STATISTICS 231 COURSE NOTES

Original notes by Jerry Lawless

Winter 2013 Edition

Contents

INTRODUCTION TO STATISTICAL SCIENCE **2**

1.1 Statistical Science 2
1.2 Collection of Data 3
1.3 Data Summaries 7
1.4 Probability Distributions and Statistical Models 12
1.5 Data Analysis and Statistical Inference 16
1.6 Statistical Software 19
1.7 A More Detailed Example: Colour Classification by Robots 25
1.8 Appendix. The R Language and Software 30
1.9 Problems 35

MODEL FITTING, MAXIMUM LIKELIHOOD ESTIMATION, AND MODEL CHECKING **43**

2.1 Statistical Models and Probability Distributions 43
2.2 Estimation of Parameters (Model Fitting) 47
2.3 Likelihood Functions From Multinomial Models 54
2.4 Checking Models 56
2.5 Problems 60

PLANNING AND CONDUCTING EMPIRICAL STUDIES **66**

3.1 Empirical Studies 66
3.2 Planning a Study 70
3.3 Data Collection 72
3.4 Problems 73

STATISTICAL INFERENCE: ESTIMATION **75**

4.1 Introduction 75
4.2 Some Distribution Theory 79

4.3 Confidence Intervals for a Parameter 88
4.4 Problems 98

STATISTICAL INFERENCE: TESTING HYPOTHESES **106**
5.1 Introduction 106
5.2 Testing Parametric Hypotheses with Likelihood Ratio Statistics 110
5.3 Hypothesis Testing and Interval Estimation 115
5.4 Problems 116

GAUSSIAN RESPONSE MODELS **121**
6.1 Introduction 121
6.2 Inference for a single sample from a Gaussian Distribution 126
6.3 General Gaussian Response Models 131
6.4 Inference for Paired Data 136
6.5 Linear Regression Models 142
6.6 Model Checking 149
6.7 Problems 151

TESTS AND INFERENCE PROBLEMS BASED ON MULTINOMIAL MODELS **161**
7.1 Introduction 161
7.2 Goodness of Fit Tests 162
7.3 Two-Way Tables and Testing for Independence of Two Variables 164
7.4 Problems 169

CAUSE AND EFFECT **173**
8.1 Introduction 173
8.2 Experimental Studies 175
8.3 Observational Studies 177
8.4 Problems 178

References and Supplementary Resources **182**

Statistical Tables **183**

APPENDIX. ANSWERS TO SELECTED PROBLEMS **184**

A Short Review of Probability **188**

INTRODUCTION TO STATISTICAL SCIENCE

1.1 Statistical Science

Statistical science, or statistics, is the discipline that deals with the collection, analysis and interpretation of data, and with the study and treatment of variability and of uncertainty. If you think about it, you soon realize that almost everything we do or know depends on data of some kind, and that there is usually some degree of uncertainty present. For example, in deciding whether to take an umbrella on a long walk we may utilize information from weather forecasts along with our own direct impression of the weather, but even so there is usually some degree of uncertainty as to whether it will actually rain or not. In areas such as insurance or finance, decisions must be made about what rates to charge for an insurance policy, or whether to buy or sell a stock, on the basis of certain types of data. The uncertainty as to whether a policy holder will have a claim over the next year, or whether a stock's price will rise or fall, is the basis of financial risk for the insurer and the investor.

In order to increase our knowledge about some area or to make better decisions, we must collect and analyze data about the area in question. To discuss general ways of doing this, it is useful to have terms that refer to the objects we are studying. The words "population", "phenomenon" and "process" are frequently used below; they are simply catch-all terms that represent groups of objects and events that someone might wish to study.

Variability and uncertainty are present in most processes and phenomena in the real world. Uncertainty or lack of knowledge is the main reason why someone chooses to study a phenomenon in the first place. For example, a medical study to assess the effect of a new drug for controlling hypertension (high blood pressure) may be conducted by a drug company because they do not know how the drug will perform on different types of people, what its side effects will be, and so on. Variability is ever-present; people have varying degrees of hypertension, they react differently to drugs, they have different physical characteristics. One might similarly want to study variations in currency or stock values, variation in sales for a company over time, or variation in hits and response times for a commercial web site. Statistical science deals both with the study of variability in processes and phenomena, and with good (i.e. informative, cost-effective) ways to collect and analyze data about such processes.

There are various possible objectives when one collects and analyzes data on a population, phenomenon, or process. In addition to pure "learning" or furthering knowledge, these 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 their scientific knowledge, but also to provide information to legislators or groups 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 or in determining costs; statistical process control is used to increase the quality and productivity of manufacturing 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 does not equal information, and it is the purpose of statistical science 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.

Mathematical models are used to represent many phenomena, populations, or processes and to deal with problems that involve variability we utilize probability models. These have been introduced and studied in your first probability course, and you have seen how to describe variability and solve certain types of problems using them. This course will focus more on the collection, analysis and interpretation of data, but the probability models studied earlier will be heavily used. The most important part of probability for this course is the material dealing with random variables and their probability distributions, including distributions such as the binomial, hypergeometric, Poisson, multinomial, normal and exponential. You should review your previous notes on this material.

Statistical science is a large discipline, and this course is only an introduction. Our broad objectives are to discuss the collection, analysis and interpretation of data, and to show why this is necessary. By way of further introduction we will outline important statistical topics, first data collection, and then probability models, data analysis, and statistical inference. We should bear in mind that study of a process or phenomenon involves iteration between model building, data collection, data analysis, and interpretation. We must also 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.

1.2 Collection of Data

The objects of study in this course are usually referred to as either populations or processes. In essence a *population* is just some collection of units (which can be either real or imagined), for example, the

collection of persons under the age of 18 in Canada as of September 1, 2012 or the collection of car insurance policies issued by a company over a one year period. A *process* is a mechanism by which output of some kind is produced; units can often be associated with the output. For example, hits on a website constitute a process (the “units" are the distinct hits), as do 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).

Populations or processes are studied by defining *variates* or *variables* which represent characteristics of units. These are usually numerical-valued and are represented by letters such as x, y, z. For example, we might define a variable y as the number of car insurance claims from an individual policy holder in a given year, or as the number of hits on a website over a specified one hour period. The values of y vary across the units in a population or process, this variability which generates uncertainty and makes it necessary to study populations and processes by collecting data about them. By "data" we mean here the values of the variates for specific units in the population.

In planning for the collection of data about some phenomenon, we must carefully specify what the objectives of doing this are. Then, the feasibility of obtaining information by various means must be considered, as well as to what extent it will be possible to answer questions of interest. This sounds simple but is usually difficult to do well, especially with limited resources.

There are several ways in which data are commonly obtained. One 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 exactly what "available" data represent, and they may be unsuitable for serious analysis. For example, people who voluntarily provide data in a survey may not be representative of the population at large. Statistical science stresses the importance of obtaining data so that they will be “objective" and provide maximal information. Three broad approaches are often used to do this:

(i) **Sample Surveys** The object of many studies is a finite population of some sort (e.g. all persons over 19 in Ontario; all cars produced by GM in the past year). In this case information may be obtained by selecting a “representative” sample of individuals from the population and studying them. Representativeness of the sample is usually achieved by selecting the sample members randomly from those in the population. Sample surveys are widely used in government statistical studies, economics, marketing, public opinion polls, sociology, and other areas.

(ii) **Observational Studies** An observational study is one in which data are collected about a process or phenomenon (over which the observer has no control) in some objective way, often over some period of time. For example, in studying risk factors associated with a disease such as lung cancer, one might investigate all such cases (or perhaps a random sample of them) that occur

over a given time period. A distinction between a sample survey and an observational study is that for the latter the "population" of interest is usually infinite or conceptual. For example, in investigating risk factors for a disease we would prefer to think of the population 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 collecting the data) exercises some degree of control over the process being studied. This usually takes the form of the experimenter being able to control certain factors in the process. For example, in an engineering experiment to quantify the effect of temperature on the performance of personal computers, we might decide to run an experiment with 40 PC's, ten of which would be operated at each of the temperatures 10, 20, 30, and 40 degrees Celsius.

The three types of studies described above are not mutually exclusive, and many studies involve aspects of two or more 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, people's smoking habits and the population referred to 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 possibly contact all persons in the population, we decide to select a random sample of n persons. The data to be obtained from each person might consist of their age, sex, place of residence, occupation, whether they currently smoke, and some additional information about their smoking habits and how long they have smoked (if they are smokers or ex-smokers).

Note that we have to decide how large n should be, and how we are going to obtain our random sample. The latter question is, in particular, very important if we want to ensure that our sample is indeed "representative" of the 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 about 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 (or only a very small fraction of) packages contain less than the required amount.

Consider, for example, soft drinks sold in nominal 26 ounce bottles. Because of inherent variation in the bottle filling process (what might some sources of this be?), the amount of liquid x that goes into

a bottle varies over a small range. Note that the manufacturer would like the variability in x to be as small as possible, and for bottles to contain at least 26 ounces. Suppose that the manufacturer has just added a new filling machine to increase the plant's capacity and wants to compare the new machine with the older ones. She decides to do this by sampling some filled bottles from each machine and accurately measuring the amount of liquid x in each bottle; this will be an observational study.

How exactly should the data be collected? The machines may "drift" over time (i.e. the average or the variability in the values of x may vary systematically up or down over time) so we should randomly select bottles over time from each machine; we would have to decide how many, and over what time periods to collect them.

Example 1.2.3 Clinical trials in medicine

In medical studies of the treatment of disease, it is common to compare alternative treatments in experiments called clinical trials. Consider, for example, persons who are considered 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 persons daily doses of aspirin (call this Treatment 1) and others a daily dose of a placebo, that is, an inactive compound, given in the same form as the aspirin (call this Treatment 2). The two groups of persons were then followed for a period of time, and the number of strokes in each group was observed.

This sounds simple, but there are several important points. For example, patients should be assigned to receive Treatment 1 or Treatment 2 in some random fashion so as to avoid unconscious bias (e.g. doctors might otherwise tend to put persons at higher risk 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 variables should we measure as the basis for our data? 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 the 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, credit cards held), spending, and other variates. Based on this data, the company wishes to select persons whom it considers have a good chance of responding positively to the mailout. 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

We noted in previous section that data consisting of measurements on variables $x, y, z, \ldots$ of interest are collected when we study a phenomenon or process. Data in raw form can be difficult to comprehend, especially if the volume is great or if there are large numbers of variables. Many methods of summarizing data so they can be more easily understood have been developed. There are two main types: graphical and numerical. We will consider a few important data summaries here.

The basic setup is as follows. Suppose that data on a variable x 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 x-value as $x_1, x_2, \ldots, x_n$. We might also collect data on a second variate y for each unit, and we would denote the values as $y_1, y_2, \ldots, y_n$. We often refer 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 samples or data sets, and refer to n as the *sample size*.

First we describe some graphical summaries of data sets like this, and then we describe some numerical summaries.

1.3.1 Numerical Summaries

Numerical data summaries are useful for describing features of a data set $\{y_1, \ldots, y_n\}$. Important ones are

- The *mean* (also called the sample mean) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- the *(sample) variance* $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- the *(sample) standard deviation* $s_y = \sqrt{s_y^2}$
- the *percentiles and quantiles*: the p'th quantile or $100p$'th percentile is a y-value $Q(p)$ such that a fraction p of the values in the data set are below $Q(p)$. The values $Q(.5)$, $Q(.25)$ and $Q(.75)$ are called the median, the lower quartile, and the upper quartile respectively. In fact, quantiles

are not uniquely defined for all p-values in $(0, 1)$ for a given data set, and there are different conventions for defining quantiles and percentiles. For example what is the median of the values $\{1, 2, 3, 4, 5, 6\}$? What is the lower quartile? The different conventions for defining quantiles become identical as n becomes large.

The mean and the percentiles and quantiles are easily understood. The variance and standard deviation measure the variability or "spread" of the y-values in a data set, which is usually an important characteristic. Another way to measure variability is in terms of the distance between a "low" and "high" percentile, for example $Q(.10)$ and $Q(.90)$.

A final numerical summary is a *frequency table*. This is closely related to a histogram and, in fact, is just a table showing the interval I_j and their frequencies f_j, as used in a histogram. For example, for the 200 male height measurements in Example 1.4.2, the frequency table corresponding to the bottom-left histogram in Figure 1.1 is shown in Table 1.3.1.

Table 1.3.1 Frequency Table of Male Heights (in m.)

Interval (I_j)	Frequency(f_j)
$[1.55, 1.60)$	2
$[1.60, 1.65)$	13
$[1.65, 1.70)$	48
$[1.70, 1.75)$	64
$[1.75, 1.80)$	42
$[1.80, 1.85)$	25
$[1.85, 1.90)$	6
Total	200

1.3.2 Graphical Summaries

We consider the first two types of plots for a data set $\{y_1, y_2, ..., y_n\}$ of numerical values. These are called histograms and cumulative frequency plots.

Histograms

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

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

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 on each interval; the height of the rectangle for I_j is chosen so that the rectangle's area is proportional to f_j. Two main types of histogram are used:

(a) a “standard" histogram where the range of y is taken to be finite and the I_j are of equal length. The height of the rectangle is taken to be 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, which we call the *relative frequency* for I_j. Note that in this case the sum of the areas of all the rectangles in the histogram equals one.

Example 1.3.2
Figure 1.1 shows relative frequency histograms based on each of two samples, (a) heights of 200 females randomly selected from workers aged 18 - 60 in New Zealand, and (b) heights of 200 males, selected from the same population. Heights are recorded in metres; the female heights range from 1.45 to 1.78m (57.1 to 70.1 in.) and the males heights from 1.59 to 1.88m (62.6 to 74.0 in.).

To construct a histogram, we have to choose the number (k) 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 $f_j \geq 1$). Software packages are used to produce histogram plots (see Section 1.6) and they will either automatically select the intervals for a given data set or allow the user to specify them.

The visual impression from a histogram can change somewhat according to the choice of intervals. In Figure 1.1, the left-hand panels use 7 intervals and the right-hand use 17 for females and 15 for males. Note that the histograms give a picture of the distribution of y values in the two samples. For both females and males the distributions are fairly symmetrical-looking. To allow easy comparison of female and male height distributions we have used the same y scale (x-axis) for males and females. Obviously, the distribution of male heights is to the right of that for female heights, but the “spread" and shape of the two distributions is similar.

Example 1.3.3 Different shapes of distributions can occur in data on a variable y. Figure 1.2 shows a histogram of the lifetimes (in terms of number of km driven) for the front brake pads on 200 new mid-size cars of the same type. Notice that the distribution is less symmetrical than the ones in Figure 1.1; the brake pad lifetimes have a rather long right-hand 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 a data set $\{y_1, y_2, ..., y_n\}$ is to count the number or proportion of values in the set which are smaller than any given value. This gives a function

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

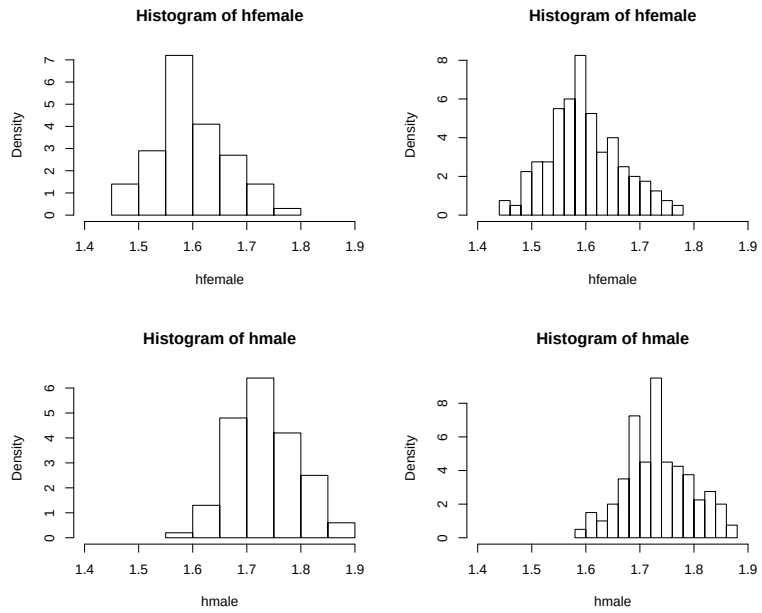


Figure 1.1: Histograms for Femaile and Male Heights. Sample sizes=200.

Software will produce such functions for a given data set. This is conveniently done by first ordering the y_i's $(i = 1, \ldots, n)$ to give the ordered values $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$. Then, we note that $\tilde{F}(y)$ is a "staircase" or "step" function that is easily obtained from the ordered values. If the data values $y_i (i = 1, \ldots, n)$ are all different, then $\tilde{F}(y_{(j)}) = j/n$.

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

$$\tilde{F}(y) = \begin{cases} 0 & y < 1.5 \\ .25 & 1.5 \leq y < 2.2 \\ .50 & 2.2 \leq y < 3.4 \\ .75 & 3.4 \leq y < 5.0 \\ 1.00 & y \geq 5.0 \end{cases}$$

Example 1.3.5 Figure 1.3 shows the cumulative relative frequency plots $\tilde{F}(y)$ for (1) the sample of female heights, and (b) the sample of male heights in Example 1.3.1.

A cumulative frequency plot does not show the "shape" of the distribution of y-values in a data set quite as clearly as a histogram. However, it shows us the proportion of y-values in any given interval;

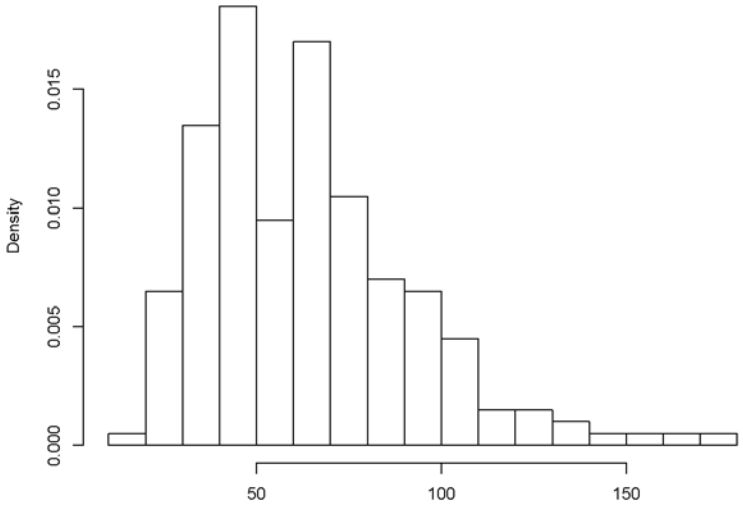


Figure 1.2: **Lifetimes (in km driven) for Front Brake Pads on 200 Cars**

the proportion in the interval $(a, b]$ is just $\tilde{F}(b) - \tilde{F}(a)$. In addition, this plot allows us to pinpoint values such as the *median* (a y-value m such that half of the data values are below m and half are above m) or the *100p'th percentile* (a y-value $Q(p)$ such that a proportion p of the data values is less that $Q(p)$), where $0 < p < 1$. For example, we see from Figure 1.3.3 that the median (or Q (.5)) height for females is about 1.60m (63.0 in) and for males, about 1.73m (68.1 in).

Other plots are also sometimes useful. The size n of a data set can be small or large. Histograms are not very useful when n is less than about 20-30, and for small samples we often just plot the locations of y-values on a line; an example is given in Section 1.7. A useful plot called the "strip-plot" for comparing two or more data sets is given next.

Box plots

Sometimes we have two or more samples of y-values, and we may wish to compare them. One way is by plotting histograms or relative frequency plots for the different samples on the same graph or page; we did this in Example 1.3.2 for the samples of female heights and male heights. The *box plot* is a plot in which only certain values based on a data set are shown, in particular the median, upper and lower quartiles (these are the 25th and 75th percentiles Q(.25) and Q(.75)), plus values equal to 1.5 times the "inter-quantile range" $Q(.75) - Q(.25)$ below $Q(.25)$ and above $Q(.75)$. Figure 1.4 shows such a plot for the female heights and male heights data sets from Example 1.3.2.

From the boxplot we can determine, for example, that approximately 75% of the females have heights less than 1.65 m. or that about 50% of males had heights between 1.7 and 1.79 m.

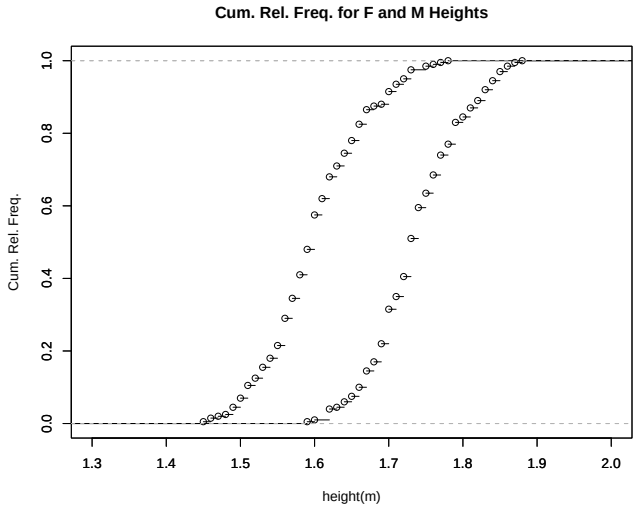


Figure 1.3: Cumulative relative frequency for Female (F) and Male (M) heights

Two-variable plots

Often we have data on two or more variables for each unit represented in a sample. For example, we might have the heights x and weights y for samples of individuals. The data set 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 i'th person in the sample.

When we have two such variables, a useful plot is a *scatter plot*, which is an $x - y$ plot of the points $(x_i, y_i),\ i = 1, \ldots, n$. This shows whether x_i and y_i tend to be related in some way. Figure 1.5 shows a scatterplot of heights x_i and weights y_i for 100 adult males. As is obvious from looking at people around us, taller people tend to weigh more, but there is considerable variability in weight across persons of the same height.

1.4 Probability Distributions and Statistical Models

Probability models are used to describe random processes. (For convenience we'll often use the single term "process" below but the terms population or phenomenon could also be inserted.) They help us understand such processes and to make decisions in the face of uncertainty. They are important in studies involving the collection and analysis of data for several reasons. These include:

(i) when studying a process scientifically, questions are often formulated in terms of *a model for the process*. The questions of primary interest do not concern the data, but the data provides a

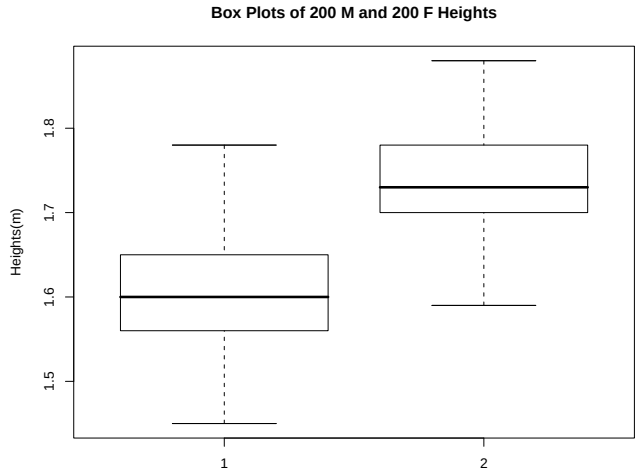


Figure 1.4: **Box Plots Based on 200 Female and 200 Male Heights in example 1.3.2. "F"=1, "M"=2.**

window to the population or the model.

(ii) the data collected in studying processes are variable, so random variables are often used in discussing and dealing with data,

(iii) studies of a process usually lead to inferences or decisions that involve some degree of uncertainty, and probability is used to quantify this,

(iv) procedures for making decisions are often formulated in terms of models,

(v) models allow us to characterize processes, and to simulate them via computer experiments or other means.

Consider a variable y associated with the units in a population or process. To describe or "model" the variability in y-values we use probability distributions, which were introduced in your first probability course. This is done as follows: let y be the value for a randomly chosen unit in the population or process. Because this value is random (we do not know which unit will be chosen) we call Y a *random variable*, and use a probability distribution to provide us with probabilities such as $P(a \leq Y \leq b)$. You should review your probability notes(a limited review is given in an appendix to this chapter) and recall that random variables are usually either *discrete* or *continuous*. A discrete random variable (r.v.) Y is one for which the range R (set of possible values) of Y is countable. A continuous r.v. is one whose range R consists of one or more continuous intervals of real numbers. For a discrete r.v. the *probability*

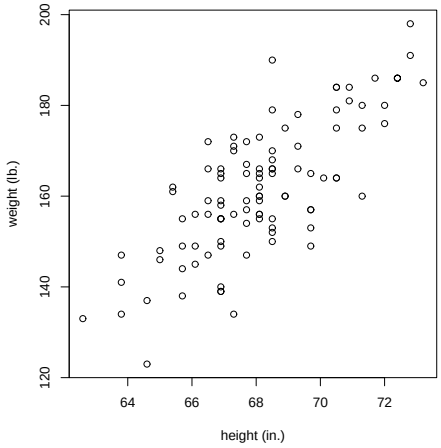


Figure 1.5: **Scatterplot of Height vs. Weight for 100 Adult Males**

function (p.f.) $f(y)$ is defined as

$$f(y) = P(Y = y) \quad \text{for certain values of } y \in R$$

where $R = \{r_1, r_2, r_3, \ldots\}$, a countable subset of $\mathbb{R}$, is the range of Y..For a continuous random variables. the *probability density function* (p.d.f) $f(y)$ is such that for any interval (a, b) contained in R,

$$P(a \leq Y \leq b) = \int_a^b f(y)dy.$$

Example 1.4.1 A Binomial Distribution

Consider a “toy" example in which a six-sided die is rolled repeatedly. This constitutes the target process, and for a study we might roll the die a total of n times. For this study, let Y denote the number of rolls that result in the number 6. We treat it as a random variable since it is subject to random variation. Assuming that the die has probability $1/6$ of turning up the face "six" on any single roll, and that the rolls are independent, the probability distribution of Y has probability function

$$P(Y = y) = f(y) = \binom{n}{y}(1/6)^y(5/6)^{n-y} \quad y = 0, 1, \ldots, n \tag{1.2}$$

This is called a *binomial distribution*. We should bear in mind that 1.2 is a model; for any real die the assumption that a six has probability $1/6$ could be slightly in error. However, 1.2 is a very accurate model that closely represents variability for most real dice.

Example 1.4.2 An Exponential Distribution

Suppose that in a population of light bulbs the random variable X represents the lifetime (say in days) of a randomly chosen bulb. The continuous exponential distribution provides a good model for many types of bulbs. For example, if the bulbs have an average lifetime of 100 days (2400 hours) operation, then a distribution with p.d.f.

$$f(x) = .01e^{-.01x} \qquad x > 0 \tag{1.3}$$

would be suitable. Using this, we can compute probabilities such as

$$P(X > 1000) = \int_{1000}^{\infty} .01e^{-.01x}dx = 0.368$$

Recall that the *cumulative distribution function* (c.d.f) is defined for a r.v. Y as

$$F(y) = P(Y \leq y) \qquad y\epsilon R \tag{1.4}$$

If Y is discrete then $F(y) = \sum_{x \leq y} f(x)$; if Y is continuous then

$$F(y) = \int_{x \leq y} f(x)dx$$

Recall also that if $h(Y)$ is some function of y, then the *expectation* (or "expected value") of $h(Y)$ is defined as

$$E[h(Y)] = \sum_{y\epsilon R} h(y)f(y) \tag{1.5}$$

if Y is discrete and as

$$E[h(Y)] = \int_{R} h(y)f(y)dy \tag{1.6}$$

if Y is continuous. Expectations are used in many settings, for example when costs, profits, or losses are associated with a random variable. The expectation $E(Y)$ is called the mean of Y and is often denoted by the Greek letter μ. The expectation $E[(Y - \mu)^2]$ is called the variance of Y and is often denoted either as $Var(Y)$ or with the Greek symbol σ^2. The square root $\sigma = \sqrt{Var(Y)}$ is called the standard deviation of Y, or $sd(Y)$.

Your previous course introduced several families of probability distributions along with processes or populations to which they are applied. Models such as the binomial, Poisson, exponential, normal (Gaussian), and multinomial will be reintroduced and used in this course. The first few problems at the end of the chapter provide a review of some models, and examples of where they are applied.

Many problems involve two or more random variables defined for any given unit. For example Y_1 could represent the height and Y_2 the weight of a randomly selected 30 year old male in some population. In general, we can think of a random variable $Y = (Y_1, Y_2, \ldots)$ as being a vector of length ≥ 1. This may make it necessary to consider multivariate probability distributions, which were introduced in your last course for discrete random variables.

In many statistical applications there is a primary variable y of interest, but there may be a number of other variables $x_1, x_2, \ldots$ that affect y, or are "related" to y in some way. In this case we often refer to y as the *"response" variable* and $x_1, x_2 \ldots$ as *"explanatory" variables* or *covariates*. Many studies are carried out for the purpose of determining how one or more explanatory variables are related to a response variable. For example, we might study how the number of insurance claims y for a driver is related to their sex, age, and type of car ($x-$ variables). One reason for studying explanatory variables is to search out cause and effect relationships. Another is that we can often use explanatory variables to improve decisions, predictions or "guesses" about a response variable. For example, insurance companies use explanatory variables in defining risk classes and determining life insurance premiums.

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. Two broad aspects of the analysis and interpretation of data may be distinguished. 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 certain features. (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 methods of doing this 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.

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 phenomenon to draw more general inferences about the process or phenomenon itself. In general, we try to use study data to draw inferences about some target population or process. This is a form of inductive inference, in which we reason from the specific (the observed data) 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. Two main types of problems will be discussed, loosely referred to as *estimation problems and hypothesis testing problems*. In the

former, the problem is to estimate some feature 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.

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. These questions can be addressed by collecting data on the populations in question.

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

Example 1.5.1 A smoking behaviour survey

Suppose that a random sample of 200 persons aged 14-20 was selected, as described in example 1.2.1. Let us focus only on the sex of each person in the sample, and whether or not they smoked. The data are nicely summarized in a two-way frequency table such as the following:

	No. of smokers	No. of non-smokers	Total
Female	32	66	98
Male	27	75	102
Total	59	141	200

If we wished to estimate, say, the proportion of females aged 14-20 in the population who smoke, we might simply use the sample proportion $P = 32/98 = .327$. However, we would also like some idea as to how close this estimate is likely to be to the actual proportion in the population. Note that if we selected a second sample of 200 persons, we would very likely find a different proportion of females who smoked. When we consider estimation problems later in the course, we will learn how to use a probability model to calculate the uncertainty for this kind of study. For now, let us merely note what kind of model seems appropriate.

Consider only the proportion of females who smoke, and suppose that we select n females at random. (This is not quite what was done in the above survey.) Then the number of women X who smoke is actually a random variable in the sense that before the data are collected, it is random. Suppose now that the population of females from which the sample is drawn is very large and that a proportion p of the population are smokers. Then the probability distribution of X is, to a very close approximation, binomial with probability function

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, ..., n$$

Knowing this will allow better estimation procedures to be developed later in the course.

Example 1.5.2 A soft drink bottle filler study

Recall example 1.2.2, and suppose that 26 ounce bottles are randomly selected from the output of each of two machines, one old and one new, over a period of one week. The bottles were selected from each machine's output at roughly the same times. Accurate measurements of the amounts of liquid in the bottles are as follows:

Old machine:									
27.8	28.9	26.8	27.4	28.0	27.4	27.1	28.0	26.6	25.6
24.8	27.1	25.7	27.9	25.3	26.0	27.3	27.4	25.7	26.9
27.3	25.2	25.6	27.0	26.2	27.3	24.8	27.1	26.7	26.8
26.6	26.6	28.6	27.0	26.6	27.3	25.9	27.6	27.6	28.3
28.0	26.4	25.4	26.7	27.8	27.4	27.3	26.9	26.9	26.9
New Machine:									
26.6	26.8	27.2	26.9	27.6	26.7	26.8	27.4	26.9	27.1
27.0	27.1	27.0	26.6	27.2	26.1	27.6	27.2	26.5	26.3
28.0	26.8	27.1	26.7	27.7	26.7	27.1	26.5	26.8	26.8
26.9	27.2	27.4	27.1	26.5	27.2	26.8	27.3	26.6	26.6
27.0	26.9	27.3	26.0	27.4	27.4	27.6	27.2	27.8	27.7

The amount of liquid X that goes into a bottle is a random variable, and a main objective of this study is to determine what the distribution of X looks like for the old machine and for the new machine, over the one week period of the study. For this to be really useful, we should check first that there are no "drifts" or time trends in the data. Figure 1.6 gives a plot of x_i vs. i (order of production) for each machine, and no trends are apparent. The random variable X is continuous and so we would like to use a continuous probability distribution as the model. It often turns out for problems involving weights or measures that a normal distribution provides a suitable model. Recall that the distribution $N(\mu, \sigma^2)$ (which we will often call in this course a *Gaussian distribution*, denoted $G(\mu, \sigma)$) has probability density function

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad -\infty < x < \infty$$

and that probabilities are obtained by integrating it; recall also that $\mu = E(X)$ and $\sigma^2 = \text{Var}(X)$ are the mean and variance of the distribution. (note: exp (a) is the same as e^a.)

