A First Course in Probability and Statistics

Nick Syring

2022-11-27

Contents

1 About **7**

2 Data Analysis and Statistical Inference **9**

2.1 Data, Experiments, and Studies 9

2.2 Data Summaries 14

2.3 Statistical Inference 18

3 Statistical Estimation **21**

3.1 Vocabulary 21

3.2 Properties of Estimators 21

3.3 Finding estimators - Method of Moments 25

3.4 Method of Maximum Likelihood 26

3.5 Properties of MLEs 29

4 Sampling Distributions **31**

4.1 Sample Mean 31

4.2 Sample Variance 33

4.3 Sampling distribution of studentized sample mean 35

4.4 Differences of Sample Means 37

4.5 Ratios of Sample Variances 38

5 Confidence Intervals **39**

5.1 Why want interval-valued estimates? 39

5.2 Normal population mean example 41

5.3 Other "Exact" CIs for normal population mean and variance parameters . . . 42

5.4 CIs for proportions . . . 47

5.5 Approximate CIs based on MLEs . . . 49

6 Hypothesis Testing 53

6.1 Notation . . . 53

6.2 Hypothesis testing outcomes . . . 54

6.3 Tests based on a normal population . . . 55

6.4 Likelihood-based Tests . . . 68

6.5 Chi Squared tests for tabulated data . . . 72

7 ANOVA 77

7.1 Analysis of variance . . . 77

7.2 Crop data in-depth example . . . 78

8 Simple Linear Regression 87

8.1 Simple linear regression model . . . 87

8.2 Example: House prices as function of house age . . . 94

9 Multiple Linear Regression 101

9.1 The Multiple Linear Regression Model . . . 101

9.2 Point estimation . . . 101

9.3 Sampling Distributions . . . 103

9.4 Inference on regression coefficients . . . 103

10 Likelihood-Based Inferences 107

10.1 Wald tests and CIs . . . 107

11 Bootstrap 115

11.1 Intro to the bootstrap . . . 115

11.2 Bootstrap Confidence Intervals . . . 119

11.3 Bootstrapping linear models . . . 120

12 Bayesian Statistics 125

12.1 Introduction 125

12.2 Bayesian methods in general 126

12.3 Bayesian methods for a single normal population 127

Chapter 1

About

This is a collection of notes intended for students studying probability and statistics at the advanced undergraduate level at Iowa State University. Specifically, these notes are modeled after course notes for STAT 588, 341, and 342. This site is a work in progress.

Some relevant textbook references include John E. Freund's Mathematical Statistics with Applications by Miller and Miller, or Modern Mathematical Statistics with Applications by Devore and Berk, although many books on this material are available.

Chapter 2

Data Analysis and Statistical Inference

In this first chapter we define and discuss some important concepts regarding data and scientific investigation.

2.1 Data, Experiments, and Studies

We encounter numerical summaries of information constantly in our everyday lives and sort through these in order to make all sorts of decisions. In this section we will formalize the concept of *data* connected to scientific study. Broadly speaking, data is anything we observe that is relevant to answering a question of interest, i.e., data is tacitly assumed to be informative about the question.

Three types of studies in which data are collected and analyzed are in designed, interventional experiments, observational studies, and exploratory studies. The following illustrations help to differentiate these two types of studies. One difference we will see is all about timing—experiments start with a question of interest and then are designed to answer the question. Exploratory data analysis is often done using data collected with no particular question in mind, or some other question besides the current one. Another difference is that experiments require interventions—imposed changes of behavior or conditions on the individuals in the experiment—while observational studies do not include such interventions.

2.1.1 James Lind's Scurvy Trial

Clinical trials are familiar designed, interventional experiments. A famous, early example is James Lind's Scurvy trial. Lind was a doctor aboard a British ship.

Several sailors with him were suffering from scurvy. He selected 12 sailors in similar, poor condition, and assigned to pairs of them 6 different treatments (the intervention). The two who received oranges and lemons to eat recovered fully; those who received apple cider fared next best. Lind specifically wanted to understand which treatments would be most effective at curing scurvy and planned an experiment to answer his question. His selection of like sailors and random, paired treatment assignment constitute early examples of randomization and control in experimental design.

2.1.2 Framingham Heart Study

Named for Framingham, Massachusetts, where the participants were first recruited, this was a long-running observational study of Americans aimed at understanding risks associated with heart disease. Participants agreed to medical testing every 3-5 years, from which the study researchers concluded a number of important findings, such as cigarette smoking substantially increases the risk of heart disease. There are no interventions; the researchers simply observe the patients and make conclusions based on how the patients choose to live, e.g., tobacco use.

