.1
190	F	1.57	58.91	23.9
191	F	1.65	70.51	25.9
192	F	1.6	71.42	27.9
193	F	1.62	59.57	22.7
194	F	1.64	57.56	21.4
195	F	1.54	61.9	26.1
196	F	1.58	84.63	33.9
197	F	1.7	66.76	23.1
198	F	1.56	75.68	31.1
199	F	1.68	72.25	25.6
200	F	1.53	56.88	24.3
201	F	1.58	66.9	26.8
202	F	1.59	50.06	19.8
203	F	1.64	69.66	25.9
204	F	1.63	87.15	32.8
205	F	1.66	76.61	27.8
206	F	1.53	62.03	26.5
207	F	1.66	88.73	32.2
208	F	1.65	85.21	31.3
209	F	1.67	81.99	29.4
210	F	1.6	77.82	30.4
211	F	1.71	84.21	28.8
212	F	1.61	69.99	27
213	F	1.65	96.92	35.6
214	F	1.6	77.57	30.3
215	F	1.71	78.37	26.8
216	F	1.58	77.39	31
217	F	1.61	64.28	24.8
218	F	1.59	85.96	34
219	F	1.57	64.58	26.2
220	F	1.64	76.92	28.6
221	F	1.72	71.89	24.3
222	F	1.59	58.9	23.3
223	F	1.64	86.07	32

224	F	1.64	78	29
225	F	1.58	66.9	26.8
226	F	1.53	61.1	26.1
227	F	1.62	59.05	22.5
228	F	1.62	83.72	31.9
229	F	1.61	76.99	29.7
230	F	1.57	61.62	25
231	F	1.72	107.09	36.2
232	F	1.61	45.36	17.5
233	F	1.67	89.8	32.2
234	F	1.67	77.25	27.7
235	F	1.6	82.94	32.4
236	F	1.66	82.12	29.8
237	F	1.58	74.64	29.9
238	F	1.71	79.54	27.2
239	F	1.64	61.32	22.8
240	F	1.59	60.17	23.8
241	F	1.61	95.91	37
242	F	1.56	62.79	25.8
243	F	1.56	48.19	19.8
244	F	1.54	69.73	29.4
245	F	1.52	89.64	38.8
246	F	1.57	57.68	23.4
247	F	1.67	75.02	26.9
248	F	1.57	40.42	16.4
249	F	1.57	53	21.5
250	F	1.68	101.61	36
251	F	1.72	110.94	37.5
252	F	1.68	65.48	23.2
253	F	1.77	73	23.3
254	F	1.65	71.6	26.3
255	F	1.41	46.72	23.5
256	F	1.54	73.99	31.2
257	F	1.67	79.48	28.5
258	F	1.72	60.06	20.3
259	F	1.72	63.01	21.3
260	F	1.61	81.65	31.5
261	F	1.52	85.95	37.2
262	F	1.61	54.95	21.2
263	F	1.55	78.56	32.7
264	F	1.57	64.58	26.2

265	F	1.51	76.84	33.7
266	F	1.69	81.11	28.4
267	F	1.69	78.54	27.5
268	F	1.58	72.65	29.1
269	F	1.48	65.49	29.9
270	F	1.66	60.07	21.8
271	F	1.47	61.37	28.4
272	F	1.63	71.2	26.8
273	F	1.71	66.38	22.7
274	F	1.59	70.79	28
275	F	1.56	73.49	30.2
276	F	1.62	70.07	26.7
277	F	1.53	61.57	26.3
278	F	1.7	74.27	25.7
279	F	1.6	45.06	17.6
280	F	1.52	67.93	29.4
281	F	1.61	53.66	20.7
282	F	1.58	64.66	25.9
283	F	1.71	66.67	22.8
284	F	1.58	72.65	29.1
285	F	1.65	79.22	29.1
286	F	1.65	74.32	27.3
287	F	1.7	85.83	29.7
288	F	1.7	67.63	23.4
289	F	1.66	77.98	28.3
290	F	1.67	85.9	30.8
291	F	1.64	67.51	25.1
292	F	1.68	60.96	21.6
293	F	1.54	64.03	27
294	F	1.58	61.41	24.6
295	F	1.68	75.64	26.8
296	F	1.64	64.82	24.1
297	F	1.65	59.62	21.9
298	F	1.66	76.05	27.6
299	F	1.6	61.7	24.1
300	F	1.65	76.5	28.1

Brakepad Lifetimes (1000km)

32.4	125.4	75.1	70.3	18.8	61.2	43.9	50.5
25.4	90.4	61.8	26.4	50.2	59.7	21.1	108.4
177	44.8	61.2	67.3	18.2	22	41.4	28.1
87.6	17.5	73.9	24.2	37.6	19.2	68.5	21.4
54.7	110.4	31.9	32.8	38.1	27.2	43	40.3
70.8	138	14.5	16.3	71.1	62.3	33.1	85.1
7.4	96.5	29.5	54.3	69.9	38.3	14.5	53.5
52.9	2.6	72.7	36.9	59.5	48.2	40.4	10.9
26.6	42.6	42.5	74.9	113.4	102.3	30.6	70.2
69.3	13.7	29.6	36.1	30.7	36.3	53.4	17.4
91.3	39.9	71.8	44.3	25.3	82.3	31.5	38
31.6	40.1	115	6.1	10.1	100.9	19.3	25.5
31.1	6.5	167.2	88.4	39.3	47.6	14.2	169.3
22	90.3	26.5	80	23.4	5.8	8.3	20
57.5	66.4	31	21.6	31.2	136.3	108.2	48
21.9	26.9	32.8	27.6	103.2	9.2	35.5	42.3
23.1	36.3	11.5	0.9	32	47.2	18.8	49.5
34.4	40	8.3	44.4	10.6	28.1	59.3	44.5
41.3	43.4	17.8	44.5	121.8	8.8	45.1	66.2
29.6	27.1	11.1	25.4	46.1	42.3	55	24.2
15.6	74.5	18.7	33.6	61.6	53.5	105.1	55.8