Before trying to use any particular model for X, it is a good idea to "look" at the data. Figure **??** shows frequency *histograms* of the data from the old and new machines, respectively. It shows that (i) the distributions of X for the old and new machines each look like they might be well described

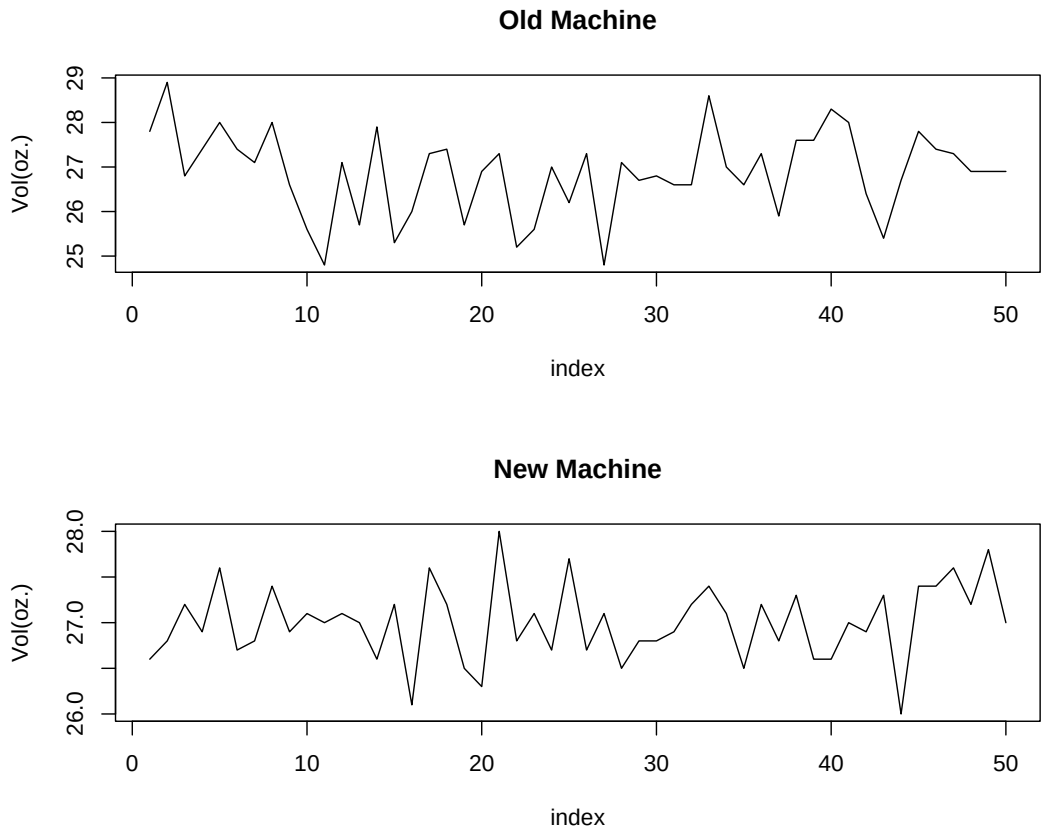


Figure 1.6: **Time Sequence Plot of Bottle Contents** y

by (different) normal distributions, (ii) the variability in the new machine's distribution is considerably less than in the old's.

After this simple bit of descriptive statistics we could carry out a more thorough analysis, for example fitting normal distributions to each machine. We can also estimate attributes of interest, such as the probability a bottle will receive less than 26 ounces of liquid, and recommend adjustments that could be made to the machines. In manufacturing processes it is important that the variability in the output is small, and so in this case the new machine is better than the old one.

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 exist many statistical software

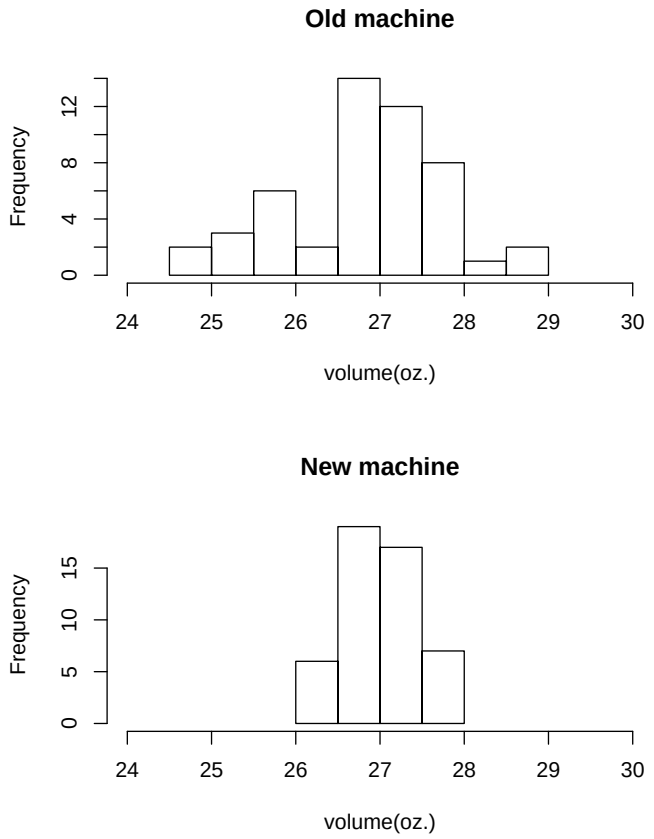


Figure 1.7: **Histograms of Data from Two Machines**

systems; some of the most comprehensive and popular are SAS, S-Plus, SPSS, Strata, Systat and R. Spreadsheet software is also useful.

In this course we will use the R software system. It is an open source package that has extensive statistical capabilities and very good graphics procedures. Its home page is at www.r-project.org. In structure it is similar to the commercial package S-Plus (Insightful Corp.); both are based on the ideas developed from the S statistical system at AT $\&$ T Bell Laboratories.

Some of the basics of R are described in the Appendix at the end of this chapter; it is very easy to use. In this course we will employ R for several purposes: to manipulate and graph data; to fit and check statistical models (distributions); to estimate quantities or test hypotheses; to simulate data from probability models.

As an introductory example we consider some data on the heights and the body-mass indexes (BMI's) of 150 males and 150 females, aged 18-60, that were collected from a random sample of

workers in New Zealand. The data are listed below, along with a few summary statistics. The BMI is often 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 types of guidelines refer to as “overweight", “underweight", etc. One that is sometimes used by public health professionals is:

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 on heights are stored in two R vectors (see the Appendix at the end of the chapter) called hmale and hfemale; the BMI measurement are in vectors bmimale and bmifemale.

```
Heights and Body-Mass Index (BMI) Measurements for 150 Males and Females

NOTE: BMI = weight(kg)/height(m)**2

MALE HEIGHTS (m)-  hmale

 [1] 1.76 1.76 1.68 1.72 1.73 1.78 1.78 1.86 1.77 1.72 1.72 1.77 1.77 1.70 1.72
[16] 1.77 1.79 1.75 1.74 1.71 1.73 1.74 1.70 1.71 1.72 1.66 1.74 1.73 1.77 1.69
[31] 1.91 1.77 1.81 1.74 1.87 1.76 1.69 1.87 1.78 1.70 1.78 1.84 1.82 1.77 1.72
[46] 1.80 1.72 1.69 1.78 1.69 1.80 1.82 1.65 1.56 1.64 1.60 1.82 1.73 1.62 1.77
[61] 1.81 1.73 1.74 1.75 1.73 1.71 1.63 1.72 1.74 1.75 1.72 1.83 1.77 1.74 1.66
[76] 1.93 1.81 1.73 1.68 1.71 1.69 1.74 1.74 1.79 1.68 1.71 1.74 1.82 1.68 1.78
[91] 1.79 1.77 1.74 1.78 1.86 1.80 1.74 1.69 1.85 1.71 1.79 1.74 1.80 1.64 1.82
[106] 1.66 1.56 1.80 1.68 1.73 1.78 1.69 1.57 1.64 1.67 1.74 1.89 1.77 1.75 1.84
[121] 1.66 1.71 1.75 1.75 1.64 1.73 1.79 1.74 1.83 1.80 1.74 1.81 1.80 1.66 1.75
[136] 1.82 1.80 1.81 1.71 1.59 1.71 1.79 1.80 1.70 1.77 1.78 1.64 1.70 1.86 1.75
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 1.56    1.71   1.74 1.744    1.79 1.93
FEMALE HEIGHTS (m)-  hfemale
 [1] 1.60 1.56 1.61 1.64 1.65 1.58 1.71 1.72 1.72 1.61 1.72 1.52 1.47 1.61 1.64
[16] 1.60 1.67 1.76 1.57 1.60 1.59 1.61 1.59 1.61 1.56 1.68 1.61 1.63 1.58 1.68
```

[31] 1.51 1.64 1.52 1.59 1.62 1.64 1.65 1.64 1.67 1.56 1.77 1.55 1.71 1.71 1.54
[46] 1.60 1.67 1.58 1.53 1.64 1.63 1.60 1.64 1.67 1.54 1.65 1.57 1.59 1.58 1.58
[61] 1.67 1.53 1.69 1.64 1.54 1.66 1.71 1.58 1.60 1.52 1.41 1.51 1.56 1.65 1.68
[76] 1.55 1.60 1.57 1.73 1.58 1.53 1.58 1.53 1.66 1.57 1.54 1.69 1.62 1.65 1.64
[91] 1.61 1.67 1.64 1.57 1.70 1.66 1.61 1.62 1.58 1.67 1.67 1.69 1.53 1.70 1.65
[106] 1.56 1.79 1.70 1.61 1.56 1.65 1.59 1.62 1.71 1.57 1.72 1.58 1.70 1.70 1.66
[121] 1.60 1.54 1.60 1.68 1.68 1.67 1.57 1.61 1.64 1.57 1.72 1.48 1.60 1.66 1.60
[136] 1.58 1.65 1.59 1.57 1.53 1.60 1.64 1.57 1.59 1.68 1.61 1.66 1.52 1.67 1.65

```
Min. 1st Qu. Median  Mean 3rd Qu. Max.
1.41   1.573   1.61 1.618   1.667 1.79
```

MALE BMI- bmimale

[1] 20.6 30.3 28.5 18.9 37.5 29.1 27.7 26.1 27.9 34.7 26.8 28.9 25.6 23.7 30.0
[16] 28.6 27.5 30.4 22.7 24.7 26.2 28.0 35.5 22.7 26.5 26.4 30.2 24.7 24.8 25.9
[31] 24.3 25.7 21.7 24.9 30.1 29.3 23.6 27.0 33.6 29.0 26.4 28.0 25.6 31.0 27.7
[46] 23.1 25.4 24.9 29.7 24.5 28.5 25.1 32.7 27.5 25.1 24.0 26.0 30.2 27.0 26.3
[61] 29.7 21.7 26.7 26.3 34.2 23.5 26.0 26.5 26.4 22.8 22.3 22.5 23.7 27.4 31.0
[76] 28.7 27.2 25.1 25.1 27.9 26.8 23.9 30.9 28.8 27.5 26.8 23.4 32.4 25.6 24.0
[91] 34.0 30.8 32.0 31.8 23.3 28.0 22.8 23.9 23.2 32.5 23.1 32.6 24.7 27.2 23.7
[106] 27.1 22.1 22.6 18.3 25.6 22.3 28.6 21.8 26.1 26.6 22.9 29.3 33.7 30.2 29.2
[121] 33.5 26.2 26.7 27.7 26.5 25.5 27.9 30.1 34.9 28.7 29.1 27.8 34.1 24.2 27.9
[136] 27.8 25.5 25.6 24.1 23.8 30.1 23.5 27.5 27.1 25.1 28.2 35.2 32.4 30.7 21.3

```
Min. 1st Qu. Median  Mean 3rd Qu. Max.
18.3    24.7  26.75 27.08    29.1 37.5
```

FEMALE BMI- bmifemale

[1] 23.4 21.2 31.2 27.1 25.9 26.8 28.8 24.3 36.2 37.0 37.5 37.2 28.4 20.7 25.1
[16] 18.9 20.4 27.7 30.1 27.9 19.8 27.0 23.3 17.5 25.8 23.2 21.2 26.8 25.9 21.6
[31] 34.2 20.3 37.9 32.7 22.7 25.9 35.6 32.0 32.2 19.8 23.3 32.7 22.7 22.8 27.0
[46] 21.3 22.7 23.7 25.6 21.4 32.8 30.3 29.0 27.7 29.4 26.3 26.2 28.0 29.1 24.6
[61] 28.4 22.5 33.2 29.6 26.1 27.8 26.8 26.8 32.4 38.8 23.5 33.7 30.2 29.1 26.8
[76] 36.4 19.0 24.5 23.1 33.9 26.5 31.0 26.1 29.8 23.4 31.2 28.4 26.7 27.3 24.1
[91] 20.7 25.0 23.6 33.1 23.1 32.2 24.8 22.5 29.9 26.9 28.5 27.5 26.3 29.7 21.9
[106] 26.3 20.3 21.5 24.5 31.1 31.3 34.0 31.9 27.2 16.4 20.3 29.1 25.7 23.4 27.6

[121] 26.4 28.4 24.8 29.1 25.6 29.4 26.2 29.7 22.8 21.5 21.3 29.9 17.6 28.3 24.1
[136] 28.3 24.0 25.4 23.9 24.3 30.4 28.6 25.0 23.8 36.0 31.5 21.8 29.4 30.8 28.1

```
Min. 1st Qu. Median  Mean 3rd Qu. Max.
16.4   23.42   26.8 26.92    29.7 38.8
```

Methods for summarizing data were discussed in Section 1.4. Both numerical and graphical summaries can be obtained easily using R. For example, *mean* (x) and $var(x)$ produce the mean $\bar{x}$ and variance s_x^2 of a set of numbers $x_1, \ldots, x_n$ contained in the vector x. The definitions of $\bar{x}$ and s_x^2 are (see Section 1.3.2)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Using this we can find that the mean (average) heights for the 150 males in the sample is $\bar{x} = 1.74$ m (68.5 in.) and for the 150 females is $\bar{x} = 1.62$ m (63.8 in.).

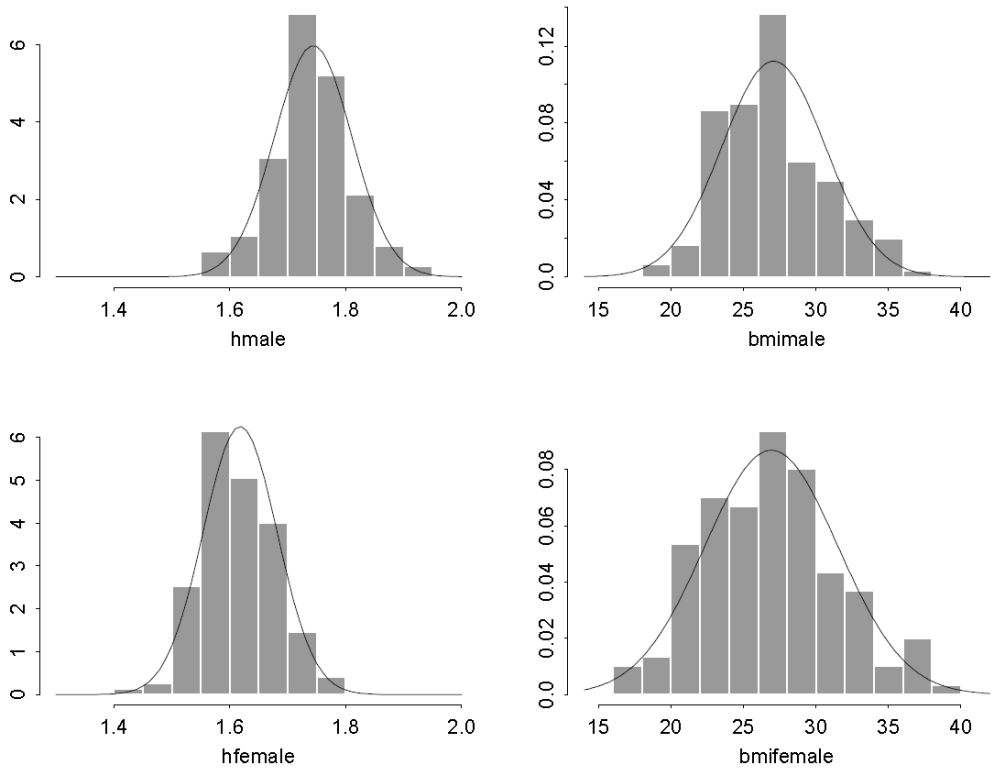


Figure 1.8: **Histograms and Models for Height and BMI data**

A histogram gives a picture of the data. Figure 1.8 shows relative frequency histograms for heights and BMI's for males and females. We also show normal distribution probability density functions

overlaid on each histogram. In each case we used a normal distribution $N(\mu, \sigma^2)$ where the mean μ and variance σ^2 were taken to equal $\bar{x}$ and s_x^2. For example, for the male heights we used $\mu = 1.74$ and $\sigma^2 = .004316$. Note from Figure 1.8 that the normal (Gaussian) distributions agree only moderately well with the observed data. Chapter 2 discusses probability models and comparisons of models and of data in more detail.

The following R code, reproduced from the Appendix, that illustrates how to look at the data and produce plots like those in Figure 1.6.1.

```
EXAMPLE: BODY-MASS INDEX DATA
The R session below describes how to take data on the BMI measurements
for 150 males and 150 females and examine them, including the possibility
of fitting Gaussian distributions to the data.

The data are in vectors bmimale and bmifemale.

> summary(bmimale)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 18.3    24.7  26.75 27.08    29.1 37.5
> summary(bmifemale)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 16.4   23.42   26.8 26.92    29.7 38.8
```

> sort(bmimale) #Sometimes its nice to look at the ordered sample
 [1] 18.3 18.9 20.6 21.3 21.7 21.7 21.8 22.1 22.3 22.3 22.5 22.6 22.7 22.7 22.8
[16] 22.8 22.9 23.1 23.1 23.2 23.3 23.4 23.5 23.5 23.6 23.7 23.7 23.7 23.8 23.9
[31] 23.9 24.0 24.0 24.1 24.2 24.3 24.5 24.7 24.7 24.7 24.8 24.9 24.9 25.1 25.1
[46] 25.1 25.1 25.1 25.4 25.5 25.5 25.6 25.6 25.6 25.6 25.6 25.7 25.9 26.0 26.0
[61] 26.1 26.1 26.2 26.2 26.3 26.3 26.4 26.4 26.4 26.5 26.5 26.5 26.6 26.7 26.7
[76] 26.8 26.8 26.8 27.0 27.0 27.1 27.1 27.2 27.2 27.4 27.5 27.5 27.5 27.5 27.7
[91] 27.7 27.7 27.8 27.8 27.9 27.9 27.9 27.9 28.0 28.0 28.0 28.2 28.5 28.5 28.6
[106] 28.6 28.7 28.7 28.8 28.9 29.0 29.1 29.1 29.2 29.3 29.3 29.7 29.7 30.0 30.1
[121] 30.1 30.1 30.2 30.2 30.2 30.3 30.4 30.7 30.8 30.9 31.0 31.0 31.8 32.0 32.4
[136] 32.4 32.5 32.6 32.7 33.5 33.6 33.7 34.0 34.1 34.2 34.7 34.9 35.2 35.5 37.5

```
> sqrt(var(bmimale))    #Get the sample standard deviations
[1] 3.555644
> sqrt(var(bmifemale))
```

```
[1] 4.602213
> par(mfrow=c(1,2))   #Sets up graphics to do two side by side plots per page
> hist(bmimale,prob=T,xlim=c(15,40))   #Relative frequency histogram; the
                                        xlim option specifies the range we want
                                        for the x-axis.
> x<- seq(15,40,.01)   #We'll use this vector to plot a Gaussian pdf
> fx<- dnorm(x,27.08,3.56)    #Computes values f(x) of the G(27.08,3.56) pdf; we
                              have estimated the distribution mean and standard
                              deviation from the sample values.
> lines(x,fx)    #This function adds points (x,fx) to the latest plot created
                   and joins them up with lines. This creates a plot of the
                   pdf overlaid on the histogram.
> hist(bmifemale,prob=T,xlim=c(15,40))  #Now do a histogram for the female
                                         data.
> fx<- dnorm(x,26.92,4.60)   #Compute pdf f(x) for G(26.92,4.60) distribution
> lines(x,fx)    # As previously
> q()   #Quit the R session.
```

1.7 A More Detailed Example: Colour Classification by Robots

Inexpensive robots and other systems sometimes use a crude light sensor to identify colour-coded items. In one particular application, items were one of five colours: White, Black, Green, Light Blue, Red. The sensor determines a light intensity measurement y from any given item and uses it to identify the colour.

In order to program the robot to do a good job, experiments are conducted on the sensor, as follows: items of different colours are passed by the sensor and the intensity readings y are recorded. Table 1 shows some typical data for 10 Red and 10 White items. Note that all Red items (or all White items) do not give the same y-values. The reasons for the variability include variation in the colour and texture of the items, variations in lighting and other ambient conditions, and variations in the angle at which the item passes the sensor.

Table 1. Light intensity measurements y for 10 Red and 10 White items

Red	47.6	47.2	46.6	46.8	47.8	46.8	46.3	46.5	47.6	48.8
White	49.2	50.1	48.8	50.6	51.3	49.6	49.3	50.8	48.6	49.8

Figure 1.9 shows a plot (called a “strip" plot) of similar data on 20 items of each of the five colours.

It is clear that the measurements for the Black items are well separated from the rest, but that there is some overlap in the ranges of the intensities for some pairs of items.

To program the robot to “recognize" colour we must partition the range for y into five regions, one corresponding to each colour. There are various ways to do this, the simplest being to choose values that minimize the number of misclassifications (incorrectly identified colours) in the data that are available. Another approach is to model the variability in y-values for each colour using a random variable Y with a probability distribution and then to use this to select the partition. This turns out to have certain advantages, and we will consider how this can be done.

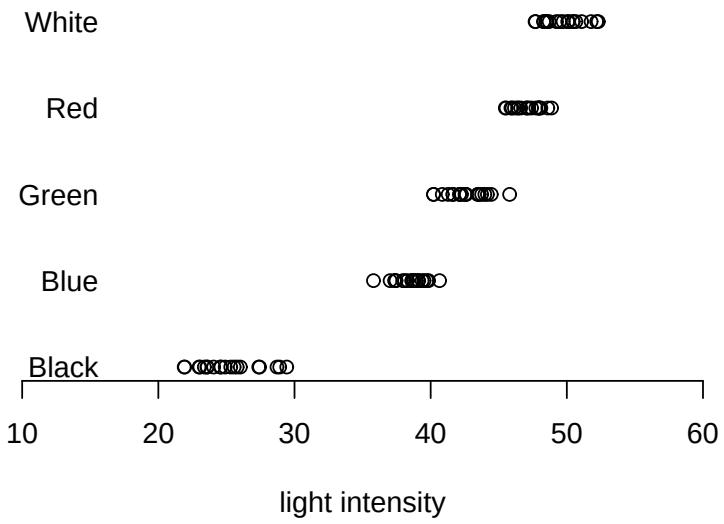


Figure 1.9: **Light Intensities for 20 Items of Each Colour**

Empirical study has shown that for the population of White items Y has close to a Gaussian distribution,

$$Y \sim G(\mu_W, \sigma_W).$$

Similarly, for Black, Green, Light Blue and Red items the distribution of Y is close to $G(\mu_B, \sigma_B)$, $G(\mu_G, \sigma_G)$, $G(\mu_L, \sigma_L)$ and $G(\mu_R, \sigma_R)$, respectively. The approximate values of μ and σ for each colour are (we will discuss in Chapter 2 how to find such values)

Black	$\mu = 25.7$	$\sigma = 2.4$
Light Blue	$\mu = 38.4$	$\sigma = 1.0$
Green	$\mu = 42.7$	$\sigma = 1.3$
Red	$\mu = 47.4$	$\sigma = 1.1$
White	$\mu = 49.8$	$\sigma = 1.2$

The operators of the equipment have set the following decision rule (partition) for identifying the colour of an item, based on the observed value y:

Black	$y \leq 34.0$
Light Blue	$34.0 < y \leq 40.5$
Green	$40.5 < y \leq 45.2$
Red	$45.2 < y \leq 48.6$
White	$y > 48.6$

We now consider a few questions that shed light on this procedure.

Question: Based on the Gaussian models, what are the probabilities an item colour is misclassified?

- This can be determined for each colour. For example,
  P(Black item is misclassified) = $P(Y > 34.0)$, where $Y \sim G(25.7, 2.4)$
  = .0003.

  P(Red item is misclassified) = 1 - $P(45.2 < Y \leq 48.6)$, where $Y \sim G(47.4, 1.1)$
  = .1604.

- Note that colours may be classified incorrectly in more than one way. For example
  P(Red is misclassified as Green) $= P(40.5 < Y \leq 45.2) = .0227$
  P(Red is misclassified as White) $= P(Y > 48.6) = .1377$.

Question: What kind of data would we collect to check that the Gaussian models are satisfactory approximations to the distributions of Y?

- We would need to randomly select items of a specific colour, then use the sensor to get an intensity measurement y. By doing this for a large number n of items, we would get measurements $y_1, ..., y_n$ which can be used to examine or fit parametric models for Y, using methods developed in later chapters.

Question: Do we need to use a Gaussian model (or any other probability distribution) for this problem?

- No. We could use a completely "empirical" approach in which we used the sensor experimentally with many items of different colours, then determined a "good" decision rule for identifying colour from the observed data. (For example, we could, as mentioned above, do this so that the total number of items misclassified was minimized.)
- To see how this would work, use R to simulate, say 20, y values for each colour, then try to pick the cut-points for your decision rule.

Question: What are some advantages of using a probability model (assuming it fits the data)?

- It allows decision rules to be obtained and compared mathematically (or numerically).
- In more complicated problems (e.g. with more types of items or with multivariate measurements) a direct empirical approach may be difficult to implement.
- Models allow comparisons to be made easily across similar types of applications. (For example, sensors of similar types, used in similar settings.)
- Models are associated with "scientific" descriptions of measurement devices and processes.

Figure 1.10 shows the probability density functions for the Gaussian distributions for Black, Light Blue, Green, Red and White items, which provides a clear picture of misclassification probabilities.

Given the models above, it is possible to determine an "optimal" partition of the y-scale, according to some criterion. The criterion that is often used is called the overall *misclassification rate*, and it is defined as follows for this problem. Suppose that among all of the items which the robot encounters over some period of time, that the fractions which are Black, Light Blue, Green, Red, and White are P_B, P_{LB}, P_G, P_R and P_W respectively (with $P_B + P_{LB} + P_G + P_T + P_W = 1$). Suppose also that instead of the values above we consider arbitrary cut-points $c_1 < c_2 < c_3 < c_4$, so that if $y \leq c_1$, the decision is "Black", if $c_1 < y \leq c_2$, the decision is "Light Blue", and so on. The overall probability a randomly selected item is classified *correctly* (CC) is

$$P(\text{CC}) = P(CC|\text{Black})P(\text{Black}) + P(CC|\text{Blue})P(\text{Blue}) + P(CC|\text{Green})P(\text{Green}) + P(CC|\text{Red})P(\text{Red}) + P(CC| \tag{1.7}$$

$$= P(Y_B \leq c_1)P_B + P(c_1 < Y_{LB} \leq c_2)P_{LB} + P(c_2 < Y_G \leq c_3)P_G + P(c_3 < Y_R \leq c_4)P_R + P(Y_W > c_4)P_W$$

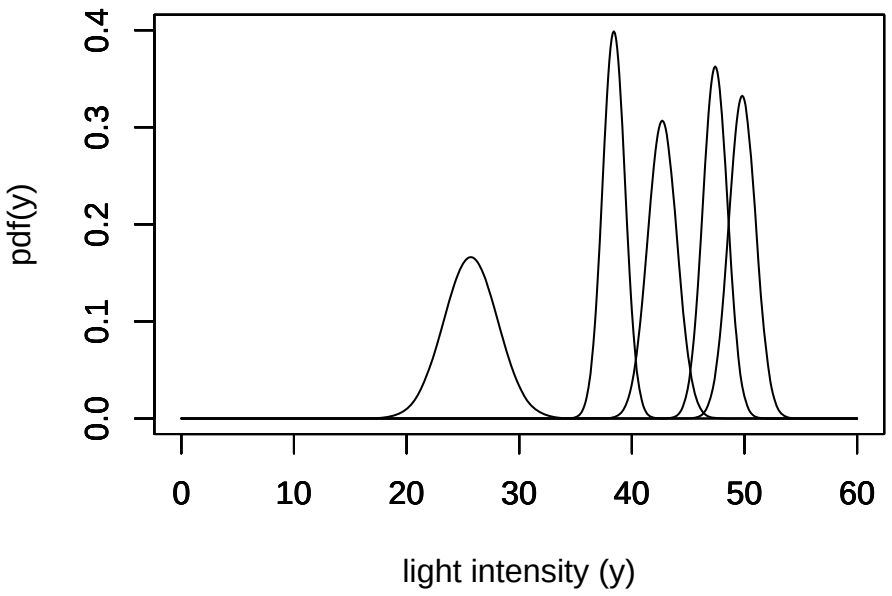


Figure 1.10: **Distributions of Light Intensities for Items of Colours**

where $Y_B, Y_{LB}, Y_G, Y_R, Y_W$ denote random variables with the $G(25.7, 2.4), G(38.4, 1.0), G(42.7, 1.3), G(47.4, 1.1), G(49.8, 1$ distributions respectively. It is not trivial to choose c_1, c_2, c_3, c_4 to maximize this (and therefore minimize the probability of incorrect classification) but it can be done numerically, for any set of values for$P_B, P_{LB}, P_G, P_R, P_W$. Note that for given values for c_1, c_2, c_3, c_4 we can readily calculate $P(\text{CC})$. For example, if $P_B = P_{LB} = .05; P_G = P_R = P_W = .3$, then with the values $c_1 = 34.0, c_2 = 40.5, c_3 = 45.2, c_4 = 48.6$ used above we get, using R to calculate the probabilities in 1.7,

$$\begin{aligned}
P(\text{CC}) &= .05[\text{pnorm}(34.0, 25.7, 2.4)] + .05[\text{pnorm}(40.5, 38.4, 1.0) - \text{pnorm}(34.0, 38.4, 1.0)] \\
&\quad + .3[\text{pnorm}(45.2, 42.7, 1.3) - \text{pnorm}(40.5, 42.7, 1.3)] \\
&\quad + .3[\text{pnorm}(48.6, 47.4, 1.1) - \text{pnorm}(45.2, 47.4, 1.1)] + .3[1 - \text{pnorm}(48.6, 49.8, 1.2)] \\
&= .05(.9997) + .05(.9821) + .3(.9723) + .3(.8396) + .3(.8413) \\
&= .895
\end{aligned}$$

Thus the probability of incorrect identification of a colour is .105. Note that if the “mix" of colours

that the robot sees changes (i.e. the values $P_B, \ldots, P_W$ change) then $P(\text{CC})$ changes. For example, if there were more Black and fewer White items in the mix, the $P(\text{CC})$ would go up. Problems 3 and 4 at the end of the chapter consider slightly simpler classification problems involving only two types of "items". It is easier to maximize the $P(\text{CC})$ in these cases. You should note that the general problem of classification based on certain data is very common. Spam detection in your emailer, or credit checking at the bank, fraud detection in a financial institution, and even legal institutions such as courts are all examples where a classification takes place on the basis of noisy data.

1.8 Appendix. The R Language and Software

1.8.1 Some R Basics

R is a statistical software system that has excellent numerical, graphical and statistical capabilities. There are Unix and Windows versions. These notes are a very brief introduction to a few of the features of R. Web resources have much more information. You can also download a Unix or Windows version of R to your own computer. R is invoked on Math Unix machines by typing R. The R prompt is $>$. R objects include variables, functions, vectors, arrays, lists and other items. To see online documentation about something, we use the help function. For example, to see documentation on the function mean(), type

```
help(mean).
```

In some cases help.search() is also helpful. The assignment symbol is $<$-: for example,

```
      x<- 15   assigns the value 15 to variable x.
To quit an R session in unix, type  q()
```

1.8.2 Vectors

Vectors can consist of numbers or other symbols; we will consider only numbers here. Vectors are defined using c(): for example,

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

defines a vector of length 5 with the elements given. Vectors and other classes of objects possess certain attributes. For example, typing

```
length(x)
```

will give the length of the vector x. Vectors of length n are often a convenient way to store data values for n individuals or units in a sample. For example, if there are variates x and y associated with any given individual, we would define vectors for x and for y.

1.8.3 Arithmetic

The following R commands and responses should explain arithmetic operations.

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

1.8.4 Some Functions

Functions of many types exist in R. Many 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; thus x and y must be the same length.) 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 mean 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
> var(x)    # Computes the (sample) variance 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
```

1.8.5 Graphs

To open a graphics window in Unix, type x11(). Note that in R, a graphics window opens automatically when a graphical function is used. There are various plotting and graphical functions. Two useful ones are

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

hist(x)    # Creates a frequency histogram based on the values in
             the vector x. To get a relative frequency histogram
             (areas of rectangles sum to one) use hist(x,prob=T).
```

Graphs can be tailored with respect to axis labels, titles, numbers of plots to a page etc. Type help(plot), help(hist) or help(par) for some information.
To save/print a graph in R using UNIX, you generate the graph you would like to save/print in R using a graphing function like plot() and type:

```
dev.print(device,file="filename")
```

where device is the device you would like to save the graph to (i.e. x11) and filename is the name of the file that you would like the graph saved to. To look at a list of the different graphics devices you can save to, type

```
help(Devices).
```

To save/print a graph in R using Windows, you can do one of two things.
a) You can go to the File menu and save the graph using one of several formats (i.e. postscript, jpeg, etc.). It can then be printed. You may also copy the graph to the clipboard using one of the formats and then paste to an editor, such as MS Word. Note that the graph can be printed directly to a printer using this option as well.

b) You can right click on the graph. This gives you a choice of copying the graph and then pasting to an editor, such as MS Word, or saving the graph as a metafile or bitmap. You may also print directly to a printer using this option as well.

1.8.6 Distributions

There are functions which compute values of probability or probability density functions, cumulative distribution functions, and quantiles for various distributions. It is also possible to generate (pseudo) random samples from these distributions. Some examples follow for the Gaussian distribution. For other distribution information, type help(Poisson),

```
help(Binomial) etc.
```

```
> y<- rnorm(10,25,5)    # Generate 10 random values from the Gaussian
                          distribution G(25,5); this is the same as
                          the normal distribution N(25,25). The values
                          are stored 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
```

1.8.7 Reading Data from a file

You can read numerical data stored in a text file called (say) data into an R vector y by typing

```
y<- scan("data")
```

Type help(scan) to see more about the scan function.

1.8.8 Writing Data or information to a file.

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).