2.1.3 Harris Bank Sex Pay Study

93 salaries of entry-level clerical workers who started working at Harris Bank between 1969 and 1971 show men were paid more than women. (From The Statistical Sleuth, reproduced from “Harris Trust and Savings Bank: An Analysis of Employee Compensation” (1979), Report 7946,Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.)

2.1.4 Large, Aggregated Data Sets

There are now many large “data sets” recording all sorts of information about consumers, e.g., data obtained by Google and other technology companies whenever consumers use their sites, apps, or devices. There are many potential use cases for such information; for instance, such data has proven helpful for targeted marketing of specific products. Such applications may be termed exploratory research—these are characterized, in part, by exploration/examination of data that was originally collected for some other purpose. In other words, the “data” was collected with either no question or some other question in mind.

2.1.5 Study Concepts

The above examples illustrate several key concepts related to scientific studies.

- Research question - There is always a reason researchers go to the trouble of collecting and analysing data; they have an important question they want to answer. For example, what can sailors do to prevent scurvy?
- Experimental units/Subjects - the research question usually references people, things, animals, or some other entity that can be studied in order to answer the question. When these are observed and measured then they are called experimental units or subjects. In the context of interventional experiments these usually refer to the units of randomization; see below.
- Data - we are inundated with information, numbers, figures, and graphs in our everyday lives. Is this data? Anything information gathered to answer a particular research question can be considered data. Relevancy to a research question is key.
- Intervention - When James Lind gave different foods to sick sailors he was making an intervention, and his goal was to study the effect of his interventions on the sailors well-being. Experiments include one or more interventions, whereas observational studies do not feature any interventions on the part of the researcher.
- Randomization - When researchers intervene, they should apply their interventions randomly with respect to subjects. In experiments the experimental units are the entities that are randomized and given interventions.
- Response/outcome variables - studies often measure multiple variables and study relationships between them. Typically the researchers expect one variable is affected by another. The response, or outcome—like the health of sailors, or pay of workers—is the variable being affected by the intervention in an experiment or by another, independent variable in an observational study.
- Control - Researchers should try to limit the effects of variables on the response that are not of interest to the study. For example, in the gender pay study, the researchers studied only entry-level workers. They *controlled* for prior experience to better isolate potential sex effects on pay.

2.1.6 Randomization, control, and causation

In experiments the researcher performs one or more interventions—such as giving patients cider versus citrus fruits in Lind's scurvy trial. The principle of *randomization* asserts that interventions in experiments should be assigned to experimental units randomly. When experimental units are heterogeneous—not all the same—it stands to reason that some of their differences apart from the intervention may impact the response to the experiment recorded by the researcher. Randomization is a way to even out these heterogeneities between groups receiving different interventions. That way, it is the intervention, rather than some other difference, which is responsible for substantially different outcomes. Randomization systematically accounts for heterogeneities, but the extent to which it works depends on the number of experimental units, the number of groups being randomized, and the presence of one or more important hetero-

geneities. For examples, consider the following: 1. Suppose in ten experimental units there is one unobserved, dichotomous trait that affects the experimental response. Three of the ten have version “0” of the trait and 7 have version “1”. Randomly split the ten into two groups of five, each group to receive a different intervention. The chance all three end up in the same group is 1/6, not ignorably small... 2. On the other hand, suppose there are 100 experimental units, half have trait “0” and half trait “1”. The chance at least 35 of either trait type end up in the same one half random split is ≈ 0.
Generally when an intervention is randomized over experimental units we interpret any significant difference in outcome/response between intervenion groups as having been *caused* by the intervention itself, as opposed to some other unobserved characteristic—these are sometimes called *lurking variables* or *confounding variables*. But, randomization is not foolproof; small sample sizes (few experimental units) and/or the presence of many confounding variables can reduce the effectiveness of randomization.

When the researcher knows about potential confounders ahead of time, the principle of *blocking* says experimental units should be representatively divided with respect to intervention across values of this variable. For example, if experimental units are humans both young and old, then the different interventions should be applied to equal numbers of young and old people. One way to accomplish this is to let age group be a *blocking factor*. In the case of a dichotomous intervention this means half of the old people will be randomly assigned to one intervention and half of the young people will be randomly assigned to one intervention—as opposed to randomly assigning half of all the experimental units to one intervention.

The principle of *control* states that intervention groups should be made as homogeneous as possible. When experiments are well-controlled researchers often assume that they can determine *causation*, and any observed differences in experimental outcome between intervention groups can be attributed to the intervention. Of course, as mentioned above, the ability of an experiment to determine causation is not all or nothing; rather, it depends on unknowns. Nevertheless, stronger controls make the results of experiments more trustworthy, and less likely to be caused by confounding variables.

Non-interventional, observational studies are not used to establish causative relationships. Rather, we say such studies establish *associations* between variables. For example, in the Framingham study, the researchers did not randomly assign individuals to groups of tobacco-users and non-users. Even though these days the evidence is quite strong that tobacco use causes heart disease, the bar for such a claim is much higher when the variable of interest—tobacco use—cannot be randomly assigned to experimental units. That’s not to say elements of control cannot be used. For instance, if enough data is collected, it is possible to compare tobacco-users and non-users with nearly the same ages, sexes, incomes, education, living in the same zip codes, etc. The more potential confounders are explicitly controlled, the closer such an observational study comes to a randomized experiment.

2.1.7 Populations and scope of inference

Whether conducting an experiment or collecting observational data, the units/subjects have to come from somewhere. The *population* refers to the set of all possible subjects—it may be finite or infinite, and it may be concrete or hypothetical. For examples, the population of current Iowa State undergraduates is well-defined, whereas the population of mouse kidney cells exists in a hypothetical sense. *Sampling* describes how subjects are obtained from the population for observation. *Random sampling* is any scheme involving selecting a subset of subjects from a larger group in some random fashion. A *simple random sample* of size n is obtained when every subset of n subjects from a total group is equally likely to be selected. Other types of random selection are possible, but we won't often consider these: - *stratified random sampling* obtains when simple random samples from separate groups/strata are combined, e.g., a 50/50 random sample stratified by male/female can be formed by taking a simple random sample of ten males and a simple random sample of 10 females from a group of 50 males and 50 females - *cluster random sampling* obtains when a larger group is subdivided into smaller groups and subgroups are selected at random, e.g., a cluster random sample of Iowa high schoolers can be obtained by choosing all high schoolers who attend one of a simple random sample of Iowa high schools.

Generally, conclusions about subjects in the study—whether it is an experiment or an observational study—may be assumed to hold for the wider population as a whole when the subjects are chosen randomly. The details of this *generalizability* of results depend on the type of random sampling conducted; we'll focus on the case of simple random sampling specifically. On the other hand, when subjects are not randomly sampled from the population, study results cannot be generalized back to the population. The reason is that the lack of randomness in selection implies some subsets of the population are more or less likely to be present in the sample of subjects observed, hence, that sample is not necessarily *representative* of the population. For an extreme example, consider a population of both young and old people and an experiment studying the effects of Covid-19. It's well-known Covid-19 is much more harmful to old people compared to young people. So this is a potential confounder. If we select only young people to study, then we certainly cannot claim the results would be similar had we studied both young and old people.

Non-random sampling schemes are quite common because they are usually easier and cheaper to implement than random sampling schemes. A *convenience sample* is just what it sounds like—a rule that selects subjects that are easy to select—such as conducting a poll of your closest friends. When a non-random sample is used, remember that the results cannot be interpreted beyond the group subjects that were observed.

Sometimes a researcher intends to study one population, but obtains data from another population. This mismatch is important to identify as it can cause bias—which simply means the answer to the researcher's question is different for the population for which data is observed compared to the intended population. As in the extreme example above, effects of Covid-19 are different in old and young populations, so the results from an experiment studying only the young are biased when viewed from the perspective of the population of old and young combined.