1.8.9

Example: Body-Mass index Data

The R session below describes how to take data on the BMI measurements for 150 males and 150 females and examine them, including the possibility of fitting Gaussian distributions to the data. The data are in vectors bmimale and bmifemale.

```
> summary(bmimale)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 18.3    24.7  26.75 27.08    29.1 37.5
> summary(bmifemale)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 16.4   23.42   26.8 26.92    29.7 38.8
```

```
> sort(bmimale)   #Sometimes its nice to look at the ordered sample
 [1] 18.3 18.9 20.6 21.3 21.7 21.7 21.8 22.1 22.3 22.3 22.5 22.6 22.7 22.7 22.8
[16] 22.8 22.9 23.1 23.1 23.2 23.3 23.4 23.5 23.5 23.6 23.7 23.7 23.7 23.8 23.9
[31] 23.9 24.0 24.0 24.1 24.2 24.3 24.5 24.7 24.7 24.7 24.8 24.9 24.9 25.1 25.1
[46] 25.1 25.1 25.1 25.4 25.5 25.5 25.6 25.6 25.6 25.6 25.6 25.7 25.9 26.0 26.0
[61] 26.1 26.1 26.2 26.2 26.3 26.3 26.4 26.4 26.4 26.5 26.5 26.5 26.6 26.7 26.7
[76] 26.8 26.8 26.8 27.0 27.0 27.1 27.1 27.2 27.2 27.4 27.5 27.5 27.5 27.5 27.7
```

```
 [91] 27.7 27.7 27.8 27.8 27.9 27.9 27.9 27.9 28.0 28.0 28.0 28.2 28.5 28.5 28.6
[106] 28.6 28.7 28.7 28.8 28.9 29.0 29.1 29.1 29.2 29.3 29.3 29.7 29.7 30.0 30.1
[121] 30.1 30.1 30.2 30.2 30.2 30.3 30.4 30.7 30.8 30.9 31.0 31.0 31.8 32.0 32.4
[136] 32.4 32.5 32.6 32.7 33.5 33.6 33.7 34.0 34.1 34.2 34.7 34.9 35.2 35.5 37.5
> sqrt(var(bmimale))    #Get the sample standard deviations
[1] 3.555644
> sqrt(var(bmifemale))
[1] 4.602213
> par(mfrow=c(1,2))   #Sets up graphics to do two side by side plots per page
> hist(bmimale,prob=T,xlim=c(15,40))   #Relative frequency histogram; the
                                        xlim option specifies the range we want
                                        for the x-axis.
> x<- seq(15,40,.01)   #We'll use this vector to plot a Gaussian pdf
> fx<- dnorm(x,27.08,3.56)    #Computes values f(x) of the G(27.08,3.56) pdf; we
                              have estimated the distribution mean and standard
                              deviation from the sample values.
> lines(x,fx)    #This function adds points (x,fx) to the latest plot created
                   and joins them up with lines. This creates a plot of the
                   pdf overlaid on the histogram.
> hist(bmifemale,prob=T,xlim=c(15,40))  #Now do a histogram for the female
                                         data.
> fx<- dnorm(x,26.92,4.60)   #Compute pdf f(x) for G(26.92,4.60) distribution
> lines(x,fx)    # As previously
> q()   #Quit the R session.
```

NOTE: You can see from the histograms and Gaussian pdf plots that the Gaussian distribution does not seem an especially good model for BMI variation. The Gaussian pdf's are symmetric whereas the distribution of BMI measurements looks somewhat asymmetric.

1.9 Problems

1. The **binomial distribution** is a discrete probability model with probability function of the form

$$f(y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y} \qquad y = 0, 1, ..., n$$

where $0 < p < 1$ and n is a positive integer. If Y is a random variable with probability function $f(y)$ we write $Y \sim \text{Bin}(n, p)$.

A woman who claims to have special guessing abilities is given a test, as follows: a deck which contains five cards with the numbers 1 to 5 is shuffled and a card drawn out of sight of the woman. The woman then guesses the card, the deck is reshuffled with the card replaced, and the procedure is repeated several times. Let Y represent the number of correct guesses by the woman.

(a) Suppose an experiment consists of 20 repetitions, or guesses. If someone guesses "randomly" each time, discuss why $Y \sim \text{Bin}(20, .2)$ would be an appropriate model.

(b) Suppose the woman guessed correctly 8 times in 20 repetitions. Calculate $P(Y \geq 8)$ if $Y \sim \text{Bin}(20, .2)$ and use the result to consider whether the woman might have a probability p of guessing correctly which is greater than .2.

(c) In a longer sequence of 100 repetitions over two days, the woman guessed correctly 32 times. Calculate $P(Y \geq 32)$ if $Y \sim \text{Bin}(100, .2)$; you can use a normal approximation if you wish. What do you conclude now?

2. The **exponential distribution** is a continuous probability model in which a random variable X has p.d.f.

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \qquad x \geq 0$$

where $\theta > 0$ is a parameter.

(a) Show that θ is the mean of X. Graph the p.d.f. of X.

(b) The exponential distribution is often found to be a suitable model for distributions of lifetimes. The 30 observations $X_1, ..., X_{30}$ below, for example, are the lifetimes (in days) of a random sample of a particular type of lightbulb, subjected to constant use:

23	261	87	7	120	14	62	47	225	71
246	21	42	20	5	12	120	11	3	14
71	11	14	11	16	90	1	16	52	95

The mean of these 30 numbers is $\overline{X} = 59.6$. It has been suggested that if an exponential model is suitable for representing the distribution of lifetimes in the population of lightbulbs from which they came, then θ in (1) should have a value of around 59.6. Why should this be so?

(c) For the exponential distribution (1) with $\theta = 59.6$, calculate

i) $p_1 = P(0 \leq X < 40)$

ii) $p_2 = P(40 \leq X < 100)$

iii) $p_3 = P(100 \leq X < 200)$

iv) $p_4 = P(X \geq 200)$

Compare the values $30p_1$, $30p_2$, $30p_3$, $30p_4$ with the actual number of observations in the four intervals $[0, 40)$, $[40, 100)$, $[100, 200)$, $[200, \infty)$, respectively. Why should these numbers agree fairly well if the exponential distribution is a suitable model?

(d) Use a graph created using R to compare the model (1) with the data observed.

3. The **normal** or **Gaussian distribution** is an important continuous probability model which describes the variation in many types of physiological measurements very well. Recall that $Y \sim G(\mu, \sigma)$ (or $Y \sim N(\mu, \sigma^2)$) means that Y has a normal distribution with mean μ and standard deviation σ (variance σ^2).

Let Y be a variate representing the systolic blood pressure of a randomly selected woman in a large population, grouped or stratified by age. Good models for Y for persons not taking any medication have been found to be:

ages 17-24 $\quad Y \sim G(118, 8)$

ages 45-54 $\quad Y \sim G(130, 9)$

(a) Plot the probability density functions for the two models on the same graph using R.

(b) For each age group find the probability a randomly selected woman has blood pressure over 140.

(c) Suppose you were given the blood pressures $Y_1, ..., Y_{25}$ for 25 women from one of the two age groups; you do not know which. Show that you could decide with near certainty which group they came from by considering the value of $\overline{Y} = (Y_1 + \cdots + Y_{25})/25$.

(d) Suppose you know that very close to 10% of the population of women are in each age group and that you want to devise a rule for "deciding" which age group a woman is in, based on knowing only her blood pressure Y. Good rules will be of the form: for some chosen value y_0, decide that

age is 17-24 $\quad$ iff $\quad Y \leq y_0$.

Assuming you are going to use this rule over and over again, find y_0 so that the fraction of decisions which are wrong is minimized.

4. A test for diabetes is based on glucose (sugar) levels Y in the blood, measured after fasting for a specified period of time. For healthy people the glucose levels have close to a Gaussian distribution with mean $\mu = 5.31\ mmol/L$ and standard deviation $\sigma = 0.58 mmol/L$. For untreated diabetics Y is (close to) Gaussian with $\mu = 11.74$ and $\sigma = 3.50$.
   A diagnostic test for diabetes can be based on a measurement Y for a given individual. Suppose that we use the cutoff value 6.5, and diagnose someone as diabetic (diagnosis to be confirmed by further tests) if $Y > 6.5$.

   (a) If a person is diabetic, what is the probability that they are diagnosed as such by the test?

   (b) What is the probability that a nondiabetic is incorrectly diagnosed as being diabetic?

   (c) The probability in part (a) is called the **sensitivity** of the test. We can increase it by choosing a cutoff value C which is less than 6.5. However, this will increase the probability of incorrectly diagnosing nondiabetics, as in part (b). Recompute the probabilities in (a) and (b) if $C = 6.0$.

   (d) What cutoff value C do we need to make the sensitivity 0.98?

   (Based on Exercise 6.6 in Wild and Seber, 1999)

5. **Normal (Gaussian) approximation for binomial probabilities**. If $Y \sim Bin(n, p)$ and n is large, the Gaussian distribution can be used to calculate approximate probabilities for Y. In particular,

$$P(Y \leq y) \doteq F\left(\frac{y - np}{\sqrt{np(1-p)}}\right) \qquad y = 0, 1, ..., n \tag{1.8}$$

   where $F(z)$ is the cumulative distribution function (cdf) for the $G(0, 1)$ distribution. A slightly better approximation is to replace y with $y + .5$ in (2).

   A mail campaign to sign up new customers for a particular credit card has in the recent past had a success rate of about $p = .012$ (i.e. about 1.2% of persons contacted sign up for the card).

   (a) If 140,000 persons are sent the offer to sign up, what is the probability at least 1600 new customers are obtained, if $p = .012$?

   (b) Suppose 2000 new customers were actually obtained. Would you conclude that this campaign was more successful than other recent ones? Explain your answer.

6. **Computer-generated (pseudo) random numbers**. Consider a procedure designed to generate sequences of digits Y_i $(i = 1, 2, ...)$ such that for each i

$$P(Y_i = y) = .1 \qquad y = 0, 1, ..., 9$$

and such that $Y_1, Y_2, ...$ are statistically independent.

   (a) Suggest two things that you might do in order to assess whether a given computer procedure satisfies the above conditions. (Assume that you can use the procedure to generate, say, 1000 digits and that you will base your assessment on these values.)

   (b) The next problem gives such a procedure*, but is theoretically a little complicated to work out its distribution mathematically. However, the distribution can be closely approximated by computer simulation. You can generate a sequence $Y_1, Y_2, \ldots Y_n$ of length n by the R command $y \leftarrow\ sample(0:9, n, replace = T)$

   Generate a sequence of 1000 digits and plot a histogram of the data. (* You can use it to check, for example, whether runs of odd and even digits are consistent with the randomness conditions.)

7. **A runs test for randomness**
   Consider a process which produces binary sequences of 0's and 1's. The process is supposed to have the property that for a sequence $Y_1, Y_2, ..., Y_n$ of arbitrary length n,

$$P(Y_i = 1) = p = 1 - P(Y_i = 0) \qquad i = 1, ..., n$$

   $Y_1, ..., Y_n$ are mutually independent.

   In probability this is called a **Bernoulli model**.

   One way to test whether a sequence "looks" like it could have come from a Bernoulli model is to consider the number of **runs** (maximal subsequences of 0's or 1's) in the sequence; let the random variable R denote the number of runs.

   (a) Consider a randomly generated sequence from the Bernoulli model and suppose there were

a 0's and b 1's $(a+b=n)$. Try to prove that, conditional on a and b,

$$P(R=r)=\begin{cases} \dfrac{2\binom{a-1}{k-1}\binom{b-1}{k-1}}{\binom{a+b}{a}} & r=2k \\ \dfrac{\binom{a-1}{k-1}\binom{b-1}{k}+\binom{a-1}{k}\binom{b-1}{k-1}}{\binom{a+b}{a}} & r=2k+1 \end{cases}$$

(b) Part (a) is challenging. An approximation to the distribution of R which is not conditional on a and b may be based on the Central Limit Theorem. Let $Y_1,...,Y_n$ be a Bernoulli sequence of 0's and 1's and define the random variables

$$U_1=1,\ \ U_i=\begin{cases} 0 & \text{if } Y_i=Y_{i-1} \\ 1 & \text{if } Y_i\neq Y_{i-1}. \end{cases}$$

Note that $R=\sum_{i=1}^{n}U_i$ (why?) and use this to find $E(R)$ and Var(R) in terms of n and p, where $p=P(Y_i=1)=1-P(Y_i=0)$.

(c) By the Central Limit Theorem, $Z=[R-E(R)]/\text{Var}(R)^{1/2}$ has a limiting $G(0,1)$ distribution as $n\to\infty$. Use this to find the approximate probability of 20 or fewer runs in a Bernoulli sequence with $n=100$, $p=.5$.

(d) Outline how you could approximate the distribution of R in part (b) by simulation. For the case $n=10, p=.5$ (i.e. sequences of length 10) generate 1000 sequences and obtain the R value for each one. Plot these in a histogram.

8. The data below show the lengths (in cm) of male and female coyotes captured in Nova Scotia.

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) Obtain relative frequency histograms of the data for the females and the males using R.

(b) Compute the sample mean $\bar{y}$ and standard deviation s_y for the female and male coyotes. Assuming $\mu = \bar{y}$ and $\sigma = s_y$, plot the pdf's for Gaussian distributions $G(\mu, \sigma)$ over top of the histograms for the females and males. (Note: if you have n measurements $y_1, \ldots, y_n$ then $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.) (Based on Table 2.3.2 in Wild and Seber 1999)

9. The data below come from a study of the completion times for a task experienced by computer users in a multiuser system. Each data point (x, y) is the result of an experiment in which x terminals connected to the same server all initiated the same task. The variable y is the average time per task.

x	y	x	y	x	y	x	y
40	9.9	40	11.9	50	15.1	65	18.6
50	17.8	10	5.5	30	13.3	65	19.8
60	18.4	30	11.0	65	21.8		
45	16.5	20	8.1	40	13.8		

(a) Make a scatterplot of the data.

(b) Which variable is the response and which is the explanatory variable?
(Based on Wild and Seber 1999, Table 3.1.2)

10. The R system contains a lot of interesting data sets. Here's how to look at a data set contained in an array *wtloss*; it show the weight on each day of a special diet for a very obese 48-year old male

who weighed 184.35 kg before starting the diet. The data set is in the MASS package which is part of R. First you need the command

> *library (MASS)*

Then the command

> *wtloss*

will display a data base called *wtloss* which has 52 rows and two columns. Column 1 gives the day of the diet and Column 2 the person's weight on that day. Obtain a scatterplot of weight vs. day by the command

> *plot (wtloss $Days, wtloss $Weight, xlab = "Day", ylab = "Weight")*

Does the weight appear to be a linear function of days on diet? Why would you expect that it would not be linear if a long time period is involved?

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 in some way. As described in Section 1.4, our interest here is in studying variability and uncertainty in populations and processes. 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 been reminded of material on random variables and probability distributions that is taught in earlier courses.

A difficult step for beginners in probability and statistics is the choice of a probability model in given situations. The choice of a model is usually driven by some combination of three factors:

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

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

3. Current data, against which models can be assessed.

In probability theory, there is a lot of emphasis on factor 1 above, and there are many “families" of probability distributions that describe certain types of situation. For example, binomial and Poisson distributions were derived as models for outcomes in repeated trials and for the random occurrence of events in time or space, respectively. The normal (Gaussian) distribution, on the other hand, is

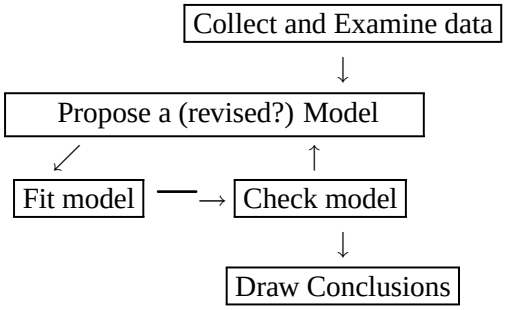
often used to represent the distributions of continuous measurements such as the heights or weights of individuals, but this is based more on past experience that such models are suitable than on factor 1 above.

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 one or more parameters θ:

$$P(Y = y) = f(y; \theta) \text{ for } y \epsilon R$$

where R 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 (let's call it $\hat{\theta}$) is often referred to as "fitting" the model or as "estimating" the value of θ. The next section gives a way to do this.

Most applications require a series 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 data 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 out other models, and perhaps look at more data, in order to work towards a satisfactory model. Doing this is usually an iterative process, which is sometimes represented by diagrams such as



Later courses in Statistics spend a lot of time on this process. In this course we will focus on settings in which the models are not too complicated, so that model formulation problems are minimized.

Before considering how to fit a model, let us review briefly some important families of distributions that were introduced in earlier courses.

Binomial Distribution

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

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

where θ is a parameter with $0 < \theta < 1$. This model arises in connection with repeated independent trials, where each trial results in either an outcome “S" (with probability θ) or “F" (with probability $1 - \theta$). If Y equals the number of S outcomes in a sequence of n trials, it has theprobability function 2.2. We write $Y \sim Bin(n, \theta)$ to indicate this.

Poisson Distribution

The discrete r.v. Y has a Poisson distribution if its p.f. is of the form

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

where θ is a parameter with $\theta > 0$. To indicate this we write $Y \sim$ Poisson (θ). It can be shown that $E(Y) = \theta$ for this model. The Poisson distribution arises in settings where Y represents the number of random events of some kind that occur over a fixed period of time, for example, the number of arrivals in a queue or the number of hits on a web site in a 1 hour period. For this model to be suitable the events must occur completely randomly in time. The Poisson distribution is also used to describe the random occurrence of events in space.

Exponential Distribution

The continuous r.v. Y has an exponential distribution if its p.d.f. is of the form

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

where θ is parameter with $\theta > 0$. To indicate this we write $Y \sim$ Exp(θ). It can be shown that $E(Y) = \theta$. The exponential distribution arises in some settings where Y represents the time until an event occurs.

Gaussian (normal) Distribution

The continuous r.v. Y has a Gaussian (also called a normal) distribution if its p.d.f. is of the form

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2} \qquad -\infty < y < \infty$$

where μ and σ are parameters, with $-\infty < \mu < \infty$ and $\sigma > 0$. It can be shown that $E(Y) = \mu, \text{Var}(Y) = \sigma^2, sd(Y) = \sigma$. We write either $Y \sim G(\mu, \sigma)$ or $Y \sim N(\mu, \sigma^2)$ to indicate that Y has this distribution. 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. The normal (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 basis for the most common model for asset prices, exchange

rates, interest rates, etc.

Multinomial Distribution

This is a multivariate distribution in which the discrete r.v.'s $Y_1, \ldots Y_K \ \ (K \geq 2)$ have the joint p.f.

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

where each $y_i (i = 1, \ldots, K)$ is an integer between 0 and n, and they satisfy 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 results in one of K types of outcomes (call them types $1, \ldots, K$), and the probability outcome i occurs is θ_i. If $Y_i \ (i = 1, \ldots, K)$ is the number of times that type i occurs in a sequence of n trials, then $(Y_1, \ldots, Y_K)$ have the joint distribution above. This is indicated by $(Y_1, \ldots, Y_K) \sim \text{Mult}(n; \boldsymbol{\theta})$.

Since $\sum_{i=1}^{K} Y_i = n$ we can if we wish rewrite $f(y_1, \ldots, y_K; \boldsymbol{\theta})$ using only $K - 1$ variables, say $y_1, \ldots, y_{K-1}$ (with y_K replaced by $y_1 - \ldots - y_{K-1}$). We see that the multinomial distribution with $K = 2$ is just the binomial distribution, where the S outcomes are type 1 (say) and the F outcomes are type 2.

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: 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 assigning numemrical values to the parameters in the model (for example μ, 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 that have been collected for the random variable in question. It is important that such data 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 were considering a Gaussian model, $Y \sim G(\mu, \sigma)$. Since $E(Y) = \mu$, we might decide to randomly select, say, 10 females from the population, to measure their weights $y_1, y_2, \ldots, y_{10}$, and then to estimate μ by the average,

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

This seems sensible (why?) and similar ideas can be developed for other parameters; in particular, note that σ must also be estimated, and think about how you might use $y_1, \ldots, y_{10}$ to do this. (Hint: what does σ or σ^2 represent in the Gaussian model?). Before we move on note that although we are estimating the parameter μ we did not write $\mu = \frac{1}{10} \sum_{i=1}^{10} y_i$! Why did we introduce a special notation $\hat{\mu}$. This serves a dual purpose, both to remind you that it is not exactly equal to the unknown value of the parameter μ, but also to indicate that it is a quantity derived from the data $y_i, i = 1, 2, ..10$ and is therefore random. A different draw of the sample $y_i, i = 1, 2, ..10$ will result in a different value for the random variable $\hat{\mu}$.

Instead of albeit sensible but nevertheless ad hoc approaches to estimation as in (2.3), it is desirable to have a general method for estimating parameters. *Maximum likelihood* is a very general method, which we now describe.

Let the (vector) random variable D represent potential data that will be used to estimate θ, and let d represent the actual observed data that are obtained in a specific application. Note that to write down $L(\theta)$, we must know (or make assumptions about) how the data d were collected. It is usually assumed here that the data consists of measurements on a random sample of population units. The *likelihood function* for θ is then defined as

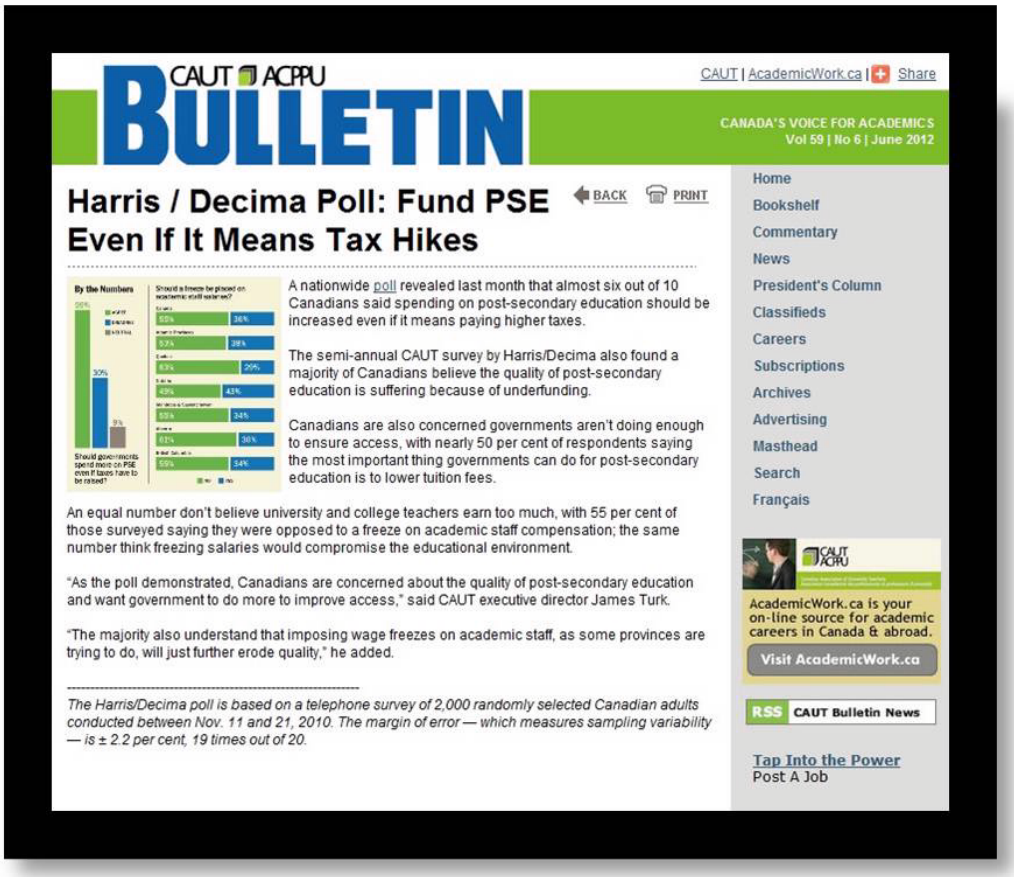
$$L(\theta) = P(D = d; \theta) \qquad \theta \in \Omega$$

where the *parameter space* Ω is the set of possible values for θ. Thus *the likelihood function is the probability that we will observe at random the observation d, considered as a function of the parameter.* Obviously values of the parameter that render our observation more probable would seem more credible than those that render is less probableso values of θ for which $L(\theta)$ is large are those that appear more consistent with the observation d. The value $\hat{\theta}$ that maximizes $L(\theta)$ for given data d is called

the *maximum likelihood estimate* (MLE) of θ. This seems like a “sensible" approach, and it turns out to have very good properties. Let us see how it works.

Example 2.2.1 (a public opinion poll).

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 in Figure **??** was conducted by Harris/Decima company under contract of the CAUT (Canadian Asso-



CAUT ACPPU BULLETIN

CAUT | AcademicWork.ca | Share

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

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

BACK PRINT

By the Numbers

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

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

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

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

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

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

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

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

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

ciation of University Teachers). This is one of semi-annual poll on Post-Secondary Education and Canadian Public Opinion, this one conducted 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".

Harris/Decima declared their result were accurate to within ± 2.2 percent 19 times out of twenty but the margin of error for regional, demographic or other subgroups is wider. What does this mean and where did these estimates and intervals come from? Suppose that the random variable Y represents

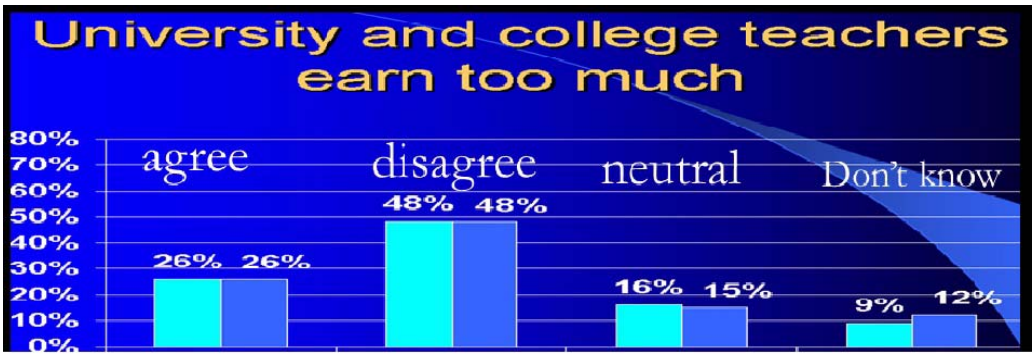


Figure 2.11: Results of the Harris/Decima poll

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:

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

where θ represents the fraction of the whole population that agree. In this case, if we select a random sample of n persons and obtain their view we have $D = Y$, and $d = y = 520$, the number that agree. Thus the likelihood function is given by

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \text{ or in this case}$$
$$\binom{2000}{520}\theta^{520}(1-\theta)^{2000-520}. \tag{2.2.1}$$

and it is easy to see that $L(\theta)$ is maximized by the value $\hat{\theta} = y/n$. (You should show this.) The value of this maximum likelihood estimate is $520/2000$ or 26%. This is easily seen from a graph of the likelihood function (2.2.1) seen in Figure 2.12 From the graph it is at least reasonable 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 this seems to contain most of the values of θ with large values of the likelihood $L(\theta)$. We will return to the constructino 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:

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

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 $D = Y$, and $d = y$ as the observed number

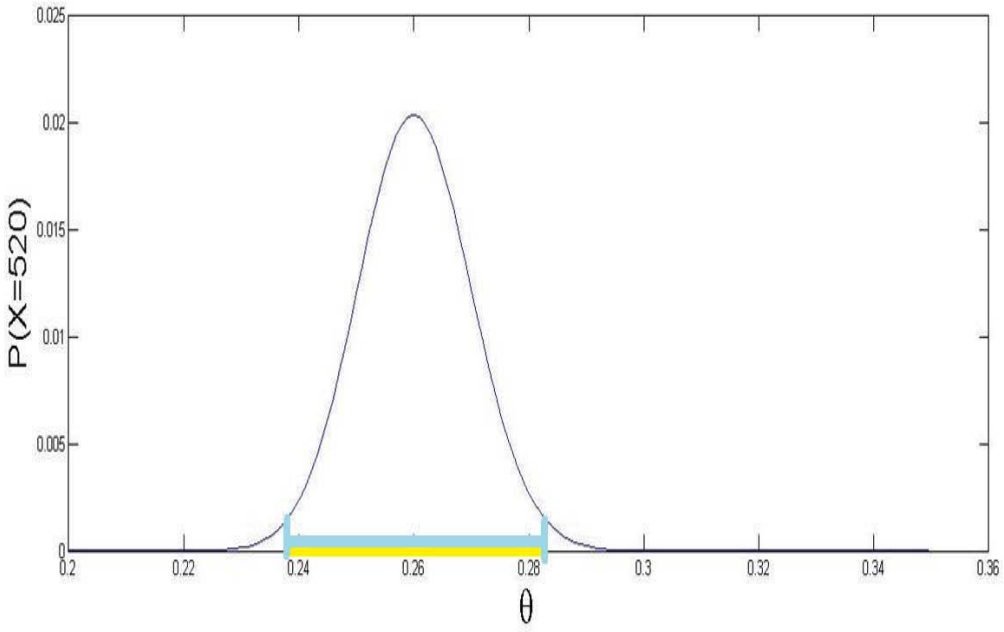


Figure 2.12: Likelihood function for the Harris/Decima poll and corresponding interval estimate for θ

infected. Thus

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \tag{2.2.1}$$

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

The likelihood function's basic properties, for example where its maximum is and its shape, is not affected if we multiply $L(\theta)$ by a constant. Indeed it is not the absolute value of the likelihood that is important but the relative values at two different values of the parameter, e.g. $L(\theta_1)/L(\theta_2)$ and these are also unaffected if we multiply $L(\theta)$ by a constant. In view of this we might define the likelihood as $P(D = d;\theta)$ or any constant multiple of it, so, for example, we could drop the term $\binom{n}{y}$ in (2.2.1) and define $L(\theta) = \theta^y(1-\theta)^{n-y}$. This function and (2.2.1) 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 the maximum so that the new maximum is 1. This rescaled version is called the **relative likelihood function**

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

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

$$\ell(\theta) = \log L(\theta) \qquad \theta \in \Omega.$$

Note that $\hat{\theta}$ also maximizes $\ell(\theta)$. (Why?) Because functions are often maximized by setting their

derivatives equal to zero[1], we can usually obtain $\hat{\theta}$ by solving the equation(s)

$$\frac{\partial \ell}{\partial \theta} = 0. \tag{2.2.2}$$

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

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

and solving $\partial \ell / \partial \theta = 0$ gives $\hat{\theta} = y/n$.

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

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta). \tag{2.2.3}$$

(You should recall from Stat 230 that if $Y_1, \ldots, Y_n$ are independent then their joint p.d.f. is the product of their individual p.d.f.'s) In addition, if we have independent data D_1 and D_2 about θ from two independent studies, then since $P(D_1 = d_1, D_2 = d_2) = P(D_1 = d_1)P(D_2 = d_2)$ with independency we can obtain the "combined" likelihood function $L(\theta)$ based on D_1 and D_2 together as

$$L(\theta) = L_1(\theta)L_2(\theta)$$

where $L_j(\theta) = P(D_j = d_j;\theta), j = 1, 2$.

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 p.d.f.

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

where $\theta > 0$. If a random sample of light bulbs is tested and gives lifetimes $y_1, ..., y_n$, then the likelihood function for θ is

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

Thus

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

[1]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 $f'(x) = 0$?

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

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

It is easily checked that this maximizes $\ell(\theta)$ and so it is the MLE.

Example 2.2.2 revisited. We can often write down a likelihood function in different ways. For the random sample of n persons who are tested for the HIV, for example, we could define for $i = 1, \ldots, n$

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

(Note: $I(A)$ is the indicator function; it equals 1 if A is true and 0 if A is false.) In this case the p.f. for Y_i is $Bin(1;\theta)$ with

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

and the likelihood function is

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

where $y = \sum_{i=1}^{n} y_i$. This is the same likelihood function as we obtained in Example 2.2.1, where we used the fact that $Y = \sum_{i=1}^{n} Y_i$ has a binomial distribution, $Bin(n,\theta)$.

Example 2.2.4 As an example involving more than one parameter, suppose that the r.v. Y has a normal (Gaussian) distribution with p.d.f.

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

(Note: exp (x) means the same as e^x.)
The random sample $y_1, ..., y_n$ then gives, with $\boldsymbol{\theta} = (\mu, \sigma)$,

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

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, σ. Solving (simultaneously)

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

we find that the MLE is $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})$, where

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

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 has close to a Poisson distribution:

$$P(Y = y) = f(y; \theta) = e^{-\theta v_i} \frac{(\theta v_i)^y}{y!}, \qquad y = 0, 1, 2, ... \tag{2.2.4}$$

where θ is the average number of bacteria per milliliter (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 get to observe is not 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 (2.2.4),

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

Suppose that n water samples, of volumes $v_1, ..., v_n$, are selected. Let $z_1, ..., z_n$ be the presence indicators. The likelihood function is then the product of $P(Z_i = z_i)$, or

$$L(\theta) = \prod_{i=1}^{n} (1 - e^{-\theta v_i})^{z_i} (e^{-\theta v_i})^{1 - z_i}$$

and

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

We cannot maximize $\ell(\theta)$ mathematically by solving $d\ell/d\theta = 0$, so we must resort to *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

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

Either maximizing $\ell(\theta)$ numerically for $\theta > 0$, or by solving $d\ell/d\theta = 0$ numerically, we find the MLE to be $\hat{\theta} = 0.478$. A simple way to maximize $\ell(\theta)$ is to plot it, as shown in Figure 2.13; the MLE can then be found by inspection or, more accurately, by iteration.

A few remarks about numerical methods are in order. Aside from a few simple models, it is not possible to maximize likelihood functions mathematically. However, there exists powerful numerical methods software 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. R and other packages are also invaluable for doing arithmetic, graphical presentations, and for manipulation of data.