2.2 Data Summaries

Data may take on many forms including sound waves, images comprised of pixels/voxels, and graphs of functions/surfaces. We will almost exclusively consider data that may be represented in a tabulated format, like the following data on weights of chickens.

```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

Regardless of the type of data, few users can make sense of all the data at once; rather, we need data summaries—numbers or plots that point out important features of the whole data set.

2.2.1 Numerical Summaries

We briefly describe summaries for categorical and numerical, continuous data.

Suppose a variable X takes categorical values, e.g., $X \in \{0, 1, 2, 3\}$. A data set consisting of observed x values, may be summarized by tabulating proportions, i.e., of 100 observations, 0.41 were $x = 0$, 0.29 were $x = 1$, 0.17 were $x = 2$ and 0.13 were $x = 3$. Binary or dichotomous variables may be also be summarized by odds and odds ratios. Suppose $Y \in \{0, 1\}$. The observations $y_1, \dots, y_n$ may be summarized by saying 60% were $y = 1$ versus 40% $y = 0$ or, equivalently, the observed odds of $y = 1$ was $0.6/0.4 = 1.5$. Suppose response Y is *blocked* by another dichotomous variable, say, undergraduate versus graduate status, and suppose the observed odds for undergraduates is 1.5 while the observed odds for graduates is 0.8. Then, the observed odds ratio for undergraduates versus graduates is $1.5/0.8 = 1.875$.

In contrast to categorical data, continuous data takes values in the real numbers $\mathbb{R}$ or some interval of real numbers. Most numerical summaries of continuous variables either measure location or dispersion. Location refers, in one way or another, to a typical or average value while dispersion refers to the spread of the values. Common measures of location are the mean, median, and trimmed mean:

- The mean of $x_1, \ldots, x_n$ is $\overline{x} = n^{-1}\sum_{i=1}^n x_i$.
- The median is denoted $\tilde{x}$. Sort the x's from least to greatest labeling them as $x_{(1)}, \ldots, x_{(n)}$. Then, $\tilde{x} = x_{(\frac{n+1}{2})}$ if n is odd and $\tilde{x} = \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right)$ if n is even.
- The $\alpha\%$ trimmed mean is the mean of the remaining x values after ignoring the smallest and largest $\alpha\%$ of values. Roughly, the $\alpha\%$ trimmed mean averages $x_{(1+\alpha n)}, \ldots, x_{n-\alpha n-1}$.
- Besides the median, we may be interested in "locations" besides the center, and these could be summarized using quantiles. Observed quantiles may be defined in a number of ways. One definition says the $\alpha \in (0,1)$ quantile is the smallest observed x value such that at least $\alpha\%$ of observed x values are no more than x.

Common measures of dispersion include the variance and standard deviation, the range, and the interquartile range:

- The observed variance of $x_1, \ldots, x_n$ is given by $\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ and may be equivalently computed as $\frac{1}{n-1}\left\{\sum_{i=1}^n (x_i^2) - n\overline{x}^2\right\}$.

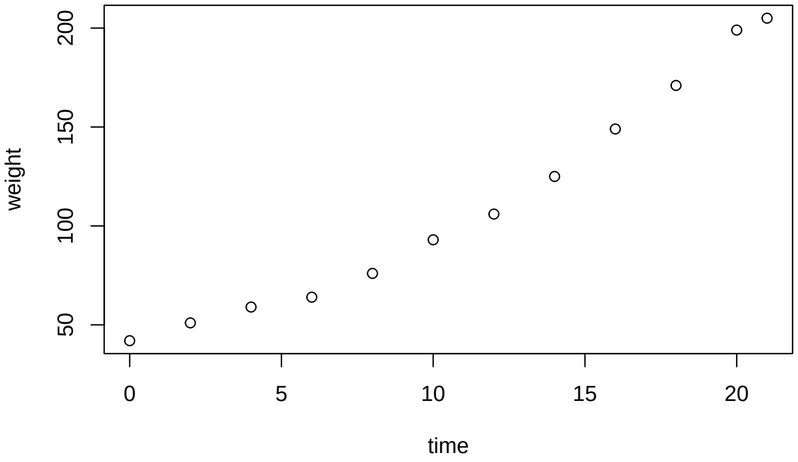
- In the calculation of variance the x values are squared, which means the variance has units equal to the squared units of x; that is, if x is measured in millimeters then its variance is measured in millimeters squared, or, e.g., dollars then dollars squared. Often the squared units are difficult to interpret. Instead, we define the standard deviation as the square root of the variance, which shares the units of the observed x variable.

- The range is the difference between maximum and minimum values, i.e., $x_{(n)} - x_{(1)}$.
- The interquartile range (IQR) is the difference between the 0.75 and 0.25 quantiles.

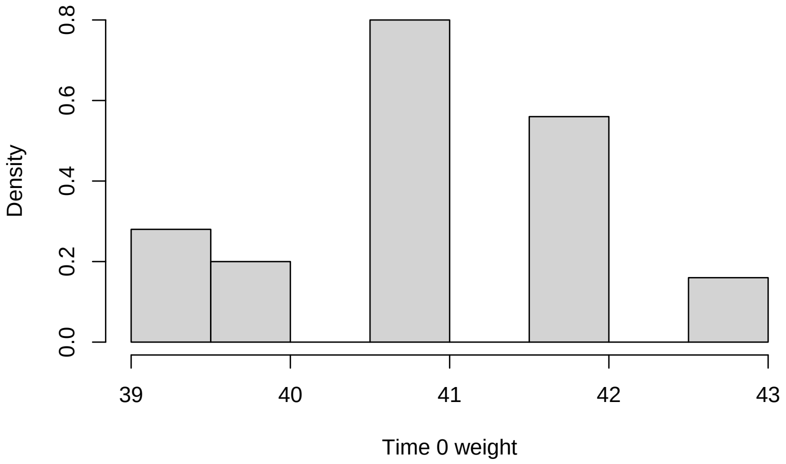
2.2.2 Visual Summaries

Plots and graphs can also be used to summarize a data set or particular observed variable. Three plots we will use extensively in the course are the scatterplot, histogram, and quantile-quantile (qq) plot.

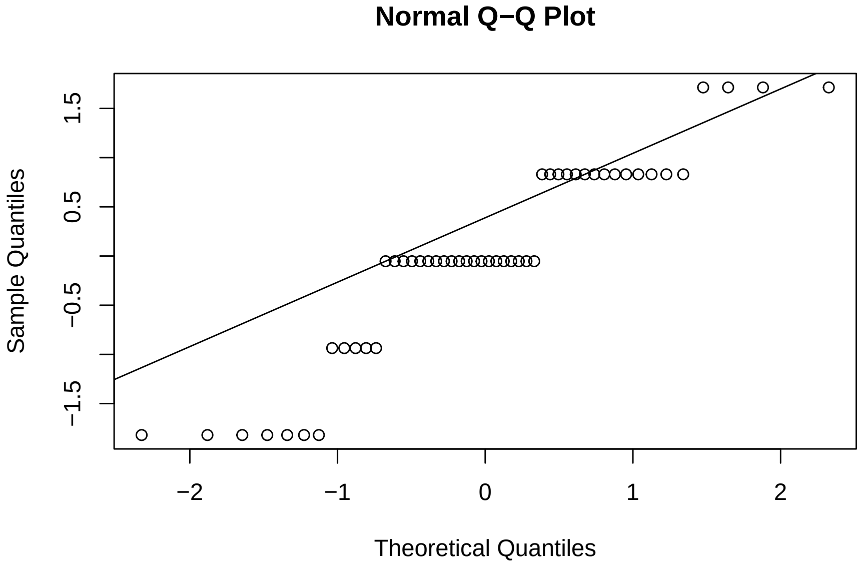
- For paired observations of two variables (x_i, y_i), $i = 1, \ldots, n$ a scatterplot displays (x_i, y_i) in the xy-plane. For example, see the plot of a chicken's weight versus time from the tabulated data set above. Scatterplots are useful for assessing relationships between variables. For example, the chick's weight increases with time, maybe in a linear or slightly quadratic fashion.



- For a single variable a histogram summarizes the distribution of its observed values by counting the numer of observations in different intervals (buckets) of values. Keep in mind that histograms with different choices of buckets may look very different. Check out this histogram of "Time 0" weights of all 23 chicks.



- A qq-plot compares the shape of a distribution of observed values to another known distribution, often the standard normal distribution. For example, make the standardizing transformation $(x_i - \overline{x})/\widehat{\sigma}_x$ where x_i is the Time 0 weight of chick i, $\overline{x}$ is the observed mean and $\widehat{\sigma}_x$ is the observed standard deviation of those values. Compute the α quantile of these values or several α values in $(0, 1)$ along with the corresponding standard normal quantiles (z-scores). Plot the pairs of α quantiles in the xy-plane. If the standardized weights are approximately normal, then the points should lie approximately on the line $y = x$. Note that extreme quantiles are always less reliably estimated, so it is typical for the ends of the "line" to fray up or down from the diagonal.



2.3 Statistical Inference

Summarizing data sets is important because no one can make sense of more than a few numbers at a time. But, data summaries cannot by themselves answer our research questions. That is because data summaries only say something about the particular data set we observe, while our research questions concern the whole population. Recall, a major aim of experimentation is generalizing from observations to population. When we make such generalizations we often refer to them as *inferences*; and, statistical inferences are characterized by their careful treatment of statistical concepts, such as hypotheses, and Type 1 and 2 errors, which we discuss below.

A hypothesis is a claim or assertion about a population. These may be simple, i.e., “the population mean is 5”, or more complex, like “the population distribution of values is equivalent to a Normal probability distribution”. Notice that both of these statements are either true or false, yes or no. A hypothesis then may be called the “null hypothesis”, and its complement (opposite) the alternative hypothesis. Which is which depends on the context. We make educated guesses about the truthfulness of a null hypothesis based on observations/data. Our educated guesses may be right or wrong, but we will never know because we will never “see” the whole population. If we reject the null hypothesis as false when it really is true, then we make a Type 1 error. The opposite, keeping the null hypothesis in favor over its alternative, when it is actually false, is a Type 2 error. Ideally, we would make no errors, but that’s not possible. In fact, the two errors have an inverse relation. For example, if we adopt the rule that we always reject the null hypothesis, then we will necessarily maximize Type 1 errors but have no Type 2 errors. And, if we take the opposite approach, then we maximize Type 2 errors while making no Type 1 errors.

Much of this course will focus on constructing tests of relevant hypotheses with the property that we limit the chance of making a Type 1 error. By chance we refer to the probability distribution of the test outcome induced by random sampling of data from the population. A test that has chance no more than α of making a Type 1 error is called a "level α test of H_0", the null hypothesis.

Chapter 3

Statistical Estimation

3.1 Vocabulary

We consider experiments in which a random sample of observations is taken from a population. The sampling distribution of observations is known up to the value(s) of some *population parameter(s)*. The goal of this chapter is to study *estimation*—approximation of these unknown parameters by *statistics*, functions of the sample data. An *estimator* is the random variable version of a statistic and has a corresponding sampling distribution. By virtue of being called an estimator we assume this statistic is "close" to the true parameter value in some sense. An *estimate* is a value of an estimator after the data is collected and the estimator computed; it is a fixed, non-random value.