2.3 Likelihood Functions From Multinomial Models

Multinomial models are used in many statistical applications. From Section 2.1, the multinomial probability function takes the form (using p_i for the probability of a type i outcome instead of θ_i)

$$f(y_1, \ldots, y_k) = \frac{n!}{y_1! \ldots y_k!} \prod_{i=1}^{k} p_i^{y_1} \qquad y_i = 0, 1, \ldots; \sum y_i = n$$

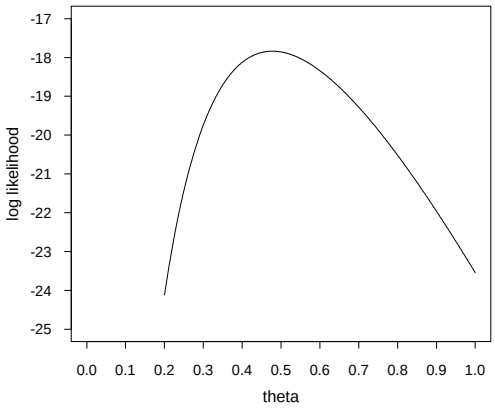


Figure 2.13: **The log likelihood function $\ell(\theta)$ for Example 2.2.**5

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

$$\hat{p}_i = y_i/n \qquad (i = 1, \ldots, k) \tag{2.3.1}$$

would be a sensible estimate. This can also be shown to be the MLE for $\boldsymbol{p} = (p_1, \ldots, p_k)$.[2]

[2]The log likelihood can be taken as (dropping the $n!/(y_1! \ldots y_k!)$ term for convenience) $\ell(\boldsymbol{p}) = \sum_{i=1}^{k} y_i \log p_i$. This is a little tricky to maximize because the p_i's satisfy a linear constraint, $\sum p_i = 1$. The theory of Lagrange multiplier methods for constrained optimization indicate that $\ell(\boldsymbol{p})$ can be maximized by solving the system of equations $\frac{\partial \ell^*(\boldsymbol{p},\lambda)}{\partial p_i} = 0 \qquad (i = 1, \ldots k)$, $\frac{\partial \ell^*}{\partial \lambda} = 0$ where

$$\ell^*(\boldsymbol{p}, \lambda) = \ell(\boldsymbol{p}) - \lambda(\sum_{i=1}^{k} p_i - 1)$$

Here, λ is called a Lagrange multiplier and it is easy to find that the solution is given by $\hat{\lambda} = n,\ \hat{p}_i = y_i/n\ (i = 1, \ldots, k)$.

Example 2.3.1 Each person is one of 4 blood types, labelled A,B,AB and O. (Which type a person is has important consequences, for example in determining who they can donate blood to for a transfusion.) Let p_1, p_2, p_3, p_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 $y_1 + y_2 + y_3 + y_4 = 400$).

Note that the random variables Y_1, Y_2, Y_3, Y_4 that represent the number of type A,B,AB,O persons we might get in a random sample of size $n = 400$ follow a multinomial distribution, Mult$(400; p_1, p_2, p_3, p_4)$. The MLE's from the observed data are therefore

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

(As a check, note that $\sum \hat{p}_i = 1$). These give estimates of the population fractions p_1, p_2, p_3, p_4. (Note: studies involving much larger numbers of people put the values of the p_i's for Caucasians at close to $p_1 = .448,\ p_2 = .083,\ p_3 = .034,\ p_4 = .436$.)

In some problems the multinomial parameters $p_1, \ldots, p_k$ may be functions of fewer than $k - 1$ parameters. The following is an example.

Example 2.3.2 Another way of classifying a person's blood is through their "M-N" type. Each person is one of 3 types, labelled MM,MN and NN and we can let p_1, p_2, p_3 be the fraction of the population that is each of the 3 types. According to a model in genetics, the p_i's can be expressed in terms of a single parameter θ for human populations:

$$p_1 = \theta^2,\ p_2 = 2\theta(1-\theta),\ p_3 = (1-\theta)^2$$

where θ is a parameter with $0 < \theta < 1$. In this case we would estimate θ from a random sample giving y_1, y_2 and y_3 persons of types MM, MN and NN by using the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{3} p_i^{y_i} \\ &= [\theta^2]^{y_1}[2\theta(1-\theta)]^{y_2}[(1-\theta)^2]^{y_3} \\ &= 2^{y_1}\theta^{2y_1+y_2}(1-\theta)^{y_2+2y_3} \end{aligned}$$

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 Y that represent measurement or variates in a population or process. A typical model has probability density function

(p.d.f) $f(y;\theta)$ if Y is continuous, or probability function (p.f.) $f(y;\theta)$ if Y is discrete, where θ is 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 $I_j = [a_{j-1}, a_j),\ \ j = 1, \ldots, k$. From our model $f(y;\theta)$ we can compute the values

$$\hat{p}_j = P(a_{j-1} \leq Y < a_j) \quad j = 1, \ldots, k. \tag{2.4.1}$$

If the model is suitable, these values should be "close" to the values of the relative frequencies $\tilde{p}_j = f_j/n$ 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 r.v's. An example of each type follows.

Example 2.4.1 Suppose that an exponential model for a positive-valued continuous r.v. Y has been proposed, with p.d.f.

$$f(y) = .01e^{-.01y}, \qquad y > 0 \tag{2.4.2}$$

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

$$\begin{array}{cccccccccc} 10, & 32, & 15, & 26, & 157, & 99, & 109, & 88, & 39, & 118, \\ 61, & 104, & 77, & 144, & 338, & 72, & 180, & 63, & 155, & 140 \end{array}$$

For illustration purposes, let us partition the range of Y into 4 intervals [0,30),[30,70),[70,140), [140,∞). The probabilities $\hat{p}_j$ from the model (2.4.2) are for $j = 1, \ldots, 4$,

$$\hat{p}_j = \int_{a_{j-1}}^{a_j} 0.1e^{-.01y}dy = e^{-.01a_{j-1}} - e^{-.01a_j}$$

and we find $\hat{p}_1 = .261, \hat{p}_2 = .244, \hat{p}_3 = .250, \hat{p}_4 = .247$, (the numbers add to 1.002 and not 1.0 because of round-off). The relative frequencies $\tilde{p}_j = f_j/20$ from the random sample are $\tilde{p}_1 = .15, \tilde{p}_2 = .25, \tilde{p}_3 = .30, \tilde{p}_4 = .30$. 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 $\tilde{p}_1$ 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. Problem 2 in Chapter 1 discusses an

application involving the exponential distribution where this is the case. 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 $p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$, 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 $\tilde{p}_1 = .17, \tilde{p}_2 = .46, \tilde{p}_3 = .37$, 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 can define the r.v. Y to be 1,2,3 according to whether a person is MM, MN or NN.) Since θ 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 θ from the observed data is of multinomial form:

$$L(\theta) = [\theta^2]^{17}[2\theta(1-\theta)]^{46}[(1-\theta)^2]^{37}$$

where $0 < \theta < 1$. Collecting terms, we find

$$\ell(\theta) = \log L(\theta) = 80 \log \theta + 120 \log (1-\theta) + \text{constant}$$

and $d\ell/d\theta = 0$ gives the MLE $\hat{\theta} = .40$. The model-based probabilities for I_1, I_2, I_3 are thus

$$\hat{p}_1 = \hat{\theta}^2 = .16, \hat{p}_2 = 2\hat{\theta}(1-\hat{\theta}) = .48, \ \hat{p}_3 = (1-\hat{\theta})^2 = .36.$$

These agree quite closely with $\tilde{p}_1, \tilde{p}_2, \tilde{p}_3$ and 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 $\tilde{p}_j = f_j/n$. 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 frequencies f_j from the sample as random variables, then they have a multinomial distribution, Mult $(n; p_1, \ldots, p_k)$, where p_j is the "true" value of $P(a_{j-1} \leq Y < a_j)$ in the population. In addition, any single f_j has a binomial distribution, $B\text{in}(n, p_j)$. This means we can assess how variable either f_j or $\tilde{p}_j = f_j/n$ is likely to be, in a random sample. From Stat 230, if n is large enough then the distribution of f_j is approximately normal, $N(np_j, np_j(1-p_j))$. It then follows that

$$P\left(np_j - 1.96\sqrt{np_j(1-p_j)} \leq f_j \leq np_j + 1.96\sqrt{np_j(1-p_j)}\right) = .95$$

and thus (dividing by n and rearranging)

$$P\left(-1.96\sqrt{\frac{p_j(1-p_j)}{n}} \leq \tilde{p}_j - p_j \leq 1.96\sqrt{\frac{p_j(1-p_j)}{n}}\right) = .95 \qquad (2.4.3)$$

This allows us to get a rough idea for what constitutes a large discrepancy between an observed relative frequency $\tilde{p}_j$ and a true probability p_j. For example when $n = 20$ and p_j is about .25, as in Example 2.4.1, we get from (2.5.3) that

$$P(-.19 \leq \tilde{p}_j - p_j \leq .19) = .95$$

That is, it is quite common for $\tilde{p}_j$ to differ from p_j by up to .19. The discrepancy between $\tilde{p}_1 = .15$ and $p_1 = .261$ in Example 2.4.1 is consequently not unusual and does not suggest the model is inadequate.

For larger sample sizes, $\tilde{p}_j$ will tend to be closer to the true value p_j. For example, with $n = 100$ and p_j =.5, (2.4.3) gives

$$P(-.10 \leq \tilde{p}_j - p_j \leq .10) = .95$$

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 (2.4.3) is concerned. This is not quite correct but (2.4.3) will still serve as a rough guide. We are also ignoring that we have picked the largest of the values $\tilde{p}_j = p_j$, 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 cumulative frequency functions $\tilde{F}(y)$ which are also called empirical c.d.f.'s, respectively.

The histogram plot for a continuous r.v. 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 p.d.f. $f(y; \theta)$ for the proposed model. The area under the p.d.f. 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 $[a_{j-1}, a_j)$, which equals $\tilde{p}_j$. The plots in Figure 1.6.1 for the height and body-mass index data in Section 1.6 are of this type.

For a discrete r.v. Y we plot a 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 cumulative frequency function or empirical c.d.f. (ECDF) $\tilde{F}(y)$ described by (1.3.1) in Section 1.3 and then to superimpose on this a plot of the model-based c.d.f., $F(y; \theta)$. If the model is suitable, the two curves should not be too far apart.

Example 2.4.3 Figure 2.14 shows a plot of the data on female heights from Section 1.6. We show (a) a relative frequency histogram, with the $G(1.62, 0.0637)$ p.d.f. superimposed (the MLE's were $\hat{\mu} = 1.62,\ \hat{\sigma} = 0.0637$, from the data), and (b) a plot of the ECDF with the $G(1.62, 0.0637)$ c.d.f. 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 - p.d.f. plot the data are grouped in forming the histogram. However, the histogram and p.d.f. show the distribution of heights more clearly. Neither plot suggests strongly that the Gaussian model is unsatisfactory. Note that the R function *ecdf* can be used to obtain $\tilde{F}(y)$.

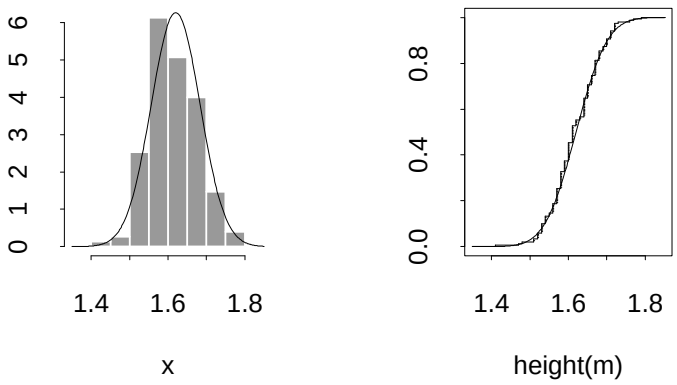


Figure 2.14: **Model and Data Comparisons for Female Heights**

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 λ per minute then the probability of x transactions in a time interval of length t minutes is

$$P(X = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x = 0, 1, 2, ...$$

   (a) The numbers of transactions received in 10 separate one minute intervals were as follows: 8, 3, 2, 4, 5, 3, 6, 5, 4, 1.
   Write down the likelihood function for λ and find the m.l.e. $\hat{\lambda}$.

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

   (c) Use R function *rpois()* with the value $\lambda = 4.1$ 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 $E(X)$ =Var$(X) = \lambda$ 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.

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

   ii) 100 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.
   The m.l.e. $\hat{\theta}$ is therefore the same in each case: what is it?

   (b) Suppose n people came to a blood donor clinic. Assuming $\theta = .10$, how large should n be to make the probability of getting 10 or more B- type donors at least .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 $x_1, x_2,$ and x_3 persons of types MM, MN, and NN respectively, find the MLE $\hat{\theta}$ in the model.

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 genders 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 m.l.e. $\hat{\alpha}$ of α. What does this give if $n = 50$ with $n_1 = 16$, $n_2 = 16$, $n_3 = 18$?

   (c) Does the model here 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 the scheme r animals are caught, tagged, and then released. Later on n animals are caught and the number X of these that bear tags are noted. The idea is to use this information to estimate N.

(a) Show that $P(X = x) = \begin{pmatrix} r \\ x \end{pmatrix} \begin{pmatrix} N-r \\ n-x \end{pmatrix} / \begin{pmatrix} N \\ n \end{pmatrix}$, under suitable assumptions.

(b) For observed r, n and x find the value $\hat{N}$ that maximizes the probability in part (a). Does this ever differ much from the intuitive estimate $\tilde{N} = rn/x$? (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 X in a family, for a large population of families:

$$P(X = k) = \alpha^k \qquad k = 1, 2, \ldots$$
$$P(X = 0) = (1 - 2\alpha)/(1 - \alpha)\ .$$

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

(a) Suppose that n families are selected at random and that f_k is the number of families with k children $(f_0 + f_1 + \ldots = n)$. Obtain the m.l.e. $\hat{\alpha}$.

(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 X represent the number of children in the family. Show that

$$P(X = k) = ck\alpha^k \qquad k = 1, 2, \ldots$$

and determine c.

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

k:	1	2	3	4
f_k:	22	7	3	1

Obtain the m.l.e. $\hat{\alpha}$ and a 10% likelihood interval. Also estimate the probability a couple has 0 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 λ per second. In n time periods of varying lengths $t_1, t_2, ..., t_n$ (seconds), the numbers of particles emitted (as determined by an automatic counter) were $y_1, y_2, ..., y_n$ respectively.

(a) Give an estimate of λ 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 λ based on these data, and describe how you could find the maximum likelihood estimate $\hat{\lambda}$.

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:

- r (where $0 \leq r \leq n$) pieces fail, at times $x_1, ..., x_r$.
- $n - r$ pieces are still working after time C.

(a) For the exponential model in Section 2.1, show that

$$P(X > C) = \exp(-C/\theta).$$

(b) Give the likelihood function for θ based on the observed data described above. Show that the m.l.e. is

$$\hat{\theta} = \frac{\sum_{i=1}^{r} x_i + (n-r)C}{r}.$$

(c) What does part (b) give when $r = 0$? 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 $r = 5$ failed before 400 hours, when the test was terminated, with $\sum_{i=1}^{5} x_i = 450$ hours. Find $\hat{\theta}$.

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

$$P(Y = y|x) = e^{-\lambda(x)}\frac{\lambda(x)^y}{y!} \qquad y = 0, 1, 2, ...$$

where $\lambda(x) = \exp(\alpha + \beta x)$, are useful.

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

   (b) Can $\hat{\alpha}$ and $\hat{\beta}$ be found in algebraic form?

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

   (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, write down the likelihood function for $\alpha,\ \beta$ based on the multinomial distribution, and maximize it to obtain $\hat{\alpha}$ and $\hat{\beta}$.

   (b) Compute expected frequencies for each of the four categories. 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 x 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) It is often found that the log lifetimes, $Y = log\ X$, are well modelled by a Gaussian distribution. Fit such a model to the data, and then produce a plot of a relative frequency histogram of the x-data with the p.d.f. for X superimposed on it, using $\hat{\mu}$ and $\hat{\sigma}$ for the values of μ and σ.
(Note: The p.d.f. of $X = \exp(Y)$ can be found from results in STAT 230, and is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2} \qquad x > 0.\Bigg)$$

(b) Suppose that instead of the model in part (a), you assumed that $X \sim G(\mu, \sigma)$. Repeat the procedure in part (a). Which model appears to fit the data better?

PLANNING AND CONDUCTING EMPIRICAL STUDIES

3.1 Empirical Studies

An empirical study is one which is carried out to learn about some real world population or process. Several examples have been given in the preceding two chapters, but we have not yet considered the various aspects of a study in any detail. 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 poor study can be worthless or even misleading.

It is helpful to break the process of planning and conducting a study into a series of parts or steps. We describe below a formulation with the acronym PPDAC, proposed by Jock Mackay and Wayne Oldford of University of Waterloo. Other similar formulations exist, and it should be remembered that their purpose is to focus on the essential aspects of an empirical study. Although they are presented as a series of steps, many studies, as well as the overall objective of learning about a population or process, involve repeated passes through one or more steps.

The steps or aspects of a study that PPDAC refers to are as follows:

- **P**roblem: a clear statement of the study's objectives
- **P**lan: the procedures used to carry out the study, including the collection of data.
- **D**ata: the physical collection of the data, as per the plan
- **A**nalysis: the analysis of the data collected
- **C**onclusion: conclusions that are drawn from the study

We will discuss the Plan, Data, Analysis and Conclusion steps in sections following this one. First, we consider some aspects of the Problem step.

The objectives of a study may be difficult to state precisely in some cases, because when trying to learn about a phenomenon we are venturing into the unknown. However, we must state as clearly as possible what we hope to learn, and (looking ahead) what the “output" or type of conclusions from our study will be. Here are some terms that describe certain problems or objectives:

- causative: this means that an objective is to study (possible) causal relationships among factors or variables. For example, we might want to study whether high amounts of fat in a person’s diet increase their risk of heart disease, or whether a drug decreases the risk of a person having a stroke.

- descriptive: this means that an objective is to describe the variability or other characteristics of certain variables, or to describe relationships among variables. (There is no attempt to consider causal connections.) For example, we may wish to estimate the percentage of persons who are unemployed this month in different regions of Canada, and to relate this to their level of education.

- analytic or inferential: this means that an objective is to generalize from the study to a larger context or process. This is what is known as inductive inference. Causative objectives are one type of analytic objective.

- technological: this refers to objectives involving predictions or decisions (e.g. develop a model or algorithm to identify someone from biometric data, or to decide whether to sell a stock)

A distinction can often be made between what we call the **target population (or process)** and the **study population (or process)**. The former is what we really want to study, and what our objectives are based on. The latter is what the units in the study are actually selected from. Sometimes the target and study populations are the same but often they are not; in that case we aim to make the study population as “similar" as possible to the target population. This will make generalizations from the study to the target population plausible.

At the Problem step we must also consider what the “units" of the population or process are, what variables will be used to study the units, and what **characteristics or attributes** of the variables we want to examine. This can be difficult, especially in areas where the amount of “hard" scientific background is limited, because it involves **quantification and measurement** issues. Some things can be quantified and measured easily and accurately (e.g. a person’s height, weight, age or sex) but many cannot. For example, how can we measure the amount of fat in a person’s diet? Public opinion polls or surveys are similarly tricky to design questions for, and we may not be sure what an answer represents. For example (see Utts 20012), in a survey the following two questions were asked in the order shown:

1. How happy are you with life in general?

2. How often do you normally go out on a date?

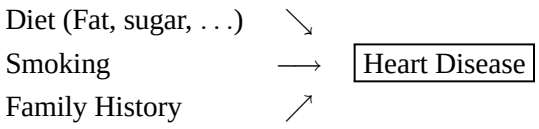
There was almost no relationship between the respondents' replies to the two questions. However, when the order of the questions was reversed there was a strong relationship, with persons going out frequently generally being happy. (Why might this occur?)

The units or individuals that are selected for a study are called a **sample**, because they usually are a subset of all the units in the study population. In so far as it is possible, we attempt to select a random sample of units from the study population, to ensure that the sample is "representative'. Drawing a random sample of units from a population or process can be difficult logistically. (Think about what this would involve for some of the studies discussed in chapters 1 and 2.) If our sample is not representative of the study or target population, conclusions we draw may be inaccurate or incorrect; this is often referred to as study **bias**. Also, ideally we have,

$$\text{Sample} \subset \text{Study Population} \subset \text{Target Population}$$

but in some cases the study population may not be a subset of the target population but only related to it in some way. For example, our target population for a carcinogenicity study might be humans, but experimental studies have to be conducted on laboratory mice or other animals. It is difficult to generalize from a study on animals to a human population.

The selection of explanatory variables for a study is also important. In many problems there is a large number of factors that might be related (in a causal way or not) to a response variable, and we have to decide which of these to collect data on. For example, in an observational study on heart disease we might include factors in the person's diet, their age, weight, family history of heart disease, smoking history and so on. Sometimes **cause and effect** diagrams are used to help us sort out factors and decide what to measure, e.g.



We'll conclude this section with a couple of examples that illustrate aspects of the Problem step.

Example 3.1.1. Blood thinners and risk of stroke
A study is to be undertaken to assess whether a certain drug that "thins" the blood reduces the risk of stroke. The main objective is thus causative. The target population is hard to specify precisely; let us

just say it consists of persons alive and at risk of stroke now and in the future. The study population will consist of individuals who can be “selected" to participate in the study; individual persons are the units in the populations. In most such studies these are persons who come to a specific group of doctors or medical facilities for care and treatment, and this may or may not be representative of the target population. To make the study generalizable at least to the study population, we would want the sample from the study population to be representative of it. We try to achieve this by drawing a random sample of persons, but in medical studies this is complicated by the fact that a person has to agree to participate if selected.

The choice of response variables depends on the study objectives but also partly on how long the study will be conducted. Some examples of response variables y are

(i) the number of strokes for a person over a specified “followup" period, say 3 years

(ii) the time until the first stroke occurs (if it does)

(iii) I(a stroke occurs within 3 years)

(iv) I(a stroke leading to major loss of function occurs within 3 years).

Once response variables have been defined, a specific objective is to compare attributes based on the distributions of these variables for persons who receive the drug and for persons who do not. In addition, to avoid study bias we randomly assign either the drug or a “placebo" (which cannot be distinguished from the drug) to each person. You can think of this being done by flipping a coin: if its Heads a person gets the drug, if Tails they get the placebo. As discussed earlier in Example 1.2.3, it is also best not to let the person (or their doctor) know which “treatment" they are receiving. Finally, one might want to also define and measure explanatory variables considered to be risk factors for stroke.

Example 3.1.2 Predicting success in university courses

Universities use certain variables to help predict how a student will do in their university program, and to help decide whom to accept for a program. One objective might be to maximize the percentage of students who pass first year, or who graduate from the program; another might be to have a formula to use in the admissions process. Empirical studies for this problem usually collect data on certain “explanatory" variables x (e.g. for Math students these might include high school course marks, contest marks, information (or a rating) for the person’s high school, and so on) for some collection of students (e.g. all students entering the university program over some period). Response variables y might include university course marks or averages, or whether a person graduates. Obviously such a study would involve collecting data over several years.

Note that for the admissions problem the target population is actually the population of students who apply to the university program, whereas the study population consists only of these persons who

get admitted to the program (and who accept). This mismatch between the target and study populations means that there is some doubt about how persons not admitted (but whose x-variables are known) would do in the program. For the problem of predicting a person's success once they are admitted, we can take the target population to be all students admitted (and accepting) over some time period. The study population is more "representative" in this case. However, for this (and for admissions decisions), there is still a discrepancy between study and target populations: data on past years' students must be used for a target population consisting of this (or a future) year's students.

3.2 Planning a Study

To plan a study we need to specify (i) the study population and its units, (ii) how (and how many) units will be "selected" for the study, and (iii) how response and explanatory variables will be measured. This assumes that at the Problem stage we have specified what variables are to be considered.

Recall from Chapter 1 that there are various types of studies. Three main types are:

(i) **Surveys or random samples** from a finite population of units. In this case there is a population of, say, N units (e.g. persons, widgets, transactions) and we randomly select $n(< N)$ for our study. Public opinion polls or surveys on health, income, and other matters are often of this type, as are random audits and quality inspections.

(ii) **Experimental studies**. In this case the person(s) conducting the study exercise control over certain factors, or explanatory variables. These studies are often hard or expensive to conduct, but are important for demonstrating causal relationships. Clinical trials involving medical treatments (e.g. does a drug lower the risk of stroke?) and engineering experiments (e.g. how is the strength of a steel bolt related to its diameter?) are examples of such studies.

(iii) **Observational studies**. In this case the data are collected on a study population or process over which there is no control. The units selected for the study are often selected in some random way, so as to make the selected units "representative". For example, we might collect data on a random set of transactions in an audit study, or on a random sample of persons buying a particular product in a marketing study. In medicine, we might compare the "risk profiles" of persons diagnosed with a disease (e.g. diabetes) with persons not diagnosed with the disease.

Two main questions with observational studies are whether the study population is representative of the target population, and whether the units chosen for the study are representative of the study population. It is sometimes impossible to generalize, or even to draw any conclusions, from observational data.

Selecting random samples from a population or process can be challenging due to logistical problems as noted above. In addition, there is the question whether a process is stable over time. This affects both the objectives of a study and the way units are sampled and measured. Entire books are devoted to ways of drawing random or representative samples in different settings.

In settings where the population or process has a finite number of units, N, we can draw a random sample by measuring the units (say as $1, 2, \ldots, N$) and then drawing n numbers (usually without replacement). This identifies the units selected for the sample. This can of course be hard or impossible to implement in many settings; consider for example how you could select a sample of n persons aged 18-25 and living in Canada on some specific date.

It is often taken for granted that measurements, or data, are “reliable". In fact, measurement of the variables in a study may be subject to errors, and this should be considered in planning and analyzing a study. If measurement errors are too severe, it may not even be worth doing the study. Consider the following example.

Example 3.2.1 Piston diameters

Pistons for car engines must have very precise dimensions, but they vary slightly in key dimensions such as diameter because of variability in the manufacturing process. Let Y denote the deviation of a piston diameter from its desired value, and suppose that $Y \sim G(0, \sigma)$. The value of σ determines the variability of the piston diameters; suppose that for the process to be acceptable it is necessary that $\sigma \leq 1$ (in some unspecified units). We can assess whether $\sigma \leq 1$ by randomly selecting some pistons and measuring their y-values. However, suppose y is measured using a crude tool so that what we observe is not y, but $y^* = y + e$, where e is the measurement error. If e is independent of y and $Var(e) = \tau^2$, then $Var(Y^*) = Var(Y) + Var(e) = \sigma^2 + \tau^2$ and $sd(Y^*) = (\sigma^2 + \tau^2)^{1/2}$. If τ is large enough (e.g. suppose if $\tau = 1$) then $Var(Y^*) = (\sigma^2 + 1)^{1/2}$, so even if σ is very small it “looks" from our measurements that $sd(Y^*) > 1$. In order to assess whether $\sigma \leq 1$, the measurements y^* alone would be useless, and a more precise measurement process (method) is needed.

Most measurement methods have some degree of error. In many cases this has no significant effect, but the only way to be sure is to understand (and study, if necessary) the measurement processes being used. Note that in some problems the measurement may involve a definition that is somewhat vague or subject to error. For example, if y is the number of incidents of sexual abuse that a person was subjected to over a 5 year period, then we should consider the definition of an “incident" and whether everyone interprets this the same way.

Finally, the number of population units to be included in a study is a crucial issue, since it is directly related to the amount of information that will be obtained. This topic will be discussed in subsequent chapters so we will not consider it here.

3.3 Data Collection

The previous sections noted the need to define study variables clearly and to have satisfactory methods of measuring them. It is difficult to discuss data collection except in the context of specific examples, but we mention a few relevant points.

- errors can occur in recording data or entering it in a data base, as well as in measurement of variables
- in many studies the “units" must be tracked and measured over a fairly long period of time (e.g. consider a stroke study in which persons are followed for 3 years)
- 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 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 conclusions
- in some studies the amount of data may be massive, so data base design and management is important.

3.4 Analysis and Conclusions

The remainder of this course is focussed on statistical methods of data analysis and inference, so we won’t discuss it in any detail here. Statistical analysis is a huge area, but there are a few major types of analysis:

- **estimation** of characteristics of populations or processes (e.g. the unemployment rate in Ontario, the probability a certain type of person has a stroke, the degree of association between dietary fat and heart disease)
- **testing hypotheses** (e.g. Is dietary fat related to heart disease? Does a drug prevent strokes? Is one computer sort algorithm better than another?)
- **model building** (e.g. produce a model relating lifestyle risk factors to the probability of a stroke; produce a model relating the strength of a steel bolt to its diameter; produce a model that can

be used to identify customers for a targeted marketing campaign.) Models are often used for classifying objects or units, or for making decisions.

- **exploratory analysis**: looking at data so as to uncover special structure or relationships. This is often done without any specific objectives (e.g. hypotheses to test or attributes to estimate) in mind. “Data mining" is an evocative term for this type of activity.

Statistical analysis uses a wide array of numerical and graphical methods. Several major topics in statistical analysis are introduced in Chapters 4 to 8 which follow. Specific applications are used to illustrate the methods and how conclusions are drawn from the analysis. Although we do not discuss it much in the remaining chapters, we must remember that well planned and conducted studies are important for drawing reliable conclusions.

3.4 Problems

1. Suppose you wish to study the smoking habits of teenagers and young adults, in order to understand what personal factors are relative 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 what variables you would collect data on.

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 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 Section 1.6, where heights x and weights y of individuals are to be recorded, discuss sources of variation due to the measurement of x and y on any individual.

STATISTICAL INFERENCE: 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 quantities and, importantly, for determining the uncertainty in an estimate.

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

Example 4.1.1. Suppose we want to estimate quantities associated with BMI for some population of individuals (e.g. Canadian males age 21-35). If the distribution of BMI values y 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 quantity associated with the BMI distribution. For example,

(i) $\mu = E(Y)$ is the average BMI in the population

(ii) μ is also the median BMI (50th percentile)

(iii) The 10th and 90th percentiles (.10 quantiles) for BMI are $y_{.10} = \mu - 1.28\sigma$ and $y_{.90} = \mu + 1.28\sigma$ (To see this, note for example that $P(Y \leq \mu - 1.28\sigma) = P(Z \leq -1.28) = .10$, where $Z = (Y - \mu)/\sigma \sim G(0, 1)$.)

(iv) The fraction of the population with BMI over 35.0 is $p = 1 - \Phi(\frac{35.0-\mu}{\sigma})$, where $\Phi(z)$ is the c.d.f for a $G(0,1)$ random variable.

Thus, if we collected a random sample of, say, 150 individuals and found maximum likelihood estimates (MLE's) $\hat{\mu} = 27.1, \hat{\sigma} = 3.56$ then estimates of the quantities in (i)-(iv) would be: (i) and (ii)$\hat{\mu} = 27.1$, (iii) $\hat{y}_{.10} = \hat{\mu} - 1.28\hat{\sigma} = 22.54,\ \hat{y}_{.90} = 31.66$, and (iv) $\hat{p} = .0132$.

The preceding example raises several issues, if we think about it.

- Where do we get our probability distribution? What if it isn't 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. In the data introduced in Section 1.5 it was not in fact clear that a Gaussian model was suitable when y is BMI. We will consider other models later.

- 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 variables (e.g. measurement error) must be included in the model. In the BMI example it has been assumed that a BMI was measured without error for a random sample of units (persons) from the population. In these notes it is typically assumed 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.

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

The third point is what we will focus on in this chapter; it is assumed that we can deal with the first two points with ideas introduced in Chapters 1 and 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 an **estimate** of θ, denoted as $\hat{\theta}$, is some function of the observed sample $y_1, \ldots, y_n$:

$$\hat{\theta} = g(y_1, \ldots, y_n). \tag{4.1.1}$$

Maximum likelihood provides a general method for obtaining estimates, but other methods also 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 MLE if Y has a Gaussian distribution but because of the Central Limit Theorem it is a good estimate of μ more generally. Thus, while we will use ML 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 point. 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 Section 1.6 gave five estimates $\hat{\theta} = \bar{y}$ of $E(Y)$: 1.723, 1.743, 1.734, 1.752, 1.736 (meters).The variability in estimates obtained from repeated samples of the same size is termed a sampling distribution. More precisely, we define this as follows. Let the r.v.'s $Y_1, \ldots, Y_n$ represent the observations in a random sample, and associate with the estimate $\hat{\theta}$ given by (4.1.1) a random variable

$$\tilde{\theta} = g(Y_1, \ldots, Y_n). \tag{4.1.2}$$

We call $\tilde{\theta}$ the **estimator** of θ corresponding to $\hat{\theta}$. (We will always use $\hat{\theta}$ to denote an estimate and $\tilde{\theta}$ to denote the corresponding estimator.) The distribution of $\tilde{\theta}$ is called the **sampling distribution** of the estimator.

Since $\tilde{\theta}$ is a function of the r.v.'s $Y_1, \ldots, Y_n$ 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 $\tilde{\theta}$ (we can think of this as describing an estimation **procedure**, if we wish) 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

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

for μ. At this point, we recall some probability theory which says that the distribution of $\bar{Y}$ is also Gaussian, $G(\mu, \sigma/\sqrt{n})$. Let us now consider the probability that $|\tilde{\mu} - \mu|$ is less than or equal to some specified value Δ. We have

$$\begin{aligned} pr(|\tilde{\mu} - \mu| \leq \Delta) &= P(\mu - \Delta \leq \bar{Y} \leq \mu + \Delta) \\ &= P(\frac{-\Delta\sqrt{n}}{\sigma} \leq Z \leq \frac{\Delta\sqrt{n}}{\sigma}). \end{aligned} \tag{4.1.3}$$

where $Z = (\bar{Y} - \mu)/(\sigma/\sqrt{n}) \sim G(0,1)$. Clearly, as n increases, the probability (4.1.3) approaches 1. 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 Section 1.5, and that we take $\Delta = .01$. That is, we want to find the probability that $|\tilde{\mu} - \mu|$ is no more than .01 meters. Assuming $\sigma = .07(m)$, which is roughly what we estimated it to be in Section 1.5, (4.1.3) gives the following results for a sample of size $n = 50$ and for a sample of size $n = 100$:

$$\begin{aligned} n = 50 \quad &= P(|\tilde{\mu} - \mu| \leq .01) = P(-1.01 \leq Z \leq 1.01) = .688 \\ n = 100 \quad &= P(|\tilde{\mu} - \mu| \leq .01) = P(-1.43 \leq Z \leq 1.43) = .847 \end{aligned}$$

This indicates that a large sample is "better" in the sense that the probability is higher that $\tilde{\mu}$ will be within .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 $y_1, \ldots, y_n$: we merely give probabilities like the above for the associated estimator, which indicate 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 in the estimator $\tilde{\mu}$ 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. 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. For illustration, consider the case where a variable Y has a $G(\mu, \sigma)$ distribution in the population, with $\mu = 100$ and $\sigma = 50$, and suppose we take samples $y = (y_1, \ldots, y_n)$ of size n, giving $\hat{\mu} = \bar{y}$. We can investigate the distribution of $\tilde{\mu}$ by simulation, by

1. (i) generating a sample of size n; in R this is done by

$$y \leftarrow rnorm(n, 100, 50).$$

(ii) computing $\hat{\mu} = \bar{y}$ from the sample; in R this is done by

$$ybar \leftarrow mean(y),$$

and then repeating 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 $\tilde{\mu}$, and we can study it by plotting a histogram or other plot of the values.