For example, consider an experiment sampling n iid random samples from a Normal population with unknown mean μ and variance 1. Here μ is the parameter we wish to estimate. An intuitive estimator is the sample mean statistic $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. When the data is collected we refer to the calculated sample mean—the estimate of μ—as lowercase $\overline{x}_n$.

3.2 Properties of Estimators

It will become apparent that estimators are not unique—there are very often several seemingly reasonable estimators for a single parameter. Therefore we ought to be concerned with choosing the "best" estimator according to some criteria. Different criteria will lead to different "best" estimators.

3.2.1 Bias and Unbiasedness

Consider a generic parameter denoted θ and imagine an estimator for θ called $\widehat{\theta}_n$, which is a statistic and, hence, a random variable. We say $\widehat{\theta}_n$ is *unbiased* if $E(\widehat{\theta}_n) = \theta$ where the expectation is taken with respect to the sampling distribution of $\widehat{\theta}_n$. For example, the sample mean $\overline{X}_n$ is unbiased for the population mean μ (so long as the population mean exists):

$$\begin{aligned} E(\overline{X}_n) &= E\left(n^{-1}\sum_{i=1}^{n} X_i\right) \\ &= n^{-1}\sum_{i=1}^{n} E(X_i) \\ &= n^{-1}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned}$$

An unbiased estimator commits no systematic errors in estimating the parameter. In contrast, a biased estimator of a univariate real-valued parameter systematically underestimates or overestimates the parameter. Consider biased $\widehat{\theta}_n$ such that $E(\widehat{\theta}_n) = \theta + 1$. We **expect** $\widehat{\theta}_n$ to be 1 more than θ—hence we expect it to overestimate.

3.2.2 Minimum Variance Unbiased Estimators

Besides a finite mean, estimators often have a finite variance. And, intuitively, we would tend to prefer an estimator with low variance to one with high variance, particularly if both are unbiased. If we insist on an unbiased estimator, then the "best" unbiased estimator is the unbiased estimator with smallest variance among all unbiased estimators.

It is not obvious how one would find the lowest variance estimator among all unbiased estimators. One result, due to CR Rao and Harald Cramer, provides a lower bound on the variance of an unbiased estimator. This lower bound can be checked by computing the variance of a given unbiased estimator, and, if they match, this implies the given estimator is the MVUE. We describe this procedure below.

Let $f(x;\theta)$ denote the density of the data and let θ denote a univariate parameter. Let $\ell(x;\theta) := \log(f(x;\theta))$, the natural logarithm of the density. Define the *Fisher Information* for one data point by

$$I(\theta) = E\left[\left(\frac{\partial \ell(x;\theta)}{\partial \theta}\right)^2\right].$$

The Fisher information for a random sample of size n is n times $I(\theta)$. Then, Cramer and Rao showed that if $\hat{\theta}_n$ is unbiased, its variance cannot be smaller than

$$V(\hat{\theta}_n) \geq [nI(\theta)]^{-1}$$

where θ is the true parameter value.

Example: Let $X_1, \ldots, X_n$ be a random sample from a normal population with variance 1 and mean μ and consider the estimator $\overline{X}_n$. The logarithm of the density function is

$$\ell(f(x;\mu)) = -\log(\sqrt{2\pi}) - \tfrac{1}{2}(x-\mu)^2$$

with $\mu-$derivative $x - \mu$. Therefore, the Fisher Information is

$$I(\mu) = E[(X-\mu)^2] = 1$$

since it is, by definition, equal to the variance. The Cramer-Rao lower bound is

$$V(\hat{\theta}_n) \geq \frac{1}{n},$$

and we know $V(\overline{X}_n) = 1/n$ so the sample mean $\overline{X}_n$ is, indeed, the MVUE for this experiment.

3.2.3 Mean Squared Error and Bias-Variance tradeoff

As described above a common strategy is to select an unbiased estimator, preferably one with low (or the lowest) variance. On the other hand, one may prefer a biased estimator over an unbiased one if the bias is low and there is substantial reduction in variance. One way to choose estimators that balance bias and variance is to consider their mean squared error (MSE), defined by

$$MSE(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2] = Bias(\theta_n)^2 + V(\hat{\theta}_n).$$

It is left as an exercise to the reader to show the MSE equals the sum of estimator variance and squared bias. For the above reasons the estimator minimizing the MSE may be preferable even to the MVUE.

Example: Estimation of a normal population variance The usual variance estimator is the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. It can be checked this estimator is unbiased. And, it is a bit of a pain, but it can be shown that $V(S_n^2) = \frac{2\sigma^4}{n-1}$. Now, consider an alternative estimator that is a constant multiple of S_n^2, say cS_n^2 for some $c > 0$. The MSE of this estimator is

$$\begin{aligned} MSE(cS_n^2) &= V(cS_n^2) + Bias(cS_n^2)^2 \\ &= c^2\frac{2\sigma^4}{n-1} + [E(cS_n^2) - \sigma^2]^2 \\ &= c^2\frac{2\sigma^4}{n-1} + (c\sigma^2 - \sigma^2)^2 \\ &= c^2\frac{2\sigma^4}{n-1} + \sigma^4(c-1)^2. \end{aligned}$$

Differentiate w.r.t. c to find

$$\frac{\partial MSE}{\partial c} = 2c\frac{2\sigma^4}{n-1} + 2(c-1)\sigma^4.$$

Set this equal to zero and solve for c. We get

$$c = \frac{2\sigma^4}{\frac{4\sigma^4}{n-1} + 2\sigma^4} = \frac{n-1}{2+n-1} = \frac{n-1}{n+1}.$$

This means the minimim MSE estimator (at least among those that are multiples of S_n^2) is actually $\frac{1}{n+1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

3.2.4 Consistency

Besides avoiding systematic estimation errors and having low variance, we would expect that as more and more data is collected an estimator should get better and better—and get “closer” to the true parameter, in some sense. This intuition is captured mathematically by *consistency.* An estimator is consistent if for any $c > 0$, however small,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > c) = 0.$$

Dissecting this definition from the inside out we first note $|\hat{\theta}_n - \theta|$ is the random estimation error. Then, $|\hat{\theta}_n - \theta| > c$ says the error is at least c. Since the error is a random variable we attach a probability to the chance the error is at least c, which is $P(|\hat{\theta}_n - \theta| > c)$. And, consistency says this probability must vanish as we accumulate data. So, the chance of an error of any size c or bigger vanishes. Taking complements, this is equivalent to saying

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < c) = 1,$$

which means $\hat{\theta}_n$ is within c of θ with probability going to 1.

One way to show an estimator is consistent is to show it is unbiased and has variance that vanishes as $n \to 0$. One example is the sample mean, which is unbiased and has variance σ^2/n. Then, consistency follows by Chebyshev's inequality, which says: for any r.v. X with finite mean and variance (μ, σ^2),

$$P(|X - \mu| > c) \leq \frac{\sigma^2}{c^2}$$

for any $c > 0$. In the context of estimation, we have

$$P(|\hat{\theta}_n - E(\hat{\theta}_n)| > c) \leq \frac{Var(\hat{\theta}_n)}{c^2}.$$

Now, suppose $\hat{\theta}_n$ is unbiased and its variance vanishes in n. Then, the above statement says

$$P(|\hat{\theta}_n - \theta| > c) \leq s_n.$$

for a sequence s_n satisfying $\lim_{n\to\infty} s_n = 0$. Checking the definition we see this means $\hat{\theta}_n$ is consistent. A similar argument can work for biased estimators as well, provided the bias also vanishes as $n \to \infty$. Such estimators are called *asymptotically unbiased*. One such example is the minimum MSE estimator of σ^2 from the previous section. It has bias $\sigma^2(\frac{n-1}{n+1} - 1)$ which has limit zero.

3.3 Finding estimators - Method of Moments

So far we've discussed desirable properties of estimators but not where these estimators come from. How do we find estimators in the first place? One strategy is based on the fact population parameters often are related to population moments. Then, we can find estimators by replacing population moments by sample moments and referring back to the relationship between the moments and parameters. The simplest example of the method of moments is estimation of the population mean by the sample mean.

Example: Suppose a population is modeled as a Gamma distribution with shape and rate parameters (α, β). The population mean is α/β and the population variance is α/β^2. This means the population second raw moment is $\alpha/\beta^2 + (\alpha/\beta)^2$. We can estimate (α, β) by matching the sample and population raw moments as follows:

$$\begin{aligned} n^{-1}\sum_{i=1}^n X_i &= \alpha/\beta^2 \\ n^{-1}\sum_{i=1}^n X_i^2 &= \alpha/\beta^2 + (\alpha/\beta)^2. \end{aligned}$$