Exercise: Generate $k = 100$ samples this way, and plot a histogram based on the values $\bar{y}_1, \ldots, \bar{y}_{100}$.

The approaches illustrated in the preceding examples can be used generally. Given an estimator $\tilde{\theta}$, we can consider its sampling distribution and compute probabilities of the form $P(|\tilde{\theta} - \theta| \leq \Delta)$.

We need to be able to find the distributions of estimators and other variables. We now review some probability results from Stat 230 and derive a few other results that will be used in dealing with estimation and other statistical procedures.

4.2 Some Distribution Theory

4.2.1 Moment Generating Functions

Let Y be a random variable. The **moment generating function** (m.g.f) of Y is defined as

$$M(t) = E(e^{tY}),$$

assuming that this expectation exists for all t in some open set $(-a, a)$ of real numbers $(a > 0)$. If Y is continuous with p.d.f. $f(y)$ then

$$M(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy \tag{4.2.1}$$

and if Y is discrete with p.f. $f(y)$ then

$$M(t) = \sum_{y} e^{ty} f(y) \tag{4.2.2}$$

where the integral and sum in (4.2.1) and (4.2.2) are over the range of Y.

The m.g.f. is a transform, that is, a function $M(t)$ that is obtained from the function $f(y)$. Not all probability distributions have m.g.f.'s (since (4.2.1) or (4.2.2) may not exist in some cases) but if the m.g.f. exists, it uniquely determines the distribution. That is, a m.g.f $M(t)$ can arise from only one function $f(y)$.

Example 4.2.1. Binomial Distribution
The m.g.f.[3]. for the distribution $Bin(n, p)$ is $M(t) = (pe^t + 1 - p)^n$

We now give some simple theorems about m.g.f.'s, with applications. Moment generating functions are so-called because from them we can derive the moments of a random variable as follows.

Theorem 4.2.1[4] Let the r.v. Y have m.g.f. $M(t)$. Then for $r = 1, 2, 3, \ldots$

$$E(Y^r) = \frac{d^r M(t)}{dt^r}|_{t=0} = M^{(r)}(0) = \text{Coefficient of } \frac{t^r}{r!} \text{ in the power series representation of } M(t).$$

[3] $M(t) = \sum_{y=0}^{n} e^{ty}\binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1-p)^{n-y} = (pe^t + 1 - p)^n$ by the binomial theorem.

[4] **Proof**: For simplicity consider a continuous r.v. Y. Then $M^{(r)}(t) = \frac{d^r}{dt^r} \int_{-\infty}^{\infty} e^{ty} f(y) dy = \int_{-\infty}^{\infty} \frac{d^r (e^{ty})}{dt^r} f(y) dy = \int_{-\infty}^{\infty} y^r e^{ty} f(y) dy$. Therefore $M^{(r)}(0) = \int_{-\infty}^{\infty} y^r f(y) dy = E(Y^r)$.

Example 4.2.3[5] **Mean and Variance of Binomial Distribution**
For $Y \sim Bin(n,p)$, $E(Y) = np$, and $Var(Y) = np(1-p)$.

Theorem 4.2.2[6] Let X and Y be r.v.'s related by the fact that $Y = aX + b$, where a and b are constants. Then

$$M_y(t) = e^{bt} M_x(at), \tag{4.2.3}$$

where $M_x(t)$ is the m.g.f for X and $M_y(t)$ is the m.g.f. for Y.

Example 4.2.3 MGF's for Gaussian Distributions

The moment generating function of $Y \sim G(\mu, \sigma)$ is given by

$$M_y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \tag{4.2.4}$$

Proof. First consider $Z \sim G(0,1)$. If we can find $M_z(t)$ then we can get $M_y(t)$, where $Y \sim G(\mu, \sigma)$, from it. This is because $Z = (Y-\mu)/\sigma \sim G(0,1)$, or $Y = \sigma Z + \mu$. Thus

$$M_y(t) = e^{\mu t} M_z(\sigma t) \tag{4.2.5}$$

by Theorem 4.2.2. To find $M_z(t)$, we must evaluate

$$M_z(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

This can be obtained by "completing the square" in the exponent of the integrand: since

$$tz - \frac{1}{2}z^2 = -\frac{1}{2}[(z-t)^2 - t^2]$$

we get

$$\begin{aligned} M_z(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(z-t)^2 - t^2]} dz \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz = e^{\frac{1}{2}t^2}, \end{aligned}$$

since the integral is the integral of the p.d.f. for $G(t,1)$ and therefore equals 1. Now, using (4.2.3), we get

$$M_y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \tag{4.2.6}$$

as the m.g.f. for $Y \sim G(\mu, \sigma)$.

■

Exercise: Verify using (4.2.4) that $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

---

[5] **Proof:** we have $M(t) = (pe^t + 1 - p)^n$. This gives $M^{(1)}(t) = npe^t(pe^t + 1 - p)^{n-1}$ and $M^{(2)}(t) = npe^t(pe^t + 1 - p)^{n-1} + n(n-1)(pe^t)^2(pe^t + 1 - p)^{n-2}$. Therefore $E(Y) = M^{(1)}(0) = np$ and so $E(Y^2) = M^{(2)}(0) = np + n(n-1)p^2$. Finally $Var(Y) = E(Y^2) - [E(Y)]^2 = np(1-p)$.

[6] **Proof**: $M_y(t) = E(e^{tY}) = E(e^{t(aX+b)}) = E(e^{bt}e^{(at)X}) = e^{bt}E(e^{(at)X}) = e^{bt}M_x(at)$

Sums of Independent Random Variables We often encounter sums or linear combinations of independent random variables, and m.g.f.'s are useful in dealing with them.

Theorem 4.2.3 Let $Y_1, \ldots, Y_k$ be mutually independent random variables with m.g.f.'s $M_1(t), \ldots, M_k(t)$. Then the m.g.f. of

$$S = \sum_{i=1}^{k} Y_i$$

is

$$M_s(t) = \prod_{i=1}^{k} M_i(t)$$

Proof.

$$\begin{aligned} M_s(t) = E(e^{tS}) = E\left(e^{t\sum Y_i}\right) &= E\left(\prod_{i=1}^{k} e^{tY_i}\right) \\ &= \prod_{i=1}^{k} E\left(e^{tY_i}\right) \quad \text{since the } Y_i's \text{ are independent} \\ &= \prod_{i=1}^{k} M_i(t) \end{aligned}$$

■

This allows us to prove a fundamental property of Gaussian random variables. If we take linear combinations (such as averages or sums) of (independent) Gaussian random variables, the result is also a Gaussian random variable. This is one reason that the Gaussian distribution is so popular. For example in finance it is a great convenience in a model if the returns from individual investments are each normally distributed, so too is the total return from the portfolio.

Theorem 4.2.4 Let $Y_1, \ldots, Y_k$ be independent r.v.'s with $Y_i \sim G(\mu_i, \sigma_i)$. Then if $a_1, \ldots, a_k$ are constants, the distribution of

$$W = \sum_{i=1}^{k} a_i Y_i$$

is $G\left(\sum_{i=1}^{k} a_i\mu_i, \left(\sum_{i=1}^{k} a_i^2\sigma_i^2\right)^{1/2}\right)$.

Proof. By Theorems 4.2.2 and 4.2.3 we have the m.g.f. of a_iY_i as $M_{a_iY_i}(t) = M_{Y_i}(a_it) = exp(a_i\mu_it + \frac{1}{2}a_i^2\sigma_i^2t^2)$.

Thus

$$M_w(t) = \prod_{i=1}^{k} M_{a_i Y_i}(t)$$
$$= exp\left[\left(\sum_{i=1}^{k} a_i\mu_i\right)t + \frac{1}{2}\left(\sum_{i=1}^{k} a_i^2\sigma_i^2\right)t^2\right]$$

This is the m.g.f. for a $G\left(\sum_{i=1}^{k} a_i\mu_i, \left(\sum_{i=1}^{k} a_i^2\sigma_i^2\right)^{1/2}\right)$ r.v., which proves the theorem.

■

Corollary[7] Suppose that $Y_1, \ldots, Y_n$ are independent r.v.'s with $Y_i \sim G(\mu, \sigma)$. If $\bar{Y} = \sum Y_i/n$, then

$$\bar{Y} \sim G(\mu, \sigma/\sqrt{n}).$$

The χ^2 (chi-squared) Distribution

The χ^2 distribution arises in statistics and probability not as a model, but as a distribution derived from Gaussian r.v.'s. It is a continuous family of distributions on $(0, \infty)$ with p.d.f.'s of the form

$$f(x) = \frac{1}{2^{r/2}\Gamma(r/2)} x^{(r/2)-1} e^{-x/2} \qquad x > 0 \tag{4.2.7}$$

where r is a parameter with values in $\{1, 2, \ldots\}$. To denote that X has p.d.f. (4.2.8) we write $X \sim \chi^2_{(r)}$. The parameter r is referred to as the "degrees of freedom" (d.f.) parameter. The function $\Gamma()$ in (4.2.7) is the gamma function, defined as follows:

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx \tag{4.2.8}$$

for positive real numbers a. It is easily shown that for integers $a = 1, 2, 3 \ldots$ we get $\Gamma(a) = (a-1)!$ and integration by parts shows that for all $a > 1$,

$$\Gamma(a) = (a-1)\Gamma(a-1)$$

Problem 11 at the end of the chapter gives some results for the χ^2 distributions, including the fact that its m.g.f. is

$$M(t) = (1-2t)^{-r/2} \tag{4.2.9}$$

and that its mean and variance are $E(X) = r$ and $Var(X) = 2r$. The c.d.f. $F(x)$ can be given in closed algebraic form for even values of r, but tables and software give the functions's values. In R the

[7] **Proof**: Consider theorem 4.2.4 with $\mu_i = \mu, \sigma_i = \sigma^2$ and $a_i = 1/n$ for $i = 1, \ldots, n$.

functions $dchisq(x,r)$ and $pchisq(x,r)$ give the p.d.f. $f(x)$ and c.d.f. $F(x)$ for the $\chi^2_{(r)}$ distribution. A table with selected values is given at the end of these notes.

We now give a pair of important results. The first shows that when we add independent chi-squared random variables, the sum is also chi-squared, and the degrees of freedom simply sum.

Theorem 4.2.5[8] Let $W1,\ldots,W_n$ be independent r.v.'s with $W_i \sim \chi^2(n_i)$. Then $S = \sum_{i=1}^{n} W_i$ has a χ^2 distribution with degrees of freedom $\sum_{i=1}^{n} n_i$, i..e. $S \sim \chi^2_{(\sum_{i=1}^{n} n_i)}$.

The next result shows why the chi-squared distribution is important whenever we study Gaussian distributed random variables. It arises as the square of a standard normal random variable.

Theorem 4.2.6 If $Z \sim G(0,1)$ then the distribution of $W = Z^2$ is $\chi^2_{(1)}$.

Proof. The m.g.f. of W is

$$\begin{aligned} M_w(t) &= E(e^{tW}) \\ &= E(e^{tZ^2}) \\ &= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2(1-2t)} dz \end{aligned}$$

This integral exists (is finite) provided $1-2t > 0$, i.e., $t < 1/2$. Making the change of variable $y = z(1-2t)^{1/2}$ we get

$$\begin{aligned} M_w(t) &= (1-2t)^{-1/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \\ &= (1-2t)^{-1/2} \end{aligned}$$

This is the m.g.f. for $\chi^2_{(1)}$, so W must have a $\chi^2_{(1)}$ distribution. ■

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 be chi-squared!

Corollary:[9] If $Z_1,\ldots,Z_n$ are mutually independent $G(0,1)$ random variables and $S = \sum_{i=1}^{n} Z_i^2$,then $S \sim \chi^2_{(n)}$.

---

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

[9] **Proof**: By the theorem, each X_i^2 has a $\chi^2_{(1)}$ distribution. Theorem 4.2.5 then gives the result.

Limiting Distributions and Convergence in Probability

Sometimes we encounter a sequence of random variables $Y_1, Y_2, \ldots$ whose distributions converge to some limit. Random variables are said to converge in distribution if their corrresponding cumulative distribution functions converge. The following definition states this more precisely.

Definition (convergence in distribution): Let $\{Y_n : n = 1, 2, \ldots\}$ be a sequence of r.v.'s with c.d.f.'s $F_n(y), n = 1, 2, \ldots$. If there is a c.d.f. $F(y)$ such that

$$\lim_{n\to\infty} F_n(y) = F(y) \tag{4.2.10}$$

at all points y at which F is continuous, then we say that the sequence of r.v.'s has a **limiting distribution** with c.d.f. $F(y)$. We often write this as $Y_n \xrightarrow{d} Y$, where Y is a r.v. with c.d.f. $F(y)$.

A major use of limiting distributions is as **approximations**. In many problems the c.d.f. $F_n(y)$ may be complicated or intractable, but as $n \to \infty$ there is a much simpler limit $F(y)$. If n is sufficiently large, we often just use $F(y)$ as a close approximation to $F_n(y)$.

We will state and use some limiting distributions in these notes. A famous limiting distribution called the Central Limit Theorem is contained in the following theorem. In essence the theorem states that we may approximate the distribution of the sum of independent random variables using the normal distribution. Of course the more summands, the better the approximation. A way to prove the theorem is provided in Problem 4 at the end of the chapter.

Theorem 4.2.7 (Central Limit Theorem). Let $Y_1, \ldots, Y_n$ be independent r.v.'s, each having mean $E(Y_i) = \mu$ and variance $Var(Y_i) = \sigma^2$. Let $\bar{Y} = \sum_{i=1}^{n} Y_i/n$, and consider the "standardized" r.v.

$$Z_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}}$$

Then as $n \to \infty$ the distribution of Z_n converges to $G(0, 1)$.[10]

Remark: For n sufficiently large we may use the $G(0, 1)$ distribution to approximate the distribution of Z and thus to calculate probabilities for $\sum_{i=1}^{n} Y_i$. Recall that constant times a normal random variable is also normally distributed so the theorem also asserts that $\bar{Y}$ has a Gaussian distribution with mean μ and standard deviation $\sqrt{\sigma^2/n}$. The accuracy of the approximation depends on n (bigger is better) and also on the distributions of $Y_1, Y_2, \ldots$.[11]

[10] It is an interesting question as to whether this implies the p.d.f. of Z converges to the normal p.d.f. Is it possible for a sequence $F_n(x)$ to converge to a limit $F(x)$ and yet their derivatives do not converge, $F_n'(x) \nrightarrow F(x)$?

[11] Suppose Y_j is your winning in a lottery on week j and Y_j is either 1 million (with probability 10^{-6} or 0 with probability $1 - 10^{-6}$. How close do you think the distribution of $\sum_{i=1}^{100} Y_i$ is to a normal distribution?

Remark: It is possible to use the theorem to approximate discrete as well as continuous distributions. A very important case is the binomial distribution: let $Y_n \sim Bin(n,p)$ be a binomial r.v. Then as $n \to \infty$ the limiting distribution of

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

is $G(0,1)$. This is proved by noting that Y_n can be written as a sum of independent r.v.'s:

$$Y_n = \sum_{i=1}^{n} X_i$$

where $X_i \sim Bin(1,p)$, with $\mu = E(X_i) = p$ and $\sigma^2 = Var(X_i) = p(1-p)$. The result above is then given by Theorem 4.2.7.

Exercise: Use the limiting Gaussian approximation to the binomial distribution to evaluate $P(20 \leq Y \leq 28)$ and $P(16 \leq Y \leq 32)$, when $Y \sim Bin(60,.4)$. Compare the answers with the exact probabilities, obtained from the R function $pbinom(x,n,p)$. (Note: Recall that when approximating a **discrete distribution** we can use a **continuity connection**; this means here that we consider $P(19.5 \leq Y \leq 28.5)$ and $P(15.5 \leq Y \leq 32.5)$ when we apply the normal approximation.)

Besides convergence in distribution, there is another concept called **convergence in probability** which we will mention briefly. A sequence of random variables $\{X_n : n, 1, 2, \ldots\}$ is said to converge in probability to the constant c if,

$$\lim_{n\to\infty} P\{|X_n - c| \geq \epsilon\} = 0 \text{ for all } \epsilon > 0 \tag{4.2.11}$$

Loosely, this implies that however we define our tolerance ϵ, the probability that X_n is nearly equal to c (i.e. using this tolerance) gets closer and closer to 1. A major application of this concept is in showing that certain estimators $\tilde{\theta}_n$, based on a sample of size n, converge to the true value θ of the parameter being estimated as $n \to \infty$ (i.e. as sample size becomes arbitrarily large). Convergence in probability will not be used much in this course, but is an important tool in more advanced discussions of statistical theory. In fact we can show that the definition of convergence in distribution in (4.2.10)[12] to a constant is equivalent to convergence in (4.2.11)

4.2.2 Interval Estimation Using Likelihood Functions

The estimates and estimators discussed in Section 4.1 are often referred to as **point estimates (and estimators)**. This is because they consist of a single value or "point". The discussion of sampling

[12] What is the cumulative distribution function $F(y)$ of the constant c and at what points y is it continuous?

distributions shows how to address the uncertainty in an estimate, but we nevertheless prefer in most settings to make this uncertainty an explicit part of the estimate. This leads to the concept of an **interval estimate**, which takes the form

$$\theta \in (L, U) \quad \text{or} \quad L \leq \theta \leq U,$$

where $L = h_1(y_1, \ldots, y_n)$ and $U = h_2(y_1, \ldots, y_n)$ are based on the observed data. Notice that this provides an interval with endpoints L,and U both of which depend on the data. With random variables replacing the observed data, the endpoints $\tilde{L} = h_1(Y_1, \ldots, Y_n)$ and $\tilde{U} = h_2(Y_1, \ldots, Y_n)$ are random variables and there is a specific probability (hopefully large) that the parameter falls in this random interval, given by

$$P(\tilde{L} < \theta < \tilde{U}).$$

This probability, the coverage probability, gives an indication how good the interval estimate is. For example if it is 0.95, this means that 95% of the time (i.e. 95% of the different samples we might draw), the parameter falls in the interval $(\tilde{L}, \tilde{U})$ so we can be reasonably safe in assuming on his occasion, and for this dataset, it does so. In general, uncertainty in an interval estimate is explicitly stated by giving the interval estimate along with the probability $P\left[\theta \in (\tilde{L}, \tilde{U})\right]$, when $\tilde{L} = h_1(Y_1, \ldots, Y_n)$ and $\tilde{U} = h_2(Y_1, \ldots, Y_n)$ are the random variables associated with L and U.

The likelihood function can be used to obtain interval estimates for parameters in a very straightforward way. We do this here for the case in which the probability model involves only a single scalar parameter θ. Models with two or more parameters will be considered later. Individual models often have constraints on the parameters, so for example in the Gaussin distribution while the mean can take any real number $-\infty < \mu < \infty$ the standard deviation must be positive, i.e. $\sigma > 0$. Similarly in 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 before we often multiply the likleihood function by a convenient scale, resulting in the relative likelihood.

Definition: Suppose θ is scalar and that some observed data (say a random sample $y_1, ..., y_n$) have given a likelihood function $L(\theta)$. The **relative likelihood function** $R(\theta)$ is then defined as

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

where $\hat{\theta}$ is the m.l.e. (obtained by maximizing $L(\theta)$) and Ω is the parameter space. Note that $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$.

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

Actually, $\{\theta : \ R(\theta) \geq p\}$ is not necessarily an interval unless $R(\theta)$ is unimodal, but this is the case for all models that we consider. 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. The main challenge is to decide what p should be; we show later that choosing p in the range $.10 - .15$ is often useful. If you return to the likelihood for the Harris/Decima poll in Figure 2.12, note that 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 in the range 0.10-0.15.

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 $Y \sim \text{Bin}(n, \theta)$ is a reasonable model and the observed data $Y = y$ gives the likelihood function

$$L(\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \qquad 0 \leq \theta \leq 1.$$

We find $\hat{\theta} = y/n$ and then

$$R(\theta) = \frac{\theta^y (1-\theta)^{n-y}}{\hat{\theta}^y (1-\hat{\theta})^{n-y}} \qquad 0 \leq \theta \leq 1.$$

Figure 4.15 shows the $R(\theta)$ functions for two polls:

Poll 1: $n = 200$, $y = 80$

Poll 2: $n = 1,000$, $y = 400$.

In each case $\hat{\theta} = .40$, but the relative likelihood function is more "concentrated" around $\hat{\theta}$ for the larger poll (Poll 2). The $.10$ likelihood intervals also reflect this:

Poll 1: $R(\theta) \geq .1$ for $.33 \leq \theta \leq .47$

Poll 2: $R(\theta) \geq .1$ for $.37 \leq \theta \leq .43$.

The graph also shows the **log relative likelihood function**,

$$r(\theta) = \log R(\theta) = \ell(\theta) - \ell(\hat{\theta}),$$

where $\ell(\theta) = \log L(\theta)$ is the log likelihood function. It's often convenient to compute $r(\theta)$ instead of $R(\theta)$ and to compute a p likelihood interval by the fact that $R(\theta) \geq p$ if $r(\theta) \geq \log p$.

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 the relative likelihood function $R(\theta)$ or $r(\theta) = logR(\theta)$. 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

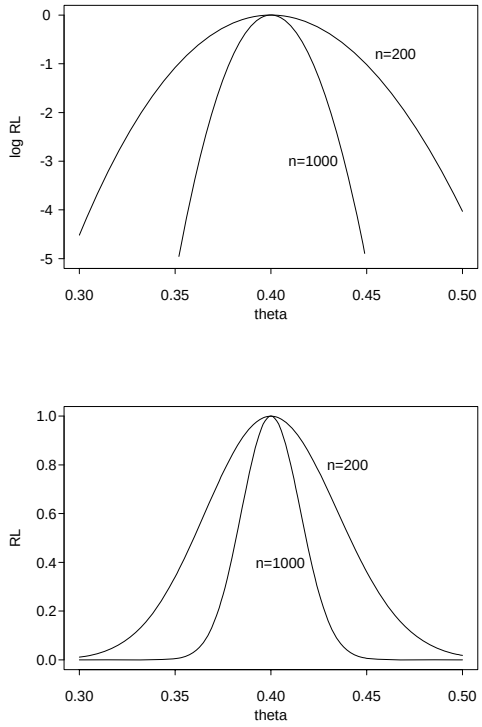


Figure 4.15: **Relative Likelihood and log Relative Likelihood Functions for a Binomial Parameter**

the choice of p. Sometimes it is argued that values like $p = .10$ or $p = .05$ 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.

4.3 Confidence Intervals for a Parameter

To start we consider the **coverage probability** for any interval estimate, as follows.

In general, a likelihood interval or any other interval estimate for θ takes the form $[L(\mathbf{Y}), U(\mathbf{Y})]$, where $\mathbf{Y}$ stands for the data the estimate is based on. Let the true unknown value of θ be θ_0. We now ask: "What is the probability that $L(\mathbf{Y}) \leq \theta_0 \leq U(\mathbf{Y})$"? Since $\mathbf{Y}$ represents a random sample of some kind this probability can be found by working with the probability distribution for $\mathbf{Y}$. The value

$$C(\theta_0) = P\{L(\mathbf{Y}) \leq \theta_0 \leq U(\mathbf{Y})\}$$

is called the **coverage probability** for the interval estimate. In practice, we try to find intervals for which $C(\theta_0)$ is fairly close to 1 (values .90, .95 and .99 are often used) while keeping the interval fairly short. Such interval estimates are called **confidence intervals** and the value of $C(\theta_0)$ is also called the **confidence coefficient**.

To show that such intervals exist, consider the following simple example.

Example 4.4.1 Suppose that a random variable Y has a $G(\mu_0, 1)$ distribution. That is, $\mu_0 = E(Y)$ is unknown but the standard deviation of Y is known to equal 1. Let the r.v.'s $Y_1, ..., Y_n$ represent a random sample of y-values, and consider the interval $(\overline{Y} - 1.96n^{-1/2},\ \overline{Y} + 1.96n^{-1/2})$, where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ is the sample mean. Since $\overline{Y} \sim G(\mu_0, n^{-1/2})$, we find that

$$P(\overline{Y} - 1.96n^{-1/2} \leq \mu_0 \leq \overline{Y} + 1.96n^{-1/2}) = P(-1.96 \leq \frac{\overline{Y} - \mu_0}{n^{-1/2}} \leq 1.96)$$
$$= P(-1.96 \leq Z \leq 1.96) = .95,$$

where $Z \sim G(0, 1)$. Thus the interval $(\overline{Y} - 1.96n^{-1/2},\ \overline{Y} + 1.96n^{-1/2})$ is a confidence interval for μ with confidence coefficient (coverage probability) .95.

It is important to note the interpretation for a confidence interval: if the **procedure** in question is used repeatedly then in a fraction $C(\theta_0)$ of cases the interval will contain the true value θ_0. If we actually took a sample of size 16 in Example 4.4.1 and found that the observed mean was $\bar{y} = 10.4$, then the **observed .95 confidence interval** would be $(\bar{y} - 1.96/4, \bar{y} + 1.96/4)$, or (9.91, 10.89). We do not say that the probability of the observed interval $9.91 \leq \mu_0 \leq 10.89$ is .95, but we have a high degree of **confidence** that this interval contains μ_0.

Confidence intervals tend to 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. We noted this characteristic earlier for likelihood intervals, and we show a bit later that likelihood intervals are a type of confidence interval. Note also that the coverage probability for the interval in the example did not depend on μ_0; we have $C(\mu_0) = .95$ for all μ_0. This is a highly desirable property because we'd like to know the coverage probability while not knowing the value of the parameter (θ_0). We next consider a general way to find confidence intervals which have this property.

Pivotal Quantities and Confidence Intervals

Definition: A **pivotal quantity** $Q = g(Y_1, ..., Y_n, \theta)$ is a function of a random sample $Y_1, ..., Y_n$ and θ such that the distribution of the r.v. Q is fully known. That is, the distribution does not depend on θ or other unknown information.

The motivation for this definition is that if the relationship $a \leq g(Y_1, ..., Y_n, \theta) \leq b$ can be rewritten as $L(Y_1, ..., Y_n) \leq \theta \leq U(Y_1, ..., Y_n)$, then $[L(Y_1, ..., Y_n), U(Y_1, ..., Y_n)]$ is a confidence interval for θ.

This is because

$$P\{a \le g(Y_1, ..., Y_n, \theta) \le b\}$$
$$= P\{L(Y_1, ..., Y_n) \le \theta \le U(Y_1, ..., Y_n)\} = \alpha$$

for some value α. Note that the coverage probability (confidence coefficient) α depends on a and b, but for any given values of a and b it can be found from the known distribution of $Q = g(Y_1, \ldots, Y_n, \theta)$. The value of α does not depend on the value of θ.

Example 4.4.2 Suppose $Y \sim G(\mu, \sigma_0)$ where μ is unknown but σ_0 is known. Then if $Y_1, ..., Y_n$ is a random sample, we know

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

so it 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 get a .95 confidence interval for μ we just need to find values a and b such that $P(a \le Q \le b) = .95$, and then

$$.95 = P(a \le \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \le b)$$
$$= P(\overline{Y} - \frac{b\sigma_0}{\sqrt{n}} \le \mu \le \overline{Y} - \frac{a\sigma_0}{\sqrt{n}}),$$

so that

$$\overline{Y} - \frac{b\sigma_0}{\sqrt{n}} \le \mu \le \overline{Y} - \frac{a\sigma_0}{\sqrt{n}} \tag{4.4.1}$$

is a .95 confidence interval for μ. Note that there are infinitely many pairs (a, b) giving $P(a \le Q \le b) = .95$. A common choice is $a = -1.96$, $b = 1.96$; this gives the interval $(\overline{Y} - 1.96\sigma_0/\sqrt{n}, \overline{Y} + 1.96\sigma_0/\sqrt{n})$ which turns out to be the shortest possible .95 confidence interval. Another choice would be $a = -\infty$, $b = 1.645$, which gives the interval $(\overline{Y} - 1.645\sigma_0/\sqrt{n}, \infty)$. This is useful when we are interested in getting 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 term it an **approximate pivotal quantity**.

Example 4.4.3. Polls Consider Example 4.3.1 discussed earlier, where $Y \sim \text{Bin}(n, \theta)$. For large n we know that $Q_1 = (Y - n\theta)/[n\theta(1-\theta)]^{1/2}$ is approximately $G(0, 1)$. It can also be proved that the

distribution of

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

where $\hat{\theta} = Y/n$, 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,

$$\begin{aligned} .95 &\doteq P(-1.96 \leq Q \leq 1.96) \\ &= P\left(\hat{\theta} - 1.96\left[\frac{\hat{\theta}(1-\hat{\theta})}{n}\right]^{1/2} \leq \theta \leq \hat{\theta} + 1.96\left[\frac{\hat{\theta}(1-\hat{\theta})}{n}\right]^{1/2}\right). \end{aligned}$$

Thus

$$\hat{\theta} \pm 1.96\left[\frac{\hat{\theta}(1-\hat{\theta})}{n}\right]^{1/2} \tag{4.4.2}$$

gives an approximate .95 confidence interval for θ. As a numerical example, suppose we observed $n = 100$, $y = 18$ in a poll. Then (4.4.2) becomes $.18 \pm 1.96[.18(.82)/100]^{1/2}$, or $.115 \leq \theta \leq .255$.

Remark: It is important to understand that confidence intervals may vary quite a lot when we take repeated samples. For example, 10 samples of size $n = 100$ which were simulated for a population where $p = 0.25$ gave the following .95 confidence intervals for p: .20 - .38, .14 - .31, .23 - .42, .22 - .41, .18 - .36, .14 - .31, .10 - .26, .21 - .40, .15 - .33, .19 - .37.

When we get a .95 confidence interval from a single sample, it will include the true value of p with probability .95, but this does not necessarily mean another sample will give a confidence interval that is similar to the first one. If we take larger samples, then the confidence intervals are narrower and will agree better. For example, try generating a few samples of size $n = 1000$ and compare the confidence intervals for p.

Likelihood-Based Confidence Intervals

It turns out that likelihood intervals are approximate confidence intervals and sometimes they are exact confidence intervals. Let $R(\theta) = L(\theta)/L(\hat{\theta})$ and define the quantity

$$\Lambda = -2\log R(\theta) = 2\ell(\hat{\theta}) - 2\ell(\theta). \tag{4.4.3}$$

This is called the **likelihood ratio statistic**. The following result can be proved:

If $L(\theta)$ is based on 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 where $R(\theta)$ is close to 1 (i.e. Λ is close to 0), we

get confidence intervals by working from the probability

$$P(\chi^2_{(1)} \leq c) \doteq P(\Lambda \leq c)$$
$$= P(L(Y_1, ..., Y_n) \leq \theta \leq U(Y_1, ..., Y_n)) \quad (4.4.4)$$

Example 4.4.4 Consider the binomial model in Examples 4.3.1 and 4.4.3 once again. We have, after rearrangement,

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

To get a .95 confidence interval for θ we note that $P(\chi^2_{(1)} \leq 3.841) = .95$. To find the confidence interval we have to find all θ values satisfying $\Lambda \leq 3.841$. This has to be done numerically, and depends on the observed data. For example, suppose that we observe $n = 100$, $y = 40$. Thus $\hat{\theta} = .40$ and the observed value of Λ is a function of θ,

$$\lambda(\theta) = 80\log(.4/\theta) + 120\log\left(\frac{.6}{1-\theta}\right).$$

Figure 4.16 shows a plot of $\lambda(\theta)$ and the line $g(\theta) = 3.841$ exactly, from which the confidence interval can be extracted. Solving $\lambda(\theta) \leq 3.841$, we find that $.307 \leq \theta \leq .496$ is the .95 confidence interval. We could also use the approximate pivotal quantity in Example 4.4.3 for this situation. It gives the .95 confidence interval (4.4.2), which is $.304 \leq \theta \leq .496$. The two confidence intervals differ slightly (they are both based on approximations) but are extremely close.

We can now see the connection between likelihood intervals and confidence intervals. The likelihood interval defined by $R(\theta) \geq p$ is the same as the confidence interval defined by $\Lambda(\theta) \leq -2\log p$. For a .95 confidence interval we use $\Lambda(\theta) \leq 3.841$ (since $P(\chi^2_{(1)} \leq 3.841) = .95$), which corresponds to $R(\theta) \geq .147$. Conversely a .10 likelihood interval given by $R(\theta) \geq .1$ corresponds to $\Lambda(\theta) \leq 4.605$. Since $P(\chi^2_{(1)} \leq 4.605) = .968$, we see that a **.10 likelihood interval is a confidence interval with approximate confidence coefficient .968**. Normally in statistical work, however, we use confidence intervals with (approximate) confidence coefficients .90,.95 or .99, and we usually employ $\Lambda(\theta)$) rather then $R(\theta)$ in discussions about likelihood-based interval estimates.

4.4.3 Choosing a Sample Size

We have seen in examples 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

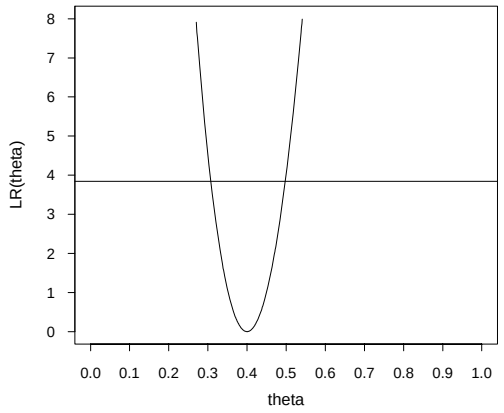


Figure 4.16: **Likelihood Ratio Statistic for Binomial Parameter**

θ from a binomial experiment in which the response variable Y has a $Bin(n, \theta)$ distribution. We will assume that the approximate pivotal quantity

$$Q = \frac{Y - n\theta}{[n\hat{\theta}(1-\hat{\theta}]^{1/2}} \simeq G(0,1) \tag{4.4.5}$$

introduced in Example 4.4.3 will be used to get confidence intervals for θ (Using the likelihood ratio statistic leads to a more difficult derivation and in any case, for large n the LR confidence intervals are very close to those based on Q.) Here is a criterion that is widely used for choosing the size of n: pick n large enough so a .95 CI for θ is no larger than $\hat{\theta} \pm .03$. Let's see why this is used and where it leads. From Example 4.4.3, we know that (see (4.4.2))

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

is an approximate .95 CI for θ. To make this CI $\hat{\theta} \pm .03$ (or even shorter, say $\hat{\theta} \pm .025$), we need n large enough that

$$1.96 \left[\frac{\hat{\theta}(1-\hat{\theta})}{n} \right]^{1/2} \leq .03$$

This can be rewritten as

$$n = \left(\frac{1.96}{.03} \right)^2 \hat{\theta}(1-\hat{\theta}).$$

We of course don't know what $\hat{\theta}$ will be once we take our sample, but we note that the worst case scenario is when $\hat{\theta} = .5$. So to be conservative, we find n such that

$$n = \left(\frac{1.96}{.03} \right)^2 (.05)^2 = 1067$$

Thus, choosing $n = 1067$ (or larger) will result in a .95 CI of the form $\hat{\theta} \pm c$, where $c \leq .03$. 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 θ (which is usually given in percentile form) satisfies $P(|\tilde{\theta} - \theta| \leq .03) = .95$, or equivalently, that the actual estimate $\hat{\theta}$ is the centre of a .95 confidence interval $\hat{\theta} \pm c$, for which $c = .03$. In practice, many polls are based on around 1050-1100 people, giving "accuracy to within 3 percent" (with probability .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 $n = 480$. In many situations this might not be sufficiently accurate for the purpose of the study, however.

Exercise: Show that to make the .95 CI $\hat{\theta} \pm .02$ or smaller, you need $n = 2401$ What should n be to make a .99 CI $\hat{\theta} \pm .02$ or less?

Remark: Very large binomial polls ($n \geq 2000$) are not done very often. Although we can in theory estimate $\hat{\theta}$ very precisely with an extremely large poll, there are two problems:

1. it is difficult to pick a sample that is truly random, so $Y \sim Bin(n, \theta)$ 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.

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 $\theta_j (j = 1, \ldots, k)$ or for functions $\psi = h(\theta_1, \ldots, \theta_k)$. For example, with a Gaussian model $G(\mu, \sigma)$ we might want to estimate μ and σ, and also the 90th percentile $\psi = \mu + 1.28\sigma$. 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. There also exist approximate pivotal quantities based on the likelihood function and m.l.e.'s. 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

$$R(\theta_1, \theta_2) = \frac{L(\theta_1, \theta_2)}{L(\hat{\theta}_1, \hat{\theta}_2)}$$

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 **p likelihood region** for (θ_1, θ_2). The concept of confidence intervals can similarly be extended to **confidence regions**.

4.3.1 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 the component that last for a period of years (or even months, in most cases) and there fore they use what are called **accelerated life tests**. These involve placing such high levels of stress on the components 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 PC's, with ambient temperature being the stress factor. As the temperature increases, the lifetimes of components tend to decrease and a temperature around 70° Celsius average lifetime 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) 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**: we know in each case only that the lifetime in question was over 600 hours. In Table 4.5.1 the asterisks denote the censored observations.

It is known from past experience that at each temperature level lifetimes follow close to an exponential distribution; let us therefore suppose that at temperature t $(t = 40, 50, 60, 70)$, component lifetimes Y have probability density function

$$f(y; \theta t) = \frac{1}{\theta t} e^{-y/\theta t} \qquad\qquad y \geq 0 \tag{4.5.1}$$

where $\theta t = E(Y)$ is the mean lifetime. The likelihood function based on a sample consisting of both censoring times and lifetimes is a little different from one based on a random sample of lifetimes. It is, for the tests at temperature t,

$$L(\theta_t) = \{\prod_{LT} \frac{1}{\theta_t} e^{-y_i/\theta t}\}\{\prod_{CT} e^{-y_i/\theta t}\} \tag{4.5.2}$$

where LT stands for the set of lifetimes yi and CT the set of censoring times yi.

Question 1 Show that for the distribution 4.5.1, $P(Y > y) = e - y/\theta_t$. Then describe how 4.5.2 is obtained.

Note that 4.5.2 can be rewritten as

$$L(\theta_t) = \frac{1}{\theta_t^r} e^{s/\theta_t} \tag{4.5.3}$$

where $r =$ number of lifetimes observed $S = \sum_{i=1}^{n} yi =$ sum of all lifetimes and censoring times.

Question 2 Assuming that the exponential model is correct, obtain m.l.e.'s $\hat{\theta}_1$ for the mean lifetime at each of the former termperature levels $t40, 50, 60, 70$. Graph the likelihood functions for θ_{40} and θ_{70} and comment on any qualitative differences.

Question 3 Check, perhaps using some kind of graph, whether the exponential model seems appropriate. Engineers used a model (called the Arrhenius model) that relates the mean lifetime of a component to the ambient temperature. This states that

$$\theta_t = exp\{\alpha + \frac{\beta}{t + 273.2}\} \tag{4.5.4}$$

where t is the temperature in degrees Celsius and α and β are parameters.

Questions 4 Make a plot of $log\hat{\theta}_t$ vs $(t + 273.2) - 1$ for the four temperatures involved in the life test experiment. Do the points lie roughly along a straight line? Give rough point estimators of α and

β. Extrapolate your plot or use your estimates of α and β to estimate the mean lifetime θ at $t = 20^\circ$ C, the normal temperature.

Question 5 A point estimate of θ at $20°$C is not very satisfactory. Outline how you might attempt to get an interval estimate based on the likelihood function. Once armed with an interval estimate, would you have many remaining qualms 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 4.5.4 applies, determine what α and β must approximately equal 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 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.

4.4 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 gender the heights Y in the population from which the samples were drawn is adequately represented by $Y \sim G(\mu, \sigma)$, obtain the m.l.e.'s $\hat{\mu}$ and $\hat{\sigma}$ in each case.

(b) Give the m.l.e.'s for the 10th and 90th percentiles of the height distribution for males and for females.

(c) Give the m.l.e.'s 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

$$\hat{Pr}(Y > 1.83) = \frac{\text{number of persons in sample with } y > 1.83}{n}$$

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 something we are in effect estimating the true value of the quantity; measurements of the same quantity on different occasions are usually not equal. A chemist has two ways of measuring a particular quantity μ; 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

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

where $w_1 = \frac{n}{\sigma_1^2}$ and $w_2 = \frac{m}{\sigma_2^2}$.

(b) Suppose that $\sigma_1 = 1$, $\sigma_2 = .5$ and $n = m = 10$. How would you rationalize to a non-statistician why you were using the estimate $\hat{\mu} = \frac{(\bar{x}+4\bar{y})}{5}$ instead of $\frac{(\bar{x}+\bar{y})}{2}$?

(c) Determine the standard deviation of $\hat{\mu}$ and of $(\bar{x} + \bar{y})/2$ under the conditions of part (b). Why is $\hat{\mu}$ a better estimate?

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.

   (a) Consider the random variable Y and estimator $\widetilde{p} = Y/n$. Describe how you could work out the probability that $-.03 \leq \widetilde{p} - p \leq .03$, if you knew the values of n and p.

   (b) Suppose that p is .40. Determine how large n should be in order to make $P(-.03 \leq \widetilde{p} - p \leq .03) = .95$. Use an approximation if you wish.

4. **Proof of Central Limit Theorem** (Special Case) Suppose $Y_1, Y_2, \ldots$ are independent r.v.'s with $E(Y_i) = \mu, Var(Y_i) = \sigma^2$ and that they have the same distribution, which has a m.g.f.

   (a) Show that $(Y_i - \mu)/\sigma$ has m.g.f. 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 m.g.f of the form $(1 + \frac{t^2}{2n} + 0(n))$, 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 m.g.f. is of the form $(1 + \frac{t^2}{2n} + 0(n))^n$. Show that as $n \to \infty$ this approaches the limit $e^{t^2/2}$, which is the m.g.f. for $G(0, 1)$. (Hint: For any real number a, $(1 + a/n)^n \to e^a$ as $n \to \infty$.)

5. 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 r.v.'s 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 i.i.d. random variables with p.d.f. $f(x; \theta)$. A point estimator $\hat{\theta}_n$ based on a random sample $X_1, \cdots, X_n$ is said to be consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

i. Let $X_1, \cdots, X_n$ be i.i.d. $U(0, \theta)$. Show that $\hat{\theta}_n$ is consistent for θ.

ii. Let $X \sim Bin(n, p)$. Show that $\hat{p}_n = X/n$ is consistent for p.

6. Refer to the definition of consistency in Problem 5(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

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

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 m.l.e. $\hat{\sigma}^2$ and show that it is not consistent. (To do this you have to find the m.l.e.'s 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?

7. Suppose that blood samples for nk 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: the nk samples are mixed together k at a time to give n pooled samples. A pooled sample will test negative if all k individuals 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 m.l.e. $\hat{\theta}$ and 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.)

8. (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 m.l.e. of α in this case. Plot the relative likelihood function with the one in (a), and compare the accuracy of estimation in the two cases.

9. 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 λ, obtain and graph the relative likelihood function $R(\lambda)$ 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 $R(\lambda)$ 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 $\Lambda(\lambda)$ as an approximate pivotal quantity to get .95 confidence intervals for λ 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?

10. The lifetime T (in days) of a particular type of lightbulb is assumed to have a distribution with p.d.f.

$$f(t;\lambda) = \frac{\lambda^3 t^2 e^{-\lambda t}}{2}, \quad t > 0;\ \lambda > 0.$$

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

$$L(\lambda) = \lambda^{3n} e^{-\lambda \sum t_i}.$$

Find the m.l.e. $\hat{\lambda}$ and the relative likelihood function $R(\lambda)$.

(b) If $n = 20$ and $\sum t_i = 996$, graph $R(\lambda)$ and determine the 10% likelihood interval for λ. What is the approximate confidence level associated with this interval?

(c) Suppose we wish to estimate the mean lifetime of a lightbulb. Show $E(T) = \frac{3}{\lambda}$. (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.

(d) The probability p that a lightbulb lasts less than 50 days is $p = P(T \leq 50) = 1 - e^{-50\lambda}[1250\lambda^2 + 50\lambda + 1]$. (Can you show this?) Thus $\hat{p} = .580$ and we can find a .95 confidence interval for p from a CI for λ. In the data referred to in part (b), 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 4.3.3 and 4.3.4). Find both intervals. What are the pros and cons of the second interval over the first one?

11. **The χ^2 (Chi-squared) distribution.** Consider the χ^2 distribution whose p.d.f. is given by $(4.2.8)$ in Section 4.2.3. If $Y \sim \chi^2_{(r)}$ then

(a) show that $f(y)$ integrates to 1 for any r in $\{1, 2, \ldots.\}$

(b) find the m.g.f of Y (see $(4.2.10)$) and use it to show that $E(Y) = r$ and $Var(Y) = 2r$.

(c) Plot the p.d.f.'s for the $\chi^2_{(5)}$ and $\chi^2_{(10)}$ distributions on the same graph.

12. In an early study concerning survival time for patients diagnosed with Acquired Immune Deficiency Syndrome (AIDS), the survival times (i.e. times between diagnosis of AIDS and death) of 30 male patients were such that $\sum_{i=1}^{30} x_i = 11,400$ 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 get an approximate .90 confidence interval for θ.

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

13. Let X have an exponential distribution with p.d.f.

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

where $\theta > 0$.

(a) Show that $Y = 2X/\theta$ has a $\chi^2_{(2)}$ distribution. (Hint: compare the p.d.f. of Y with (4.2.8).)

(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 12. Using the fact that

$$P\left(43.19 \leq \chi^2_{(60)} \leq 79.08\right) = .90$$

obtain a .90 confidence interval for θ based on U. Compare this with the interval found in 12(a). Which interval is preferred here? (Why?)

14. 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).

15. 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 ML estimators based on a sample of size n then it can be shown that when n is large,

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

is very close to $G(0, 1)$, and so it is approximately a pivotal quantity. Use Z to obtain a .99 confidence interval for μ for males and for females.

16. 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 feel about the accuracy of this estimate.

STATISTICAL INFERENCE: TESTING HYPOTHESES

5.1 Introduction

There are often hypotheses that a statistician or scientist might want to "test" in the light of observed data. Two important types of hypotheses are

(1) that a parameter vector 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 p.d.f. $f_0(y)$; we denote this as H_0: $Y \sim f_0(y)$.

The statistical approach to hypothesis testing is as follows: First, assume that the hypothesis H_0 will be tested using some random data "Data". Next, define a **test statistic** (also called a **discrepancy measure**) $D = g$(Data) that is constructed to measure the degree of "agreement" between Data 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. Third, once specific observed "data" have been collected, let $d_{obs} = g$(data) be the corresponding observed value of D. To test H, we now calculate the observed **significance level** (also called the **p-value**), defined as

$$SL = P(D \geq d_{obs}; H_0), \tag{4.5.2}$$

where the notation "; H_0" means "assuming H_0 is true". If SL 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 observed** is small. This makes the alternative explanation that 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 we have observed data that indicate poor agreement with H_0, or

(b) H_0 is false.

A SL less than about .05 provides moderately strong evidence against H_0.

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 $\theta = P$ (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 "Data" $= Y$ 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, we get $d_{obs} = 14$ and the significance level is (using the second definition of D)

$$\begin{aligned} SL &= P(D \geq 14; H_0) \\ &= P(Y \geq 44; \theta = 1/6) \\ &= \sum_{y=44}^{180} \left(\begin{array}{c} 180 \\ y \end{array} \right) (\frac{1}{6})^y (\frac{5}{6})^{180-y} = .005. \end{aligned}$$

This 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 SL is

$$\begin{aligned} SL &= P(Y \geq 35; \theta = 1/6) \\ &= \sum_{y=35}^{180} \left(\begin{array}{c} 180 \\ y \end{array} \right) (\frac{1}{6})^y (5/6)^{180-y} \\ &= .183. \end{aligned}$$

This probability is not especially small, so we conclude that there is no strong evidence against H_0. Note that we do **not** claim that H_0 is true, only that there is no evidence that it is not true.

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 1 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

$$D = |\overline{Y} - 1.00|$$

where $\overline{Y} = \sum Y_i/n$. Since $\overline{Y} \simeq G(\mu, \sigma/\sqrt{n})$, where $\mu = E(Y)$ and $\sigma^2 = \text{Var}(Y)$, we can compute the significance level (at least approximately) using a Gaussian distribution. Since we don't know σ^2 we will estimate it by the sample variance $s^2 = \sum(y_i - \bar{y})^2/(n-1)$ in the calculations below.

The samples from scales A and B above give us

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

The SL for A is (pretending $\sigma = s = 0.0230$)

$$\begin{aligned} SL &= P(D \geq 0.0061;\ \mu = 1.00) \\ &= P(|\overline{Y} - 1.00| \geq 0.0061) \\ &= P\left(\left|\frac{\overline{Y} - 1.00}{0.0230/\sqrt{10}}\right| \geq \frac{0.0061}{0.0230/\sqrt{10}}\right) \\ &= P(|Z| \geq 0.839) \ \text{ where } \ Z \sim G(0,1) \\ &= .401. \end{aligned}$$

Thus there is no evidence of bias (that is, that H_0: $\mu = 1.00$ is false) for scale A.

For scale B, however, we get

$$\begin{aligned} SL &= P\left(\left|\frac{\overline{Y} - 1.00}{.0170/\sqrt{10}}\right| \geq \frac{.0190}{.0170/\sqrt{10}}\right) \\ &= P(|Z| \geq 3.534) = .0004. \end{aligned}$$

Thus there is strong evidence against H_0: $\mu = 1.00$, 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. In fact, we can get an approximate .95 confidence interval for $\mu = E(Y)$ for scale B by using the approximate pivotal quantity

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

Since $P(-1.96 \leq Z \leq 1.96) \doteq .95$, we get the approximate .95 confidence interval $\bar{y} \pm 1.96s/\sqrt{10}$, or $0.981 \pm .011$, or $0.970 \leq \mu \leq 0.992$. Thus the bias in measuring the 1 kg weight is likely fairly

small (about 1% - 3%).

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

1. If the SL is small (close to 0) then the test indicates **strong evidence against** H_0; this is often termed “statistically significant" evidence against H_0. Rough rules of thumb are that $SL < .05$ provides moderately strong evidence against H_0 and that $SL < .01$ provides strong evidence.

2. If the SL is not small we do not conclude by saying 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 hypotheses 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 hypotheses 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**.

A drawback with the approach to testing described so far is that we are not told how to construct test statistics D. Often there are “intuitively obvious" statistics that can be used; this is the case in the examples of this section. However, in more complicated situations it is not always easy to come up with a test statistic. In the next section we show how to use the likelihood function to construct test statistics in general settings.

A final point is that once we have specified a test statistic D, we need to be able to compute the significance level (5.1.1) for the observed data. This 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 usually use approximations based on χ^2 distributions.

5.2 Testing Parametric Hypotheses with Likelihood Ratio Statistics

General Theory

First we show how test statistics can be constructed from the likelihood function for any hypothesis H_0 that is specified in terms of one or more parameters. Let "Data" represent data generated from a distribution with probability or probability density function f (Data; $\boldsymbol{\theta}$) 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 $p < k$. The dimensions of Ω and Ω_0 refer to the minimum number of parameters (or "coordinates") needed to specify points in them. We can test H_0 using as our **test statistic** the **likelihood ratio test statistic** Λ, defined as follows:

Let $L(\boldsymbol{\theta}) = f(\text{Data};\ \boldsymbol{\theta})$ be the likelihood function and let

$\hat{\boldsymbol{\theta}}$ denote the m.l.e. of $\boldsymbol{\theta}$ over Ω

$\hat{\boldsymbol{\theta}}_0$ denote the m.l.e. of $\boldsymbol{\theta}$ over Ω_0.

Now let

$$\Lambda = 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\hat{\boldsymbol{\theta}}_0) = -2\log\left\{\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right\}. \tag{5.2.1}$$

This seems like a sensible way to measure the degree of agreement between H_0 and the observed data: we look at the relative likelihood

$$R(\hat{\boldsymbol{\theta}}_0) = L(\hat{\boldsymbol{\theta}}_0)/L(\hat{\boldsymbol{\theta}})$$

of the "most likely" vector $\hat{\boldsymbol{\theta}}_0$ under H_0 (i.e. when $\boldsymbol{\theta}_0$ is in Ω_0). If this is very small then there is evidence against H_0 (think why). The reason we take $D = \Lambda$ as the test statistic instead of $R(\hat{\boldsymbol{\theta}}_0)$ is that $\Lambda = -2logR(\hat{\boldsymbol{\theta}}_0)$ takes values ≥ 0, with $\Lambda = 0$ $(R(\hat{\boldsymbol{\theta}}_0) = 1)$ representing the best agreement between H_0 and the data. Also, it can be shown that under H_0, the distribution of Λ becomes $\chi^2_{(k-p)}$ as the size of the data set becomes large. Large values of Λ_{obs} (small values of $R(\hat{\boldsymbol{\theta}}_0)$) indicate evidence **against** H_0 so the **p-value** (**significance level**) is

$$SL = P(\Lambda \geq \Lambda_{obs};\ H_0) \doteq P(\chi^2_{(k-p)} \geq \Lambda_{obs}). \tag{5.2.2}$$

(**Note**: Here we are using Λ_{obs} to represent the value of Λ obtained when we get the (numerical) data; Λ represents the r.v. when we think, as usual, of the data as random variables, before they are collected.)

Some Examples

This approach is very general and can be used with many different types of problems. A few examples follow.

(a) **A single parameter model.** Suppose Y has an exponential distribution with p.d.f. $f(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$ $(y \geq 0)$. 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

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

(b) **A model with two parameters.** Suppose $Y \sim G(\mu, \sigma)$ with p.d.f. $f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$ $(-\infty < y < \infty)$, $\boldsymbol{\theta} = (\mu, \sigma)$

Test H_0: $\sigma = \sigma_0$ based on a random sample $y_1, ..., y_n$. Thus $\Omega = \{(\mu,\sigma) : -\infty < \mu < \infty, \sigma > 0\}$, $\Omega_0 = \{(\mu, \sigma_0), -\infty < \mu < \infty\}$; $k = 2$, $p = 1$ and

$$L(\boldsymbol{\theta}) = L(\mu,\sigma) = \prod_{i=1}^{n} f(y_i;\mu,\sigma)$$

(c) **Comparison of two parameters.** Suppose we have data from two Poisson distributions with p.f.'s $f(y_{1i};\mu_1) = e^{-\mu_1}\frac{\mu_1^{y_{1i}}}{y_{1i}!}$; $f(y_{2i};\mu_2) = e^{-\mu_2}\frac{\mu_2^{y_{2i}}}{y_{2i}!}$,
where y_{1i} and y_{2i} take values in $\{0, 1, 2, \ldots\}$.

Test H_0: $\mu_1 = \mu_2$ based on two independent random samples y_{1i} $(i = 1, ..., n_1)$ and y_{2i} $(i = 1, ..., n_2)$. Thus $\boldsymbol{\theta} = (\mu_1, \mu_2)$ and $\Omega = \{(\mu_1, \mu_2) : \mu_1 > 0, \mu_2 > 0\}$, $\Omega_0 = \{(\mu,\mu) : \mu > 0\}$, $k = 2$, $p = 1$ and

$$L(\boldsymbol{\theta}) = L(\mu_1,\mu_2) = \prod_{i=1}^{n_1} f(y_{1i};\mu_1)\prod_{i=1}^{n_2} f(y_{2i};\mu_2)$$

(d) **A test about multinomial probabilities**. Consider a multinomial p.f.

$$f(y_1, ..., y_m; p_1, ..., p_m) = \frac{n!}{y_1!\cdots y_m!}p_1^{y_1}p_2^{y_2}\cdots p_m^{y_n} \quad \begin{pmatrix} 0 \leq y_j \leq n \\ \sum y_j = n \end{pmatrix}$$

Test H_0: $p_j = p_j(\boldsymbol{\alpha})$ where $\dim(\boldsymbol{\alpha}) = p < m - 1$. Thus $\boldsymbol{\theta} = (p_1, ..., p_m)$ and $\Omega = \{(p_1, ..., p_m) : 0 \leq p_j \leq 1, \sum_1^n p_j = 1\}$, $\Omega_0 = \{(p_1, .., p_m) : p_j = p_j(\boldsymbol{\alpha}) \text{ for } \boldsymbol{\alpha} \in \Omega_\alpha\}$; $k = m - 1$, $p = p$ and

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

We now consider problems that involve actual data, and describe the steps in the tests in more detail. In many testing problems it is necessary to use numerical methods to maximize likelihoods and find $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_0$. In this course and the examples below we focus on problems in which these estimates can be formed mathematically.

Example 5.2.1. Lifetimes of light bulbs. The variability in lifetimes of light bulbs (in hours, say, of operation before failure) and other electrical components is often well described by an exponential distribution with p.d.f. of the form

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

where $\theta = E(Y)$ is the average (mean) lifetime. 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 20 light bulbs was tested over a long period and 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 $\sum y_i$ 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 setup is as described in Example (a) above: the likelihood function based on a random sample $y_1, \ldots, y_n$ is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta}\, e^{-y_i/\theta} = \frac{1}{\theta^n} e^{-\sum y_i/\theta}.$$

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 $\Omega = \{2000\}$. The dimensions of Ω and Ω_0 are 1 and 0, respectively. We use the likelihood ratio statistic Λ of (5.2.1) as our test statistic D. To evaluate this we first write down the log likelihood function (noting that $n = 20$ and $\sum y_i = 38,524$ here)

$$\ell(\theta) = -20\ log\ \theta - \frac{38524}{\theta}.$$

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

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

Now we can compute the "observed" value of Λ from (5.2.1) as

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

The final computational step is to compute the significance level, which we do using the χ^2 approximation (5.2.2). This gives

$$\begin{aligned}SL &= P(\Lambda \geq 0.028; H_0 \text{ true}) \\ &= P(\chi^2_{(1)} \geq 0.028) \\ &= 0.87\end{aligned}$$

The SL is not close to zero so we conclude that there is no evidence against H_0 and against the manufacturer's claim that θ is 2000 hours. Although the m.l.e. $\hat{\theta}$ was under 2000 hours (1926.2) it was not sufficiently under to give conclusive evidence against $H_0 : \theta = 2000$.

Example 5.2.2 Comparison of Two Poisson Distributions. In problem 9 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 p.f. of the form

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

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 can be make to the analysis to deal with this.)

The number of failures in 12 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.

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

Denote the value of μ for Company A's copiers as μ_A and the value for Company B's as μ_B. It appears from the data that B's copiers fail less often, but let us assess this formally by testing the hypothesis

$$H_0 : \mu_A = \mu_B$$

This problem was sketched in Example (c) above. To handle it we consider the likelihood functions for μ_A and μ_B based on the observed data and then the likelihood formed by combining them. The likelihoods for μ_A and μ_B are

$$L_1(\mu_A) = \prod_{i=1}^{12} e^{-\mu_A} \frac{\mu_A^{y_{Ai}}}{y_{Ai}!} = e^{-12\mu_A} \mu_A^{240} \times \text{c}onstant$$
$$L_1(\mu_B) = \prod_{i=1}^{12} e^{-\mu_B} \frac{\mu_B^{y_{Bi}}}{y_{Bi}!} = e^{-12\mu_B} \mu_B^{140} \times \text{c}onstant$$

To test H_0 we view the two Poisson distributions together as one big model, with the parameter vector $\boldsymbol{\theta} = (\mu_A, \mu_B)$. In terms of the general framework for likelihood ratio tests the parameter space for $\boldsymbol{\theta}$ is $\Omega = \{(\mu_A, \mu_B) : \mu_A > 0, \mu_B > 0\}$, and under H_0 the parameter space becomes $\Omega_0 = \{(\mu_A, \mu_B) : \mu_A = \mu_B > 0\}$. The likelihood function for the "combined" model is the product of $L_1(\mu_A)$ and $L_2(\mu_B)$ since the two samples are independent:

$$L(\boldsymbol{\theta}) = L(\mu_A, \mu_B) = L_1(\mu_A) L_2(\mu_B)$$

We now carry out the steps for the test of H_0 exactly as in the previous example, except that now Ω has dimension $k = 2$ and Ω_0 has dimension $p = 1$. First, we write down the log likelihood function,

$$\ell(\boldsymbol{\theta}) = \ell(\mu_A, \mu_B) = -12\mu_A + 240\, log\, \mu_A - 12\mu_B + 140\, log\, \mu_B \tag{5.2.3}$$

Next we find $\hat{\theta}$ and $\hat{\theta}_0$. The m.l.e. $\hat{\theta}$ maximizes $\ell(\mu_A, \mu_B)$ in the unconstrained case. This can be done by solving the maximum likelihood equations

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

which gives $\hat{\mu}_A = 240/12 = 20.0$ and $\hat{\mu}_B = 140/12 = 11.667$. That is, $\hat{\boldsymbol{\theta}} = (20.0, 11.667)$. The constrained m.l.e. $\hat{\boldsymbol{\theta}_0}$ maximizes $\ell(\mu_A, \mu_B)$ under the constraint $\mu_A = \mu_B$. To do this we merely have to maximize

$$\ell(\mu_A, \mu_A) = -24\mu_A + 380\, log\, \mu_A$$

with respect to μ_A. Solving $\partial \ell(\mu_A, \mu_A)/\partial \mu_A = 0$, we find $\hat{\mu}_A = 380/24 = 15.833 (= \hat{\mu}_B)$; 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 (5.2.1) and (5.2.3) is

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

Finally, we compute the significance level for the test, which by (5.2.2) is

$$\begin{aligned}P(\Lambda \geq 26.64; H_0 \text{ true}) &= P(\chi^2_{(1)} \geq 26.64) \\ &= .25 \times 10^{-7}\end{aligned}$$

Our conclusion is that there is very strong evidence against the hypothesis; the test indicates that Company B's copiers have a lower rate of failure than Company A's copiers.

We can follow this conclusion up by giving confidence intervals for μ_A and μ_B; this indicates the magnitude of the difference in the two failure rates. (The m.l.e.'s $\hat{\mu}_A = 20.0$ average failures per month and $\hat{\mu_B} = 11.67$ failures per month differ a lot, but we also give confidence intervals in order to express the uncertainty in such estimates.)

Other hypothesis tests are considered in the remaining chapters of these notes. We conclude with some short remarks about hypothesis testing and estimation.

5.3 Hypothesis Testing and Interval Estimation

Hypothesis tests that are 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 (LR) tests the connection is immediately obvious, because the LR statistic

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

is used for both tests and confidence intervals. For a test, the significance level using (5.2.2) is

$$SL = P(\chi^2_{(1)} \geq \Lambda_{obs}(\theta_0)), \tag{5.3.1}$$

where we write $\Lambda_{obs}(\theta_0)$ to remind us that we are testing $H_0 : \theta = \theta_0$. On the other hand, to get a confidence interval for θ with confidence coefficient α we find by (4.4.4) all values θ_0 such that

$$\Lambda_{obs}(\theta_0) \leq \chi^2_{(1),\alpha} \tag{5.3.2}$$

where $\chi^2_{(1),\alpha}$ is the α quantile of a $\chi^2_{(1)}$ distribution; for example for a .95 confidence interval we use $\chi^2_{(1),.95} = 3.84$.

We now see the following by comparing (5.3.1) and (5.3.2):

- The parameter value θ_0 is inside an α confidence interval given by (5.3.2) if and only if (iff) for the test of $H_0 : \theta = \theta_0$ we have $SL \geq 1 - \alpha$.

For example, θ_0 is **inside** the .95 confidence interval (CI) iff the significance level for $H_0 : \theta = \theta_0$ satisfies $SL \geq .05$, To see this note that

$$\begin{aligned} & SL \geq .05 \\ & \Leftrightarrow P(\chi^2_{(1)} \geq \Lambda_{obs}(\theta_0)) \geq .05 \\ & \Leftrightarrow \Lambda_{obs}(\theta_0) \leq 3.84 \\ & \Leftrightarrow \theta_0 \text{ is inside } .95\ CI \end{aligned}$$

The connection between tests and confidence intervals can also be made when other test statistics beside the LR 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 $SL \geq .05$, or an α CI by finding values θ_0 such that $SL \geq 1 - \alpha$.

5.4 Problems

1. The accident rate over a certain stretch of highway was about $\lambda = 10$ 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 $\lambda = 10$; 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 λ and then testing $H_0 : \lambda = 10$. 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 $U(0, 1)$ (uniform distribution on $(0, 1)$) 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 $U(0, 1)$.

(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 PC 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 $x_1, \ldots, x_{20}$ 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: $\sum_{1}^{20} x_i = 18,698.6$)

   Failure times X_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 PC 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 $p_i = P$(digit i is drawn on any trial), $i = 0, 1, \ldots 9$. If the machines operate in a truly 'random' fashion, then we should have $p_i = .1$ $(i = 0, 1, \ldots, 9)$.

(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 value of the test statistic, D. What would you now do to test the hypothesis $p_i = .1$ $(i = 0, 1, 2, \ldots, 9)$? (Hint: You need to consider P(largest of 6 independent D's is $\geq D_{\text{obs}}$).)

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 p.f.

$$f(y_1, y_2, y_3) = \frac{n!}{y_1!, y_2!, y_3!} p_1^{y_1} p_2^{y_2} p_3^{y_3}, \qquad y_i \geq 0, \sum y_i = n$$

and since $p_1 + p_2 + p_3 = 1$ the parameter space $\Omega = \{(p_1, p_2, p_3) : p_i \geq 0, \sum p_i = 1\}$ has dimension 2. The genetic model discussed earlier specified that p_1, p_2, p_3 can be expressed in terms of only a single parameter θ $(0 < \theta < 1)$, as follows:

$$p_1 = \theta^2, \; p_2 = 2\theta(1 - \theta), \; p_3 = (1 - \theta)^2 \tag{5.4.1}$$

Consider (5.4.1) as a hypothesis H_0 to be tested. In that case, the dimension of the parameter space for (p_1, p_2, p_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 model (5.4.1) and state your conclusion.

7. **Likelihood ratio test for a Gaussian mean**. Suppose that a r.v 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) Set this up as a likelihood ratio test. (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) Derive the LR statistic Λ and show that it can be expressed as a function of $t = \sqrt{n}(\bar{y} - \mu_0)/s$, where s is the sample standard deviation and $\bar{y}$ is the sample mean. (Note: the easily proved identity

$$\sum_{i=1}^{n} (y_i - \mu_0)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu_0)^2 \tag{5.4.2}$$

can be used here.)

8. **Use of simulation to obtain significance levels.** In some testing problems the distribution of the test statistic D is so complicated it is not possible (or very difficult) to find the significance level,

$$SL = P(D \geq D_{obs}; H_0 \text{ true})$$

mathematically. In many problems computer simulation can be used as an alternative. Here is an approach that can be used with the "runs test" for randomness in a binary sequence that was discussed in Problem 7 of Chapter 1. For illustration we consider sequences of length 50.

Let R denote the number of runs in a sequence of 50 binary (0 or 1) digits. If the probability a digit is 1 is p then, from part (b) of Problem 7, Chapter 1, we have $E(R) = 1 + 98p(1-p)$. If $p = 5$ (i.e. 0 and 1 both have probability .5) then $E(R) = 25.5$. Thus, let us use

$$D = |R - 25.5|$$

as our statistic for testing the hypothesis

H_0: Digits come from a Bernoulli process with $p = .5$.

(a) Suppose you observe $R = 14$ and want to find the SL. Since $D_{obs} = 11.5$,

$$SL = P(D \geq 11.5) = P(R \leq 14) + P(R \geq 37)$$

Evaluate this in the following way:

(i) Simulate a sequence of 50 independent binary digits, with $P(0) = P(1) = .5$

(ii) Determine R and store it.

(iii) Repeat this 1000 times and determine the fraction of times that either $R \leq 14$ or $R \geq 37$. This is an approximation to SL (Why?).

Note: This problem is tricky because it requires that code be written to deal with step (ii). Step (i) can be handled in the R statistical software system by the command $x \leftarrow rbinom(50, 1, .5)$, which generates 50 independent $Bin(1, .5)$ random variables. The vector x thus is a random binary sequence of the type desired.

9. 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 $\lambda_j (j = 1, \ldots, K)$ 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 $P_j \lambda_j t$, we can estimate and compare the λ_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 $\lambda_1 = \lambda_2 = \ldots = \lambda_K$.

(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

Gaussian models for response variables Y are very widely used in applications of statistics. We have seen examples involving variables such as heights and body-mass index measurements of people previously in these notes. Many problems involve explanatory variables x (which may be a vector) that are related to a response Y; in this case we can generalize the simple Gaussian model $Y \sim G(\mu, \sigma)$ to $Y \sim G(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x}))$, where $\boldsymbol{x}$ is a vector of covariates (explanatory variables). The trick in creating models in such settings is to decide on the forms for $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$,

$$\mu(\boldsymbol{x}) = g_1(\boldsymbol{x}), \qquad \sigma(\boldsymbol{x}) = g_2(\boldsymbol{x})$$

To do this we rely on past information and on current data from the population or process in question.

Here are some examples of settings where Gaussian models could 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

$$Y \sim G(\mu_0, \sigma_0) \qquad Y \sim G(\mu_N, \sigma_N).$$

for the old and new machines. In this case there is no formula relating μ and σ to the machines; they are simply different.

Example 6.1.2 Price vs. Size of Commercial Buildings (Reference: Oldford and MacKay STAT 231 Course Notes, Ch. 16)

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.

A large 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 6.17 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 building in question was 4.47 x 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 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

$$Y_i \sim G(\mu_i, \sigma) \qquad\qquad \mu_i = \beta_0 + \beta_1 x_i$$

where β_0 and β_1 are parameters. In this model we've assumed that $\sigma_i = \sigma(x_i) = \sigma$, a constant. Alternatively, we could let it depend on x_i somehow. (Note that the scatter plot does not provide much information on how to do this, however.)

file=st241.fig611.ps,angle=0,width=

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

241 notes Lawless/st241.fig611.ps 241 notes Lawless/st241.fig611.ps

Figure 6.17: **Scatter Plot of Size vs. Price for 30 Buildings**

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 6.18

Diameter	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 $Y \sim G(\mu(x), \sigma)$, because the variability in y-values appears to be about the same for bolts of different diameters. Its not clear what the best choice for $\mu(x)$ would be; the relationship looks slightly nonlinear so presumably we want

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

or some other nonlinear function.

A **Gaussian response model** is one for which the distribution of the response variable Y, **given** the associated covariates $\boldsymbol{x}$ for an individual unit, is of the form

$$Y \sim G(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x})). \tag{6.1.1}$$

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

$$Y_i \sim G(\mu_i, \sigma_i) \qquad\qquad i = 1, \ldots, n$$

where $\mu_i = g_1(\boldsymbol{x}_i)$ and $\sigma_i = g_2(\boldsymbol{x}_i)$ for some specified functions g_1 and g_2. In many problems it is only μ_i that depends much on $\boldsymbol{x}_i$ and we then use models where $\sigma_i = \sigma$ is constant. Furthermore, in a great many situations μ_i can be written as a **linear function** of covariates. These models are called **Gaussian linear models** and are of the following form:

$$Y_i \sim G(\mu_i, \sigma) \qquad\qquad i = 1, \ldots, n \tag{6.1.2}$$

with

$$\mu_i = \sum_{j=1}^{k} \beta_j x_{ij}, \qquad\qquad i = 1, \ldots, n \tag{6.1.3}$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i1}, \ldots, x_{ik})$ is the vector of covariates associated with unit i and the β_j's are parameters.

241 notes Lawless/st241.fig612.ps 241 notes Lawless/st241.fig612.ps

Figure 6.18: **Scatter Plot of Diameter vs. Strength for Steel Bolts.**

Remark: Sometimes the model (6.1.2) is written a little differently as

$$Y_i = \mu_i + R_i \text{ where } R_i \sim G(0, \sigma).$$

This splits Y_i into deterministic (μ_i) and random (R_i) components.

These models are also referred to as **linear regression models**, and the β_j's are called the **regression coefficients**. (The term "regression" is used because it was introduced in the 19th century in connection with these models. We won't bother explaining just how the term arose.)

The model (6.1.2) plus (6.1.3) describes many situations well. The following are some illustrations.

1. $Y \sim G(\mu, \sigma)$, where Y is the height of a random female, corresponds to $\boldsymbol{x}_i = (1), \beta_1 = \mu$
2. The model in Example 6.1.2 had $\mu_i = \beta_0 + \beta_1 x_i$ where x_i was the building's size. This can be re-expressed as $\mu_i = \beta_0 x_{i0} + \beta_1 x_{i1}$ where $x_{i0} = 1, x_{i1} = x_i$ (here we've used x_{ij} with $j = 0, 1$ for simplicity.)

3. The bolt strength model in Example 6.1.3 had $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$.
   This could be re-expressed as

$$\mu_i = \beta_0 x_{i1} + \beta_1 x_{i1} + \beta_2 x_{i2}$$

   where $x_{i0} = 1, x_{i1} = x_i, x_{i2} = x_i^2$.

Now we'll consider estimation and testing procedures for Gaussian models. We begin in the next section with models that have no covariates.

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 MLE's of μ and σ^2 are

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

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

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

We now consider interval estimation and tests of μ and σ.

6.2.1 Confidence Intervals and Tests About 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 (CI's) for μ. However, σ is generally unknown. Fortunately it turns out that if we simply replace σ with either $\hat{\sigma}$ or s in Z, then we still have a pivotal quantity which we denote as T. We will write T in terms of s since the formulas below look a little simpler then, so T is defined as

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \tag{6.2.1}$$

Since s is treated as a r.v. in (6.2.1) (we'll use s to represent both the r.v. and an observed value, for convenience), T does not have a $G(0,1)$ distribution. It turns out that its distribution in this case is what is known as a **student-t** (or just "t") distribution. We'll digress briefly to present this distribution and show how it arises.

Student -t Distributions

This distribution arises when we consider independent r.v.'s $Z \sim G(0,1)$ and $U \sim \chi^2_{(\nu)}$ and then define the new r.v.

$$X = \frac{Z}{(U/\nu)^{1/2}} \tag{6.2.2}$$

Then X has a **student -t distribution with ν degrees if freedom** (d.f.), and we write $X \sim t_{(\nu)}$ to denote this. The p.d.f. of X can be shown by a bivariate change of variables method (we won't do this) to be

$$f(x) = k_\nu (1 + \frac{x^2}{\nu})^{-(\nu+1)/2} \qquad\qquad -\infty < x < \infty \tag{6.2.3}$$

where $\Gamma(\cdot)$ is the gamma function and

$$k_\nu = \Gamma\left(\frac{\nu+1}{2}\right) / \sqrt{\nu\pi}\Gamma(\nu/2) \tag{6.2.4}$$

This distribution is symmetric about $x = 0$ and for large ν is closely approximated by the p.d.f. of $G(0,1)$. Problem 1 at chapter's end considers some properties of $f(x)$.

Probabilities for the $t - distribution$ are available from tables or computer software. In R, the c.d.f. value

$$F(x) = P(t_{(\nu)} \leq x) \tag{6.2.5}$$

is calculated as $pt(x, \nu)$. For example, $pt(1.5, 10)$ gives $P(t_{(10)} \leq 1.5)$ as .918.

Confidence Intervals for μ

We can show using (6.2.2) that (6.2.1) has a t-distribution with $n-1$ degrees of freedom:

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{(n-1)} \tag{6.2.6}$$

To do this we use the fact that

(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 are independent.

(The results (ii) and (iii) are not too hard to show using multivariate calculus. Proofs are omitted here but done in Stat 330.)
Then, we see that X defined in (6.2.2) is just the r.v. T here, so (6.2.6) follows and T is a pivotal quantity.

Therefore, to get an α CI for μ we find values a_1 and a_2 such that

$$P(a_1 \leq t_{(n-1)} \leq a_2) = \alpha = P\left(a_1 \leq \frac{\bar{Y} - \mu}{s/\sqrt{n}} \leq a_2\right).$$

This converts to

$$P\left(\bar{Y} - a_2\frac{s}{\sqrt{n}} \leq \mu \leq \bar{Y} - a_1\frac{s}{\sqrt{n}}\right) = \alpha \tag{6.2.7}$$

so $(\bar{Y} - a_2 s/\sqrt{n}, \bar{Y} - a_1 s/\sqrt{n})$ is an α CI.

Example 6.2.1 Scores Y for an IQ test administered to ten year olds in a very large population have close to a Gaussian distribution $G(\mu, \sigma)$. A random sample of 10 children got test scores as follows:

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

We can 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 $P(-2.262 \leq t_{(9)} \leq 2.262) = .95$, for example, a $.95$ confidence interval for μ is $\bar{y} \pm 2.262 s/\sqrt{10}$. For the data given above $\bar{y} = 104.6$ and $s = 8.57$, so the observed confidence interval is 104.6 ± 6.13, or $98.47 \leq \mu \leq 110.73$.

Remarks:

1. Confidence intervals for μ get narrower as n increases. They are also narrower if σ is known (though this is unusual). In the limit as $n \to \infty$, the CI based on (6.2.6) is equivalent to using Z and knowing that $\sigma = s$. For example, if in Example 6.2.1 we know that $\sigma = 8.57$ then the .95 CI would use $\bar{y} \pm 1.96\sigma/\sqrt{n}$ instead of $\bar{y} \pm 2.262\ s/\sqrt{n}$ when $n = 10$. As n increases, $s/\sqrt{n}$ becomes arbitrarily close to zero so the CI's shrink to include only the point $\bar{y}$.

2. If we have a rough idea what the value of σ is, we can determine the value of n needed to make a (.95,say) CI a given length. This is used in deciding how large a sample to take in a study.

3. Sometimes we want CI's of the form $L(y_1, \ldots, y_n) \leq \mu$ or $U(y_1, \ldots, y_n) \geq \mu$. These are obtained by taking $a_1 = -\infty$ and $a_2 = \infty$, respectively, in (6.2.7). For "two-sided" intervals we usually pick $a_1 = -a_2$ so that the interval is symmetrical about $\bar{y}$.

Exercise: Show that these provide the shortest CI with a given confidence coefficient α.

Hypothesis Tests for μ

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 statistic

$$D = |T| = \frac{|\bar{Y} - \mu_0|}{s/\sqrt{n}} \tag{6.2.8}$$

Significance levels are obtained from the t-distribution: if D_{obs} is the value of D observed in a sample giving mean $\bar{y}$ and standard deviation s, then

$$\begin{aligned} SL &= P(D \geq D_{obs}; H_0 \text{ true}) \\ &= P(|t_{(n-1)}| \geq D_{obs}) \\ &= 1 - P(-D_{obs} \leq t_{(n-1)} \leq D_{obs}) \end{aligned}$$

Example 6.2.2 For the setting in Example 6.2.1, test $H_0 : \mu = 110$. With (6.2.8) the observed value of D is then

$$D_{obs} = \frac{|104.6 - 110|}{8.57/\sqrt{10}} = 1.99$$

and the significance level is

$$\begin{aligned} SL &= P(|t_{(9)}| \geq 1.99) \\ &= 1 - P(-1.99 \leq t_{(9)} \leq 1.99) \\ &= .078 \end{aligned}$$

This indicates there isn't any strong evidence against H_0. (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 (LR) statistic could also be used for testing $H_0 : \mu = \mu_0$ or for CI's for μ, but the methods above are a little simpler. In fact, it can be shown that the LR statistic for H_0 is a one-to-one function of $|T|$; see Problem 2 at the end of the chapter.

Remark: The function *t.test* in R will obtain confidence intervals and test hypotheses about μ; for a data set y use *t.test(y)*.

Confidence Intervals or Tests for σ

From the results following (6.2.6) we have that

$$U = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)} \tag{6.2.9}$$

so U is pivotal quantity and can be used to find CI's for σ^2 or σ. To get an α CI we find a_1, a_2 such that

$$P(a_1 \leq \chi^2_{(n-1)} \leq a_2) = \alpha = P\left(a_1 \leq \frac{(n-1)s^2}{\sigma^2} \leq a_2\right).$$

This converts to

$$P\left(\frac{(n-1)s^2}{a_2} \leq \sigma^2 \leq \frac{(n-1)s^2}{a_1}\right) = \alpha \tag{6.2.10}$$

so an α CI for σ is $\left(\sqrt{\frac{(n-1)s^2}{a_2}}, \sqrt{\frac{(n-1)s^2}{a_1}}\right)$. For "two-sided" CI's we usually choose a_1 and a_2 such that

$$P(\chi^2_{(n-1)} < a_1) = P(\chi^2_{(n-1)} > a_2) = \frac{1-\alpha}{2}$$

In some applications we are interested in an upper bound on σ (because small σ is "good" in some sense); then we take $a_2 = \infty$ so the lower confidence limit in (6.2.10) is 0.

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 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 .02 mm. On one such occasion the sample mean and sum of squares from the data were $\bar{y} = 25.009$ and $\sum(y_i - \bar{y})^2 = 0.002347$.

Consider getting a .95 confidence interval for σ, using the pivotal quantity

$$U = \frac{14s^2}{\sigma^2} = \frac{\sum(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{(14)}.$$

Since $P(5.63 \leq \chi^2_{(14)} \leq 26.12) = .95$, we find

$$P(5.63 \leq \frac{\sum(Y_i - \bar{Y})^2}{\sigma^2} \leq 26.12) = P\left(\left(\frac{\sum(Y_i - \bar{Y})^2}{26.12}\right)^{1/2} \leq \sigma \leq \left(\frac{\sum(Y_i - \bar{Y})^2}{5.63}\right)^{1/2}\right) = .95$$

This gives the observed confidence interval $.0095 \leq \sigma \leq .0204$. It seems plausible that $\sigma \leq .02$, though the right hand end of the .95 confidence interval is just over .02. A one-sided .95 CI is $\sigma \leq .0189$; this comes from $P(6.57 \leq \chi^2_{(14)} \leq \infty) = .95$

Hypothesis Tests for σ

Sometimes a test for $H_0 : \sigma = \sigma_0$ is of interest. One approach is to use a likelihood ratio (LR) statistic, as described in Chapter 4. It can be seen (see Problem 2) that the LR statistic Λ is a function of $U = (n-1)s^2/\sigma^2$,

$$\Lambda = U - n\,log\,(U/n) - n \tag{6.2.11}$$

This is not a one-to-one function of U but Λ is 0 when $U = n$ and is large when U/n is much bigger than or much less than 1 (i.e. when s^2/σ_0^2 is much bigger than one or much less than 1).

Since $U \sim \chi^2_{(n-1)}$, when H_0 is true, we can use this to compute exact significance levels, instead of using the $\chi^2_{(1)}$ approximation for Λ discussed in Chapter 4. The following simpler calculation approximates this SL:

1. Obtain $U_{obs} = (n-1)s^2_{obs}/\sigma_0^2$ from the observed data.
2. If $U_{obs} > n-1$ compute $SL = 2P(\chi^2_{(n-1)} \geq U_{obs})$.
   If $U_{obs} < n-1$ compute $SL = 2P(\chi^2_{(n-1)} \leq U_{obs})$.

Example 6.2.4 For the manufacturing process in Example 6.2.3, test the hypothesis $H_0 : \sigma = .008$ (.008 is the desired or target value of σ the manufacturer would like to achieve).

Note that since the value $\sigma = .008$ is outside the two-sided .95 CI for σ in Example 6.2.3, the SL for H_0 based on the test statistic Λ (or equivalently, U) will be less than .05. To find the exact SL, we follow the procedure above:

1. $U_{obs} = \frac{14s^2_{obs}}{.008^2} = 36.67$
2. $SL = 2P(\chi^2_{(14)} \geq 36.67) = .0017$

This indicates very strong evidence against H_0 and suggests that σ is bigger than .008.

6.3 General Gaussian Response Models

We now consider general models of the form (6.1.2) plus (6.1.3): $Y_i \sim G(\mu_i, \sigma)$ with $\mu_i = \sum_{j=1}^{k} \beta_j x_{ij}$ for independent units $i = 1, 2, \ldots, n$. For convenience we define the $n \times k$ matrix X of covariate values:

$$X = (x_{ij})_{n \times k} \tag{6.3.1}$$

We now summarize some results about the MLE's of the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ and σ.

Some results about M.L.E.'s of $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$ and of $\hat{\sigma}$

- Maximization of the likelihood function

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2}$$

or of $\ell(\boldsymbol{\beta}, \sigma) = \log L(\boldsymbol{\beta}, \sigma)$ gives

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}, \;\; \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

where $X_{n\times k} = (x_{ij})$, $\boldsymbol{y}_{n\times 1} = (y_1, ..., y_n)'$ and $\hat{\mu}_i = \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$. We also define

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{n}{n-k}\hat{\sigma}^2$$

- For the estimators $\tilde{\beta}_j$ $(j = 1, ..., k)$ and $\tilde{\sigma}^2$ it can be proved that

$$\tilde{\beta}_j \sim G(\beta_j, \sqrt{c_j\sigma^2}) \qquad j = 1, ..., k \tag{6.3.2}$$

$$W = \frac{n\tilde{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)} \tag{6.3.3}$$

$$W \text{ is independent of } (\tilde{\beta}_1, ..., \tilde{\beta}_k). \tag{6.3.4}$$

In (6.3.2), c_j are constants which are functions of the x_{ij}'s.

Remark: The MLE $\hat{\beta}$ is also a **least squares (LS) estimate** of β. Least squares is a method of estimation in linear models that predates maximum likelihood. Problem 16 describes least squares methods.

- Recall the distribution theory for the (student) t distribution. If $Z \sim G(0, 1)$ and $W \sim \chi^2_{(r)}$ then the r.v.

$$T = Z/\sqrt{W/r} \tag{6.3.5}$$

has a $t_{(r)}$ distribution.

This provides a way to get confidence intervals (or tests) for any β_j. Because (6.3.2) implies that

$$Z = \frac{\tilde{\beta}_j - \beta_j}{\sigma\sqrt{c_j}} \sim G(0, 1)$$

and because of (6.3.3), then (6.3.5) implies that

$$T = \frac{\tilde{\beta}_j - \beta_j}{s\sqrt{c_j}} \sim t_{(n-k)} \tag{6.3.6}$$

so T is a pivotal quantity for β_j. In addition W given in (6.3.3) is a pivotal quantity for σ^2 or σ.

Below we will consider some special types of Gaussian models which fall under the general theory. However, we'll alter the notation a bit for each model, for convenience.

Single Gaussian distribution

Here, $Y_i \sim G(\mu, \sigma)$ $i = 1, ..., n$ **i.e.** $\mu_i = \mu$ (we use μ instead of β as parameter name). This model was discussed in detail in Section 6.2, where we used

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}, \;\; W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

(**Note**: The c_1 in (6.3.2) is $1/n$ here; its easiest to get this by the fact that Var$(\overline{Y})$ = var$(\tilde{\mu}) = \sigma^2/n$, proved earlier.)

Comparing Two Gaussian Distributions $G(\mu_1,\sigma)$ and $G(\mu_2,\sigma)$

Independent samples $y_{11}, y_{12}, ..., y_{1n_1}$ from $G(\mu_1, \sigma)$
$y_{21}, y_{22}, ..., y_{2n_2}$ from $G(\mu_2, \sigma)$ are obtained.

(We use double subscripts for the Y's here, for convenience.)

Once again, stick with μ_1 and μ_2 as names of parameters. The likelihood function for μ_1, μ_2, σ is

$$L(\mu_1, \mu_2, \sigma) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_{ji} - \mu_j}{\sigma})^2}$$

Maximization gives the m.l.e.'s

$$\hat{\mu}_1 = \sum_{i=1}^{n_1} \frac{y_{1i}}{n_1} = \bar{y}_1, \;\; \hat{\mu}_2 = \sum_{i=1}^{n_2} \frac{y_{2i}}{n_2} = \bar{y}_2, \;\; \hat{\sigma}^2 = \frac{1}{n_1 + n_2} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{ji} - \hat{\mu}_j)^2$$

Note that $s^2 = \frac{1}{n_1+n_2-2} \sum_{j=1}^{2} (n_j - 1)s_j^2 = (n_1 + n_2)\hat{\sigma}^2/(n_1 + n_2 - 2)$

where $s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2$.

- To get CI's for $\beta = \mu_1 - \mu_2$ note that

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - \beta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim G(0, 1), \;\; W = \frac{(n_1 + n_2 - 2)s^2}{\sigma^2} \sim \chi^2_{(n_1+n_2-2)}$$

and so by (6.3.5)

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - \beta}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}. \tag{6.3.7}$$

Confidence intervals or tests about σ can be obtained by using the pivotal quantity W 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″ wide) of each paint were laid across a heavily traveled section of highway, in random order. After a period of time, reflectometer readings were taken for each line of paint; the higher the readings the greater the reflectivity and the visibility of the paint. The measurements of reflectivity were as follows:

Paint A: 12.5, 11.7, 9.9, 9.6, 10.3, 9.6, 9.4, 11.3, 8.7, 11.5, 10.6, 9.7
Paint B: 9.4, 11.6, 9.7, 10.4, 6.9, 7.3, 8.4, 7.2, 7.0, 8.2, 12.7, 9.2

Statistical objectives are to test that the average reflectivity for paints A and B is 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, reflectivity measurements Y_1 for paint A are $G(\mu_1, \sigma_1)$, and measurements Y_2 for paint B are $G(\mu_2, \sigma_2)$. We can test $H: \ \beta = \mu_1 - \mu_2 = 0$ and get confidence intervals for β by using the pivotal quantity

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - \beta}{s\sqrt{\frac{1}{12} + \frac{1}{12}}} \sim t_{(22)}$$

where it is assumed that $\sigma_1 = \sigma_2 = \sigma$, which is estimated by

$$s^2 = \frac{1}{22}\left[\sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2\right].$$

To test $H: \ \beta = 0$ we use the test statistic $D = |T|$. From the data given above we find

$$\begin{array}{llll} n_1 = 12 & \bar{y}_1 = 10.4 & \sum(y_{1i} - \bar{y}_1)^2 = 14.08 & s_1^2 = 1.2800 \\ n_2 = 12 & \bar{y}_2 = 9.0 & \sum(y_{2i} - \bar{y}_2)^2 = 38.64 & s_2^2 = 3.5127. \end{array}$$

This gives $\hat{\beta} = 1.4$ and $s^2 = 2.3964$, and the observed test statistic $d_{obs} = 2.22$. The significance level is then

$$SL = P(|t_{(22)}| \geq 2.22) \doteq .038.$$

This indicates there is fairly strong evidence against $H: \ \mu_1 = \mu_2$. Since $\bar{y}_1 > \bar{y}_2$, the indication is that paint A keeps its visibility better. A .95 confidence interval based on T is obtained using the fact that $P(-2.074 \leq t_{(22)} \leq 2.074) = .95$. This gives the confidence interval for $\beta = \mu_1 - \mu_2$ of $\hat{\beta} \pm 2.074 s/\sqrt{6}$, or $0.09 \leq \beta \leq 2.71$. 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 μ_1 and μ_2 (look at $\bar{y}_1$ and $\bar{y}_2$).

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 LR test, but we won't 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 exactly pivotal quantity with which to get CI's for $\beta = \mu_1 - \mu_2$, but

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - \beta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \simeq G(0,1) \tag{6.3.8}$$

is an approximate pivotal quantity that becomes exact as n_1 and n_2 become arbitrarily large.

To illustrate its use, consider Example 6.3.1, where we had $s_1^2 = 1.2800, s_2^2 = 3.5127$. 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 LR test is carried out. Nevertheless, let us use (6.3.8) to get a .95 CI for β. This gives the CI

$$(\bar{Y}_1 - \bar{Y}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

which with the data observed equals 1.4 ± 1.24, or $.16 \leq \beta \leq 2.64$. This is not much different than the interval obtained in Example 6.3.1.

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 others 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 A:	$n = 278$	$\bar{y} = 60.2$	$s = 10.16$
District B:	$n = 345$	$\bar{y} = 58.1$	$s = 9.02$

The average score is somewhat higher in district A, and we will give a confidence interval for the difference in average scores $\mu_A - \mu_B$ in a model representing this setting. This is done by thinking of the students in each district as a random sample from a conceptual 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 the means μ_A and μ_B for districts A and B are the same, or give a CI for the difference. (Achievement tests are usually designed so that the scores are approximately Gaussian, so this is a sensible procedure.)

Let us get a .95 CI for $\beta = \mu_A - \mu_B$ using the pivotal quantity (6.3.8). This gives the CI

$$\bar{y}_1 - \bar{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

which becomes $2.1 \pm (1.96)(.779)$ or $0.57 \leq \beta \leq 1.63$. Since $\beta = 0$ is outside the .95 CI (and also the .99 CI) we can conclude there is fairly strong evidence against the hypothesis that $\mu_A = \mu_B$, suggesting that $\mu_A > \mu_B$.

It is always a good idea to look carefully at the data and the distributions suggested for the two groups, however; we should not rely only on a comparison of their means. Figure 6.3.1 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 6.19 was obtained using the R function $boxplot()$.

Note that the distributions of marks for districts A and B are actually quite similar. The median (and mean) is a little higher for district A and because the sample sizes are so large, this gives a “statistically significant" difference in a test that $\mu_A = \mu_B$. However, it would be a mistake 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 being “superior" to another.

Remark: The R function *t.test* will carry out the test above and will give confidence intervals for $\mu_A - \mu_B$. This can be done with the command $t.test(y_A, y_B, var.equal = T)$, where y_A and y_B are the data vectors from A and B.

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 we give them their own 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.

241 notes Lawless/st241.fig631.ps 241 notes Lawless/st241.fig631.ps

Figure 6.19: **Box Plot of Math Test Scores for Two School Districts.**

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.

Let Y_{1i} and Y_{2i} be the heights of the male and female, respectively , in the i'th (brother, sister) pair $(i = 1, 2, \ldots, 1401)$. Assuming that the pairs are sampled randomly from the population, we can use them to estimate

$$\mu_1 = E(Y_{1i}) \qquad\qquad \mu_2 = E(Y_{2i})$$

and the difference $\beta = \mu_1 - \mu_2$. However, the heights of related persons are not independent, so to estimate β the method in the preceding section would not be strictly usable; 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. This topic is not considered in this course, but a clear picture 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

$$\beta = E(Y_{1i} - Y_{2i}).$$

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 likely 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. 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. Only discrete bivariate models were introduced in Stat 230 and we will not consider this model here (it is studied in Stat 330), except to note an important property:

$$Y_i = Y_{1i} - Y_{2i} \sim G(\beta, \sigma^2) \tag{6.4.1}$$

where $\beta = \mu_1 - \mu_2 = E(Y_{1i}) - E(Y_{2i})$. Thus, if we are interested in estimating or testing β, we can do this by considering the **within-pair differences** Y_i and using the methods for a single Gaussian model in Section 6.2.

The second Gaussian model used with paired data has

$$Y_{1i} \sim G(\mu_1 + \alpha_i, \sigma_1^2) \qquad\qquad Y_{2i} \sim G(\mu_2 + \alpha_i, \sigma_2^2)$$

where the α_i's are unknown constants. Here it is assumed that Y_{1i} and Y_{2i} are independent r.v.'s, and the α_i's represent factors specific to the different pairs. This model also gives the distribution (6.4.1), since

$$E(Y_{1i} - Y_{2i}) = \mu_1 - \mu_2 \ \ (\alpha_i \text{ cancels})$$
$$Var(Y_{1i} - Y_{2i}) = \sigma_1^2 + \sigma_2^2 = \sigma^2.$$

This model seems relevant for Example 6.4.2, where α_i refers to the i'th car type. Interestingly, the two models for (Y_{1i}, Y_{2i}) can be connected; if the α_i's are considered as Gaussian random variables in the population of pairs of units then the result is that (Y_{1i}, Y_{2i}) have a bivariate normal model.

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 $\beta = \mu_1 - \mu_2$ by working with the model (6.4.1).

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

$$\bar{y} = 4.895 \text{ in} \qquad\qquad s^2 = \frac{\sum(y_i - \bar{y})^2}{1400} = 6.5480 \text{ in}^2$$

Using the student$-t$ pivotal quantity (6.2.6), a two-sided .95 confidence interval for $\beta = E(Y_i)$ is $\bar{y} \pm 1.96s/\sqrt{n}$ where $n = 1401$. (Note that $t_{(1400)}$ is indistinguishable from $G(0, 1)$.) This gives the .95 CI 4.895 ± 0.134 inches, or $4.76 \leq \beta \leq 5.03$ in.

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 β 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 β does represent this difference. However, it might be wise to check this assumption.

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 $\hat{\beta} = \bar{y}_1 - \bar{y}_2 = 68.72 - 64.10 = 4.62$ 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 were the population can be arranged in pairs, the estimation of a difference in means, $\beta = \mu_1 - \mu_2$, can often be made more precise (shorter CI's) 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 of 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 the pivotal quantity (6.3.7) to get a confidence interval for $\beta = \mu_1 - \mu_2$. This gives the .95 CI $4.70 \leq \beta \leq 5.09$; we note that it is slightly longer than the .95 CI $4.76 \leq \beta \leq 5.03$ obtained using the pairings.

To see why the pairing is helpful in estimating β, 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 β is

$$\tilde{\beta} = \bar{Y}_1 - \bar{Y}_2$$

and we have that $E(\tilde{\beta}) = \beta = \mu_1 - \mu_2$ and

$$\begin{aligned} Var(\tilde{\beta}) &= Var(\bar{Y}_1) + Var(\bar{Y}_2) - 2Cov(\bar{Y}_1, \bar{Y}_2) \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2}{n}\sigma_{12}, \end{aligned}$$

where $\sigma_{12} = Cov(Y_{1i}, Y_{2i})$. If $\sigma_{12} > 0$, then $Var(\tilde{\beta})$ is **smaller** than when $\sigma_{12} = 0$ (i.e. when Y_{1i} and Y_{2i} are independent). 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 β. Note on the other hand that if $\sigma_{12} < 0$, then pairing is a bad idea since it increases the variance of $\tilde{\beta}$.

The following example involves an experimental study with pairing.

Example 6.4.3. Fibre in Diet and Cholesterol Level

This example comes from the Stat 231 Course Notes, chapter 15. In the study 20 subjects (who were actually volunteers from workers in a Boston hospital) with ordinary cholesterol levels were given a low-fibre diet for 6 weeks and a high-fibre diet for another 6 week period. The order in which the two diets were given was randomized for each subject (person), and there was a two-week gap between the two 6 week periods, in which no dietary fibre supplements were given. A primary objective of the study was to see if cholesterol levels are lower with the high-fibre diet.

Details of the study are given in the **New England Journal of Medicine**, volume 322 (January 18, 1990), pages 147-152, and in the Stat 231 Notes. These provide interesting comments on factors and difficulties in the design of studies on the effects of diet. 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'll let the r.v.'s 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

$$Y_i = Y_{1i} - Y_{2i} \sim G(\beta, \sigma). \qquad i = 1, \ldots, 20$$

The differences y_i are also shown in Table 6.4.1, and from them we calculate the sample mean and standard deviation

$$\bar{y} = -.020 \qquad\qquad s = 0.411$$

A .95 CI for β is found using the pivotal quantity (6.2.7) and the fact that $P(-2.093 \leq t_{(19)} \leq 2.093) = .95$. This gives the CI $\bar{y} \pm 2.093\, s/\sqrt{20}$, or $-.020 \pm .192$, or

$$-.212 \leq \beta \leq .172$$

This confidence interval includes $\beta = 0$, and there is clearly no evidence that the high fibre diet gives a lower cholesterol level.

Remark: The results here can be obtained using the R function $t.test$.

Exercise: Compute the significance level of the hypothesis $H_0 : \beta = 0$, using the test statistic (6.2.8).

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	.13	11	4.44	4.43	.01
2	2.91	2.85	.06	12	5.22	5.27	-.05
3	4.77	4.25	.52	13	4.22	3.61	.61
4	5.63	5.43	.20	14	4.29	4.65	-.36
5	3.58	4.38	-.80	15	4.03	4.33	-.30
6	5.11	5.05	.06	16	4.55	4.61	-.06
7	4.29	4.44	-.15	17	4.56	4.45	.11
8	3.40	3.36	.04	18	4.67	4.95	-.28
9	4.18	4.38	-.20	19	3.55	4.41	-.86
10	5.41	4.55	.86	20	4.44	4.38	.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 forms 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 $\boldsymbol{x}$, as described in Section 6.1. In this section we consider settings where there is a single x-variable. Problems with multiple x-variables were mentioned in Sections 6.1 and 6.3, and are considered in Stat 331. We start by summarizing results from Sections 6.1 and 6.3. Consider the model with independent Y_i's such that

$$Y_i \sim G(\mu_i, \sigma) \text{ with } \mu_i = \alpha + \beta x_i \tag{6.5.1}$$

(Note that this is of the form (6.1.2) and (6.1.3) with $\beta_1 = \alpha,\ \beta_2 = \beta,\ x_{1i} = 1,\ x_{2i} = x_i$).

- Once again, we can use the general results of Section 6.3 or just maximize the likelihood to get the MLE's: maximize
$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2}$$
to get $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}(S_{yy} - \hat{\beta}S_{xy})$
where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

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 $\sum x_i'$ close to zero. (Often we take $c = \bar{x}$, which makes $\sum x_i'$ 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 $E(Y|x)$is

$$\mu(x) = \alpha + \beta x = \alpha + \beta(x' + c) = (\alpha + \beta c) + \beta x'.$$

Thus, the intercept α changes if we redefine x, but not β. In the examples here we have kept the given definition of x, for simplicity.

We now consider how to get confidence intervals for quantities of interest. As usual, formulas are written in terms of

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \end{aligned}$$

instead of $\hat{\sigma}^2$.

Confidence Intervals for β

These are important because β represents the increase in

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

resulting from an increase of 1 in x. As well, if $\beta = 0$ then x has no effect on Y (within the constraints of this model).

From Section 6.3 we know that $\tilde{\beta} \sim G(\beta, \sqrt{c}\sigma)$ for some constant c. It's easy to show this directly here, and to obtain c. Write $\tilde{\beta}$ as

$$\begin{aligned}
\tilde{\beta} = \frac{S_{xy}}{s_{xx}} &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})(Y_i - \bar{Y})}{s_{xx}} \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{xx}} Y_i \qquad (\text{since } \sum (x_i - \bar{x})\bar{Y} = 0) \\
&= \sum_{i=1}^{n} a_i Y_i
\end{aligned}$$

where $a_i = (x_i - \bar{x})/s_{xx}$. This is a linear combination of independent Gaussian r.v.'s and so its distribution is also Gaussian, with

$$\begin{aligned}
E(\tilde{\beta}) &= \sum_{i=1}^{n} a_i E(Y_i) \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{xx}} (\alpha + \beta x_i) \\
&= \alpha \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{xx}} + \beta \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{xx}} x_i \\
&= 0 + \beta \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s_{xx}} \qquad (\text{since} \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) x_i) \\
&= \beta
\end{aligned}$$

and

$$\begin{aligned}
Var(\tilde{\beta}) &= \sum_{i=1}^{n} a_i^2 Var(Y_i) \\
&= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s_{xx}^2} \sigma^2 = \frac{\sigma^2}{s_{xx}}
\end{aligned}$$

Thus

$$\tilde{\beta} \sim G(\beta, \frac{\sigma}{\sqrt{s_{xx}}}) \tag{6.5.2}$$

and combining this with the fact that

$$W = \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{(n-2)} \tag{6.5.3}$$

and that $\tilde{\beta}$ and s^2 are independent, we get as in (6.3.6) that

$$T = \frac{\tilde{\beta} - \beta}{s/\sqrt{s_{xx}}} \sim t_{(n-2)} \tag{6.5.4}$$

This can be used as a pivotal quantity to get CI's for β, or to test hypotheses about β.

Note also that (6.5.3) can be used to get CI's or tests for σ, but these are usually of less interest than inference about β or the other quantities below.

Confidence Intervals for $\mu(x)$

We are often interested in estimating $\mu(x) = \alpha + \beta x$ for a specified value of x. We'll derive a student -t pivotal quantity for doing this.

The MLE of $\mu(x)$ has associated estimator

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \bar{Y} + \tilde{\beta}(x - \bar{x}),$$

since $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$. Thus $\mu(x)$ is a linear function of Gaussian r.v.'s (because $\bar{Y}$ and $\tilde{\beta}$ are) and so must 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 because $\bar{Y}$ and $\tilde{\beta}$ are independent (can be shown),

$$\begin{aligned} Var[\tilde{\mu}(x)] &= Var(\bar{Y}) + (x - \bar{x})^2 Var(\tilde{\beta}) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) + (x - \bar{x})^2 \frac{\sigma^2}{s_{xx}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}} \right) \end{aligned}$$

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

$$T = \frac{\tilde{\mu}(x) - \mu(x)}{s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}} \sim t_{(n-2)} \tag{6.5.5}$$

This can be used as a pivotal quantity to get CI's for $\mu(x)$.

Remark: The parameter α equals $\mu(0)$ so doesn't require special treatment. Sometimes $x = 0$ is a value of interest but often it is not. In the following example it refers to a building of area 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, $summary(lm(y \sim x))$ 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 model of the form (6.5.1) would be reasonable, so let us consider that.

We find easily that $\bar{x} = 0.954,\ \bar{y} = 549.0$ and $s_{xx} = 22.945,\ s_{xy} = -3316.68,\ s_{yy} = 489,462.62$ so we find

$$\hat{\beta} = -144.5,\ \hat{\alpha} = 686.9,\ s^2 = 364.37,\ s = 19.09.$$

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 1 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, and if we plot it on the same graph as the points (x_i, y_i) in the scatter plot Figure 6.1.1, we see it passes close to the points.

A confidence interval for β isn't 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 MLE

$$\hat{\mu}(4.47) = \hat{\alpha} + \hat{\beta}(4.47) = \$40.94$$

which we note is much below the assessed value of $\$75$ per square meter. However, one can object that there is uncertainty in $\hat{\mu}(4.47)$, and that it would be better to give a CI. Using (6.5.5) and the fact that

$P(-2.048 \leq t_{(28)} \leq 2.048) = .95$, we get a .95 CI for $\mu(4.47)$ as

$$\hat{\mu}(4.47) \pm 2.048 s \sqrt{\frac{1}{30} + \frac{(4.47 - \bar{x})^2}{s_{xx}}}$$

or $\$40.94 \pm \$26.54$, 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: since we are considering a **single** building (and not the average of all buildings) of size $x = 4.47(\times 10^5)m^2$, we must recognize that Y_i 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 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 $Y \sim G(\mu(x), \sigma)$ and, from above, that $\tilde{\mu}(x) \sim G(\mu(x),\ \sigma \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}\right)^{1/2})$. Also, Y is independent of $\tilde{\mu}(x)$ since it is not connected to the existing sample. Thus

$$\begin{aligned} Var(Y - \tilde{\mu}(x)) &= Var(Y) + Var(\tilde{\mu}(x)) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}\right) \end{aligned}$$

Thus

$$Y - \tilde{\mu}(x) \sim G(0, \sigma \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}\right)^{1/2})$$

and it also follows that

$$T = \frac{Y - \tilde{\mu}(x)}{s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}} \sim t_{(n-2)} \tag{6.5.6}$$

We can use the pivotal quantity T to get interval estimates for Y, since

$$\begin{aligned} &P(a_1 \leq T \leq a_2) \\ &= P\left(\tilde{\mu}(x) - a_2 s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}} \leq Y \leq \tilde{\mu}(x) - a_1 s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}}\right) \end{aligned}$$

We usually call these **prediction intervals** instead of confidence intervals, since Y isn't a parameter but a "future" observation.

Example 6.5.1 Revisited

Let us obtain a .95 prediction interval (PI) for Y when $x = 4.47$. Using (6.5.6) and the fact that $P(-2.048 \leq t_{(28)} \leq 2.048) = .95$, we get the .95 PI

$$\tilde{\mu}(4.47) \pm 2.048 s \sqrt{1 + \frac{1}{30} + \frac{(4.47 - \bar{x})^2}{s_{xx}}}$$

or $-6.30 \leq Y \leq 88.20$ (dollars per square meter). The lower limit is negative, which is nonsensical. This happened because we're using a Gaussian model (in which Y can be positive or negative) in a setting where Y (price) must be positive. Nonetheless, the Gaussian model fits the data well, so we'll just truncate the PI 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 PI).

Note also that the value $x = 4.47$ of interest is well outside the range of x-values (.20 - 3.26) in the data set of 30 buildings; look again at Figure 6.1.1. Thus any conclusions we reach are based on an assumption that the linear model $\mu(x) = \alpha + \beta x$ applies beyond $x = 3.26$ and out to $x = 4.47$. This 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 6.1.1 that $Var(YX)$ may be smaller for larger x-values. There is not sufficient data to check this either.

Remark: Note from (6.5.5) and (6.5.6) that CI's for $\mu(x)$ and PI's from 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 $\mu(x)$ or 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 scatter plot of points (x_i, y_i) for 30 bolts suggested a nonlinear relationship between Y and x. A bolt's strength might be expected to be proportional to its cross-sectional area, which is proportional to x^2. Figure 6.5.1 shows a plot of points (x_i^2, y_i); it looks quite linear.

Let us fit a linear model

$$Y_i \sim G(\alpha + \beta x_{1i}, \sigma) \qquad\qquad x_{1i} = x_i^2$$

to the data. We find (check these for yourself)

$$\hat{\alpha} = 1.667,\ \hat{\beta} = 2.838,\ s = 0.0515,\ S_{xx} = 0.2244$$

The fitted regression line $y = \hat{\alpha} + \hat{\beta} x_1$ is shown on the scatter plot in Figure 6.20; the model appears to fit well.

More as a numerical illustration, let us get a CI for β, which represents the increase in average strength $\mu(x_1)$ from increasing the diameter x (and therefore also $x_1 = x^2$) by 1 unit. Using the pivotal quantity (6.5.4) and the fact that $P(-2.048 \leq t_{(28)} \leq 2.048) = .95$, we get the .95 CI

$$\hat{\beta} \pm 2.048 \frac{s}{\sqrt{s_{xx}}}, \text{ or } 2.838 \pm 0.223$$

A .95 CI for the value of β is therefore (2.605, 3.051).

Figure 6.5.1 Scatter Plot of Bolt Diameter Squared vs. Strength

file=st241.fig651.ps,angle=0,width=

241 notes Lawless/st241.fig651.ps 241 notes Lawless/st241.fig651.ps

Figure 6.20: **Scatter Plot of Bolt Diameter Squared vs. Strength**

Exercise: This model could be used to predict the breaking strength of a new bolt of given diameter x. Find a PI for a new bolt of diameter $x = 0.35$.

6.6 Model Checking

There are two main components in Gaussian response models:

(i) the assumption that Y_i (given any covariates x_i) is Gaussian with constant standard deviation σ.

(ii) the assumption that $\mu_i = g(x_i)$ is a given form like (6.1.3).

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

$$\hat{r}_i = y_i - \hat{\mu}_i \qquad\qquad i = 1, \ldots, n$$

For example, if $Y_i \sim G(\alpha + \beta x_i; \sigma)$ then $\hat{r}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. 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 $G(0, \sigma)$.

Plots of residuals are 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 within a horizontal 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) 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 $\hat{\mu}_i$, this suggests that $Var(Y_i) = Var(R_i)$ is not constant, but may be larger when $\mu(x)$ is larger.

Figure 6.6.1 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.

In problems with only one x-variable, a plot of $\hat{\mu}(x)$ superimposed on the scatterplot of the data (as in Figure 6.5.1) 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 MLE 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 c.d.f. (EDF) from $\hat{r}_i^*(i = 1, \ldots, n)$ and superimpose on it the $G(0, 1)$ c.d.f. 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

$$\hat{r}_i^* = \frac{\hat{r}_i}{\hat{\sigma}} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}} \qquad\qquad i = 1, \ldots, n$$

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 6.6.1. When the $\hat{r}_i^*$'s are used the patterns in the plot are unchanged but the $\hat{r}_i^*$ values tend to lie in the range (-3,3). (think why this is so.)

Figure 6.6.1 Examples of Patterns in Residual Plots file=st241.fig661.ps,angle=0,width=

241 notes Lawless/st241.fig661.ps 241 notes Lawless/st241.fig661.ps

Figure 6.21: **Examples of Patterns in Residual Plots**

Example 6.6.1 Steel Bolts Let us define residuals

$$\hat{r}_i = y_i - \hat{\alpha} - \hat{\beta}x_i \qquad\qquad i = 1, \ldots, 30$$

for the model fitted in Example 6.5.3. Figure 6.22 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 6.5.1.

A further check on the Gaussian distribution is shown in Figure 6.23. Here we have plotted the EDF based on the 30 standardized residuals

$$\hat{r}_i^* = \frac{y_i - \hat{\alpha} - \hat{\beta} x_{1i}}{\hat{\sigma}}.$$

On the same graph is the $G(0,1)$ c.d.f. There is good agreement between the two curves.

241 notes Lawless/st241.fig662.ps 241 notes Lawless/st241.fig662.ps

Figure 6.22: **Residual Plot for Bolt Strength Mode**l

6.7 Problems

1. Student's t Distribution
   Suppose that Z and U are independent variates with

$$Z \sim N(0,1); \;\; U \sim \chi^2_{(\nu)} \, .$$

   Consider the ratio

$$X \equiv \frac{Z}{\sqrt{U \div \nu}} \, .$$

241 notes Lawless/st241.fig663.ps 241 notes Lawless/st241.fig663.ps

Figure 6.23: **EDF of Standard Residuals and** $G(0,1)$ **CDF**

Then X is a continuous variate. Its distribution is called the t (Student's) distribution with ν degrees of freedom, and we write $X \sim t_{(\nu)}$ for short. It can be shown by change of variables that X has pdf

$$f(x) = k_\nu \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for} \quad -\infty < x < \infty$$

where k_ν is a normalizing constant such that the total area under the pdf is 1:

$$k_\nu = \Gamma\left(\frac{\nu+1}{2}\right) / \sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right) \ .$$

The pdf is symmetric about the origin, and is similar in shape to the pdf of $N(0,1)$ but has more probability in the tails. It can be shown that $f(x)$ tends to the $N(0,1)$ pdf as $\nu \to \infty$.

(a) Plot the pdf for $\nu = 1$ and $\nu = 5$.

(b) Find values a, b such that

$$P\left(-a \le t_{(30)} \le a\right) = 0.98; \quad P\left(t_{(20)} \ge b\right) = 0.95 \ .$$

(c) Show that $f(x)$ is unimodal for all x.

(d) Show that as $\nu \to \infty$, $f(x)$ approaches the $G(0,1)$ limit $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.
(Note: To do this you will need to use the fact that $k_\nu \to 1/\sqrt{2\pi}$ as $\nu \to \infty$; this is from a property of gamma functions.)

2. Suppose that $Y_1, \cdots, Y_n$ 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 $T = \sqrt{n}(\overline{Y} - \mu)/s$, with s the sample standard deviation. (Note: The sample variance s^2 is defined as $\sum(y_i - \bar{y})^2/(n-1)$.)

(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 $N(\mu, \sigma^2)$, 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 $N(\mu, \sigma^2)$.

(a) An old method produced measurements with standard deviation $\sigma = .02$. 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% CI's 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 $N(\mu, \sigma^2)$ random variables, obtain .95 confidence intervals for μ and σ.

(b) Let $\bar{X}$ and $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ be the mean and variance in a sample of size n, and let X represent the weight of a future, independent, randomly selected package. Show that $X - \bar{X} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$ and then that

$$Z = \frac{X - \bar{X}}{s\sqrt{1 + \frac{1}{n}}} \sim t_{(n-1)} \, .$$

For the data above, use this as a pivotal to obtain a .95 "confidence" interval for X.

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 $\sum (y_i - \bar{y})^2 = 80$, 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:	$n = 30$	$\bar{x} = 120$	$\sum (x_i - \bar{x})^2 = 3050$
Engineering students:	$n = 30$	$\bar{y} = 114$	$\sum (y_i - \bar{y})^2 = 2937$

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, and 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 $Y_1, \cdots, Y_n$ be a random sample from $G(\mu_1, \sigma_1)$ and $X_1, \cdots, X_n$ be a random sample from $G(\mu_2, \sigma_2)$. Obtain the likelihood ratio statistic for testing $H : \sigma_1 = \sigma_2$ 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 m.l.e.), 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 CI's 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 mean features.

    (b) Estimate (give a CI) 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 CI for p

using this approach. (It is also possible to get a CI 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 $\boldsymbol{x}_i$ has the form

$$\mu_i = E(Y_i|\boldsymbol{x}_i) = g(\boldsymbol{x}_i;\boldsymbol{\beta}) \tag{6.7.1}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. Then the **least squares (LS) estimate** of $\boldsymbol{\beta}$ based on data $(\boldsymbol{x}_i, y_i),\ i = 1,\ldots,n$ is the value that minimizes the objective function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i - g(\boldsymbol{x}_i;\boldsymbol{\beta})]^2$$

Show that the LS estimate $\hat{\beta}$ of $\boldsymbol{\beta}$ is the same as the MLE of $\boldsymbol{\beta}$ in the Gaussian model $Y_i \sim G(\mu_i, \sigma)$, when μ_i is of the form (6.7.1).

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 vs. 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 $y = \log$(Pressure) vs. x =Temperature. Do the same for y =Pressure vs. x. Which is better described by a linear model? Does this confirm the theory's model?

(b) Fit a normal linear regression model for $y = \log$(Pressure) vs. 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.

20. Consider the data in Problem 9 of Chapter 1, in which the variable y was the average time to complete tasks by computer users, and x was the number of users on the system. Fit a regression model, using x as the explanatory variable. Give a .95 confidence interval for the mean of Y when there are 50 users on the system.

21. (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 r.v. 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 $V \sim G(1.60, .10)$, where V and Y are independent.

22. **Optimal Prediction**. In many settings we want to use covariates x to predict a future value Y. (For example, we use economic factors x to predict the price Y of a commodity a month from now.) The value Y is random, but suppose we know $\mu(x) = E(Y|x)$ and $\sigma(x)^2 = Var(Y|x)$.

    (a) Predictions take the form $\hat{Y} = g(x)$, where $g(\cdot)$ is our "prediction" function. Show that the minimum achievable value of $E(\hat{Y} - Y)^2$ is minimized by choosing $g(x) = \mu(x)$.

    (b) Show that the minimum achievable value of $E(\hat{Y} - Y)^2$, that is, its value when $g(x) = \mu(x)$ is $\sigma(x)^2$.
    This shows that if we can determine or estimate $\mu(x)$, then "optimal" prediction (in terms of Euclidean distance) is possible. Part (b) shows that we should try to find covariates x for which $\sigma(x)^2 = Var(Y|x)$ is as small as possible.

    (c) What happens when $\sigma(x)^2$ is close to zero? (Explain this in ordinary English.)

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 Example (d) of Chapter 5, Section 2, where data were assumed to arise from a multinomial distribution with probability function

$$f(y_1, ..., y_m; p_1, ..., p_m) = \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m} \tag{6.5.2}$$

where $0 \leq y_j \leq n$ and $\sum y_j = n$. The multinomial probabilities p_j satisfy $0 \leq p_j \leq 1$ and $\sum p_j = 1$, and we define $\boldsymbol{\theta} = (p_1, ..., p_m)$. Suppose now that we wish to test the hypothesis

$$H_0 : \quad p_j = p_j(\boldsymbol{\alpha}) \qquad j = 1, ..., m \tag{6.5.3}$$

where $\dim(\boldsymbol{\alpha}) = p < m - 1$.

The likelihood function based on (7.1.1) is anything proportional to

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} p_j^{y_j}. \tag{6.5.4}$$

Let Ω be the parameter space for $\boldsymbol{\theta}$. It was shown earlier that $L(\boldsymbol{\theta})$ is maximized over Ω by the vector $\hat{\boldsymbol{\theta}}$ with $\hat{p}_j = y_j/n$ $(j = 1, ..., m)$. A likelihood ratio test of the hypothesis (7.1.2) is based on the likelihood ratio statistic

$$\Lambda = 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\hat{\boldsymbol{\theta}}_0) = -2 \log \left\{ \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})} \right\}, \tag{6.5.5}$$

where $\hat{\boldsymbol{\theta}}_0$ maximizes $L(\boldsymbol{\theta})$ under the hypothesis (7.1.2), which restricts $\boldsymbol{\theta}$ to lie in a space $\Omega_0 \subset \Omega$ of dimension p. If H_0 is true (that is, if $\boldsymbol{\theta}$ really lies in Ω_0) then the distribution of Λ is approximately

$\chi^2_{(k-p)}$ when n is large, where $k = m-1$ is the dimension of Ω. This enables us to compute significance levels (p-values) from observed data by using the approximation

$$P(\Lambda \geq \Lambda_{\text{obs}}; H_0) \doteq P(\chi^2_{(k-p)} \geq \Lambda_{\text{obs}}). \tag{6.5.6}$$

This approximation is very accurate when n is large and none of the p_j's is too small; when the e_j's below all exceed 5 it is accurate enough for testing purposes.

The test statistic (7.1.4) can be written in a simple form. Let $\hat{\boldsymbol{\theta}}_0 = (\tilde{p}_1, ..., \tilde{p}_m) = (p_1(\tilde{\alpha}), ..., p_m(\tilde{\alpha}))$ denote the m.l.e. of $\boldsymbol{\theta}$ under the hypothesis (7.1.2). Then, by (7.1.3), we get

$$\begin{aligned} \Lambda &= 2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\hat{\theta}_0) \\ &= 2\sum_{j=1}^{m} y_j \log(\hat{p}_j/\tilde{p}_j). \end{aligned}$$

Noting that $\hat{p}_j = y_j/n$ and defining "expected frequencies" under H_0 as

$$e_j = n\tilde{p}_j, \qquad j = 1, ..., m$$

we can rewrite Λ as

$$\Lambda = 2\sum_{j=1}^{m} y_j \log(y_j/e_j). \tag{6.5.7}$$

An alternative test statistic that was developed historically before Λ is the "Pearson" statistic

$$D = \sum_{j=1}^{m} \frac{(y_j - e_j)^2}{e_j}. \tag{6.5.8}$$

This has similar properties to Λ; for example, both equal 0 when $y_j = e_j$ 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 $\chi^2_{(k-p)}$ distribution, with $k = m - 1$, 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 2.5 that one way to check the fit of a probability distribution is by comparing the p_j's (relative frequencies) with the estimates $\tilde{p}_j$ from the distributional model. This is equivalent to comparing the y_j's (observed frequencies) and the e_j's. In Section 2.5 this comparison was informal, with only a rough guideline for how closely the y_j's and e_j's (or $\hat{p}_j$'s and $\tilde{p}_j$'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 2.5.2, 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 p_1, p_2, p_3 respectively, with $p_1 + p_2 + p_3 = 1$. Genetic theory indicates, however, that the p_j's can be expressed in terms of a single parameter α, as

$$p_1 = \alpha^2 \qquad p_2 = 2\alpha(1-\alpha) \qquad p_3 = (1-\alpha)^2. \tag{7.2.1}$$

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 $(Y_1, Y_2, Y_3) \sim Mult(n = 100_i p_1, p_2, p_3)$.) The likelihood ratio test statistic is (7.1.6), but we have to find $\tilde{\alpha}$ and then the e_j's. The likelihood function under (7.2.1) is

$$\begin{aligned} L_1(\alpha) &= L(p_1(\alpha), p_2(\alpha), p_3(\alpha)) \\ &= (\alpha^2)^{17}[2\alpha(1-\alpha)]^{46}[(1-\alpha)^2]^{37} \\ &= c\alpha^{80}(1-\alpha)^{120} \end{aligned}$$

and we easily find that $\tilde{\alpha} = .40$. The expected frequencies are therefore $e_1 = 100\tilde{\alpha}^2 = 16$, $e_2 = 100[2\tilde{\alpha}(1-\tilde{\alpha})] = 48$, $e_3 = 100[(1-\tilde{\alpha})^2] = 36$. Clearly these are close to the observed frequencies y_1, y_2, y_3, and (7.1.6) gives the observed value $\Lambda_{\text{obs}} = 0.17$. The significance level is

$$P(\chi^2_{(1)} \geq 0.17) = .68$$

so there is no evidence against the model (7.2.1).

The Pearson statistic (7.1.7) usually gives close to the same value as Λ when n is large. In this case we find that $D = 0.17$.

Example 7.2.2. Continuous distributions can also be tested by grouping the data into intervals and then using the multinomial model. Example 2.5.1 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

$$f(t;\alpha) = \frac{1}{\alpha}e^{-t/\alpha}, \qquad t > 0. \tag{7.2.2}$$

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

$$p_j(\alpha) = \int_{a_j}^{b_j} f(t;\alpha)dt \qquad j = 1, ..., m \tag{7.2.3}$$

and if y_j is the number of observations (t's) that lie in I_j, then $Y_1, ..., Y_m$ follow a multinomial 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,

$$p_1 = \int_0^{100} \frac{1}{\alpha} e^{-t/\alpha} dt = 1 - e^{-100/\alpha},$$

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 (7.2.3) 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 $\tilde{\alpha} = 310.3$ and the expected frequencies $e_j = 100 p_j(\tilde{\alpha})$ given in the table are then obtained.

The likelihood ratio statistic (7.1.6) gives $\Lambda_{\text{obs}} = 1.91$. The significance level is computed as

$$P(\Lambda \geq \Lambda_{\text{obs}}; H_0) = P(\chi^2_{(5)} \geq 1.91) = .86,$$

so there is no evidence against the model (7.2.3). Note that the reason the χ^2 degrees of freedom is 5 is because $k = m - 1 = 6$ 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 expected frequencies 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 variables 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 variables 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'll consider the two types of studies in turn.

7.3.1 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 p_{ij} be the probability a randomly selected individual is combined type (A_i, B_j). Note that

$$\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} = n \qquad \sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij} = 1$$

and that the ab 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

$$H_0: \ p_{ij} = \alpha_i \beta_j \qquad i = 1, ..., a; \ j = 1, ..., b \tag{7.3.1}$$

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 (7.3.1) is the standard definition for independent events: $P(A_i, B_j) = P(A_i)P(B_j)$.

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 (7.3.1) is $p = (a-1) + (b-1) = a + b - 2$. All that needs to be done in order to use the statistics (7.1.6) or (7.1.7) 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 (7.3.1), the likelihood function for the y_{ij}'s is proportional to

$$\begin{aligned} L_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{i=1}^{a}\prod_{j=1}^{b} p_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta})^{y_{ij}} \\ &= \prod_{i=1}^{a}\prod_{j=1}^{b} (\alpha_i \beta_j)^{y_{ij}}. \end{aligned} \tag{7.3.2}$$

It is easy to maximize $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ subject to the linear constraints $\sum \alpha_i = 1$, $\sum \beta_j = 1$. This gives

$$\tilde{\alpha}_i = \frac{y_{i+}}{n}, \ \tilde{\beta}_j = \frac{y_{+j}}{n} \ \text{ and } \ e_{ij} = n\tilde{\alpha}_i\tilde{\beta}_j = \frac{y_{i+}y_{+j}}{n}, \tag{7.3.3}$$

where $y_{i+} = \sum_{j=1}^{b} y_{ij}$ and $y_{+j} = \sum_{i=1}^{a} y_{ij}$. The likelihood ratio statistic (7.1.6) for testing the hypothesis (7.3.1) is then

$$\Lambda = 2 \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} \log(y_{ij}/e_{ij}). \tag{7.3.4}$$

The significance level is computed by the approximation $P(\chi^2_{(a-1)(b-1)} \geq \Lambda_{\text{obs}})$; the χ^2 degrees of freedom $(a-1)(b-1)$ come from $k - p = (ab-1) - (a+b-2) = (a-1)(b-1)$.

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 4 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	
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
	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 (7.3.3). 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 via (7.3.3); 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 isn't an advantage with a computer; its simpler to use (7.3.3) above then. However, it suggests where the term "degrees of freedom" comes from.)

The observed value of the likelihood ratio test statistic (7.3.4) is $\Lambda_{\text{obs}} = 8.52$, and the significance level is approximately $P(\chi^2_{(3)} \geq 8.52) = .036$, 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 p_{ij} that a person in population group i is B-type B_j is the same for all $i = 1, ..., a$. That is, $p_{ij} = P(B_j|A_i)$ 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 p_{ij} satisfy

$$p_{i1} + p_{i2} + \cdots + p_{ib} = 1 \text{ for each } i = 1, ..., a \tag{7.3.5}$$

and the hypothesis we are interested in testing is

$$H_0 : \ \boldsymbol{p}_1 = \boldsymbol{p}_2 = \cdots = \boldsymbol{p}_a, \tag{7.3.6}$$

where $\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{ib})$. 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, Mult$(n_i; p_{i1}, ..., p_{ib})$.

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 (7.3.6) is exactly the same as (7.3.4), where now the expected frequencies e_{ij} are given by

$$e_{ij} = n_i \left(\frac{y_{+j}}{n} \right) \qquad i = 1, ..., a; \ j = 1, ..., b \tag{7.3.7}$$

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 4 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 3 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	
Aspirin Group	64 (75.6)	176 (164.4)	240
Placebo Group	86 (74.4)	150 (161.6)	236
	150	326	476

We can think of the persons receiving Aspirin and those receiving Placebo as two groups, and test the hypothesis

$$H_0: \ p_{11} = p_{21},$$

where $p_{11} = P$(Stroke) for a person in the Aspirin group and $p_{21} = P$(Stroke) for a person in the Placebo group. The test statistic (7.3.4) requires the expected frequencies, which are

$$e_{ij} = \frac{(y_{i+})(y_{+j})}{476} \qquad i = 1, 2.$$

This gives the values shown in the table. The test statistic then has observed value

$$\Lambda = 2\sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij}\log(y_{ij}/e_{ij}) = 5.25.$$

The approximate significance level is

$$P(\chi^2_{(1)} \geq 5.25) = .022$$

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 $p_{11} < p_{21}$.

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

$$\hat{p}_{11} = \frac{y_{11}}{n_1} = \frac{64}{240} = 0.267; \ \hat{p}_{21} = \frac{y_{21}}{n_2} = \frac{86}{236} = 0.364.$$

We can also give confidence intervals for p_{11} and p_{21}; approximate .95 confidence intervals based on earlier methods are $.211 \leq p_{11} \leq .323$ and $.303 \leq p_{21} \leq .425$. Confidence intervals for $\delta = p_{11} - p_{21}$ can also be obtained from the approximate $G(0, 1)$ pivotal quantity

$$Z = \frac{(\hat{p}_{11} - \hat{p}_{21}) - \delta}{\sqrt{\hat{p}_{11}(1 - \hat{p}_{11})/n_1 + \hat{p}_{21}(1 - \hat{p}_{21})/n_2}}.$$

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 distribution $Bin(12, p)$. 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 # 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 a binomial distribution with parameters n and $p_n = P$(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 $p_1 = p_2 = p_3 = p_4$.

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(any digit = j) = $.1 (j = 0, 1, \ldots, 9)$, 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 variables 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	
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 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
(Adapted from Stat 230 Course Notes).

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 gender (M or F) of an applicant, so we cannot ignore it in trying to see if gender 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

$$P(A|B_1C_i) > P(A|B_2C_i) \text{ for each } \quad i = 1, \ldots, k$$

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 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)
      1. $\bar{y}_1 = 11.7$ $\quad s_1 = 2.1$
      2. $\bar{y}_2 = 12.0$ $\quad s_2 = 2.4$

   (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

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

Statistical Tables

APPENDIX. ANSWERS TO SELECTED PROBLEMS

Chapter 1

1. (b) .032 (c) .003 (.002 using Gaussian approx.)
2. (c) $p_1 = .489$, $p_2 = .325$, $p_3 = .151$, $p_4 = .035$
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 $T = \sum kf_k$
   (b) $c = (1-\alpha)^2/\alpha$
   (c) $\hat{\alpha} = .195$; $\hat{Pr}(X = 0) = .758$

(d) $\hat{\alpha} = .5$

7. $\hat{\lambda} = \sum y_i / \sum t_i$

9. (a) $\hat{\alpha} = .35 \qquad \hat{\beta} = .42$

   (b) 14.7, 20.3, 27.3 and 37.7

Chapter 4

1. (a) $\hat{\mu} = 1.744$, $\hat{\sigma} = .0664$(M) $\hat{\mu} = 1.618$,$\hat{\sigma} = .0636$ (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) $\hat{\theta} = .0116$; interval approximately (.0056,.0207)

8. (a) $0 \leq \alpha \leq .548$ (b) .10 likelihood interval is now $.209 \leq \alpha \leq .490$

10. (a) $\hat{\lambda} = 3n / \sum t_i$ (b) $\hat{\lambda} = .06024$; $.0450 \leq \lambda \leq .0785$

    (c) .95 CI for λ is (.0463,.0768) and for μ is $39.1 \leq \mu \leq 64.8$

    (d) CI's are $.408 \leq p \leq .738$ (using model) and $.287 \leq p \leq .794$ (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 $285.5 \leq \theta \leq 521.3$

    (b) $197.9 \leq m \leq 361.3$

13. (b) $288.3 \leq \theta \leq 527.9$

14. (a) $.637 \leq p \leq .764$

Chapter 5

1. $SL = P(D \geq 15) = P(Y \geq 25; \lambda = 10) = .000047$

4. (a) LR statistic gives $\Lambda_{0bs} = .0885$ and $SL = .76$.

5. (a) $\Lambda_{0bs} = 23.605$ and $SL = .005$
   (b) $SL = 1 - .995^6 = .03$ now

6. $\Lambda_{0bs} = .042$ 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 $SL = 2P(\chi^2_{(9)} \leq 1.2) = .0024$ so there is strong evidence against $H : \sigma = .02$
   (b) No: testing $H : \mu = 13.75$ gives $SL < .001$
   (c) $13.690 \leq \mu \leq 13.700$ and $.0050 \leq \sigma \leq .0132$

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) $SL = .05$ (c) $SL = .07$

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) $\hat{\beta} = .02087$, $\hat{\alpha} = -1.022$, $s = .008389$
   (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 $P(\chi^2_{(1)} \geq 1.11) = .29$, $P(\chi^2_{(2)} \geq 4.22) = .12$, and $P(\chi^2_{(3)} \geq 1.36) = .71$, respectively.
   (b) LR statistic is 7.54 and $SL = P(\chi^2_{(3)} \geq 7.54) = .057$.

7. The LR statistic is 7.32 and $SL = P(\chi^2_{(2)} \geq 7.32) = .026$ 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 $SL = P(\chi^2_{(3)} \geq .57) = .90$ so there is no evidence of association.
   (b) LR statistic gives $\Lambda_{0bs} = 5.44$ and $SL = P(\chi^2_{(3)} \geq 5.44) = .14$. There is no evidence against the binomial model.

10. (a) $\Lambda_{0bs} = 10.8$ and $SL = P(\chi^2_{(3)} \geq 10.8) = .013$.

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.

A Short Review of Probability