Solving the system by substitution we have

$$\begin{aligned} \hat{\alpha} &= \frac{\hat{\mu}_2' - \hat{\mu}_1'}{\hat{\mu}_1'} \\ \hat{\beta} &= \sqrt{\frac{\hat{\mu}_2' - \hat{\mu}_1'}{(\hat{\mu}_1')^2}}, \end{aligned}$$

where $\hat{\mu}_1'$ and $\hat{\mu}_2'$ indicate the 1st and second raw sample moments $n^{-1}\sum_{i=1}^n X_i$ and $n^{-1}\sum_{i=1}^n X_i^2$.

Example: Suppose we will take a random sample of size n from a continuous uniform distribution supported on the interval (a,b) where the endpoints are unknown. We know that $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$ so that $E(X^2) = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}$. Then, we find estimators for (a,b) by solving

$$\begin{aligned} n^{-1}\sum_{i=1}^n X_i &= \frac{a+b}{2} \\ n^{-1}\sum_{i=1}^n X_i^2 &= \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}. \end{aligned}$$

With some work, you should find $(\hat{a},\hat{b}) = (\hat{\mu}_1' - \sqrt{3\hat{\mu}_2'},\, \hat{\mu}_1' + \sqrt{3\hat{\mu}_2'})$. That's not so intuitive… What about using the sample minimum and maximum…

3.4 Method of Maximum Likelihood

Again consider a random sample $X_1,\ldots,X_n$ from a population synonymous with a density function $f(x;\theta)$ for an unknown parameter θ. For the time being consider only scalar θ. The *likelihood function* is the joint PDF of the data viewed as a function of the parameter, and may be treated either as a random function or as a deterministic function depending on whether the data are treates as random variables or as observed values, so pay close attention to the context. For iid data the likelihood can be written

$$L(\theta; X_1,\ldots,X_n) = \prod_{i=1}^n f(X_i;\theta).$$

For the purpose of estimating θ we view the likelihood as a deterministic function given observations. Then, it acts as a sort of "ranking function" that provides a quantitative comparison of how well different parameter values agree with the observed data. The idea is to select as an estimate the parameter value that maximizes the likelihood/agreement with the data. To explain this concept of agreement with the data a bit more suppose the data come from a discrete population so that $f(x;\theta)$ is a PMF rather than a density—this makes the interpretation easier. Then $\prod_{i=1}^n f(X_i;\theta)$ is a the probability of observing the data for a given parameter value θ. Choosing θ to maximize this probability means selecting the distribution that gives the highest probability assignment to the data that was actually observed.

Example: Suppose our random sample comes from an Exponential distribution with rate λ. The likelihood function equals

$$\begin{aligned} L(\lambda, x_1, \ldots, x_n) &= \prod_{i=1}^{n} \lambda^{-1} e^{-x_i/\lambda} \\ &= \lambda^{-n} e^{-\frac{1}{\lambda}\sum_{i=1}^{n} x_i}. \end{aligned}$$

Take the first derivative of the likleihood with respect to λ:

$$\frac{\partial L}{\partial \lambda} = -n\lambda^{-(n+1)} e^{-\frac{1}{\lambda}\sum_{i=1}^{n} x_i} + \lambda^{-n} e^{-\frac{1}{\lambda}\sum_{i=1}^{n} x_i} \left(\lambda^{-2}\sum_{i=1}^{n} x_i\right)$$

Set $\frac{\partial L}{\partial \lambda}$ equal to zero and solve for λ to obtain the MLE:

$$\begin{aligned} \frac{\partial L}{\partial \lambda} = 0 &\Rightarrow -n\lambda^{-1} + \tfrac{1}{\lambda^2}\sum_{i=1}^{n} x_i = 0 \\ &\Rightarrow n\lambda = \sum_{i=1}^{n} x_i \\ &\Rightarrow \hat{\theta}_{MLE} = \overline{x}_n. \end{aligned}$$

Example: Suppose our random sample comes from a Uniform distribution on the interval $(0, \theta)$. The likelihood function equals

$$\begin{aligned} L(\lambda, x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\theta} 1(0 \le x_i \le \theta) \\ &= \theta^{-n} \prod_{i=1}^{n} 1(0 \le x_i \le \theta). \end{aligned}$$

We cannot simply maximize this likelihood function by taking the first derivative because the indicator functions are not everywhere differentiable w.r.t. θ. Instead, note that the function θ^{-n} is monotonically decreasing in θ; so, this function prefers small θ. On the other hand, the function $\prod_{i=1}^{n} 1(0 \le x_i \le \theta)$ is constant and equal to 1 so long as $\theta \ge \max_{i=1,\ldots,n} x_i$; otherwise, this function is zero and so is the likelihood (and the likelihood cannot be less than zero!). This means we should choose $\theta = \max_{i=1,\ldots,n} x_i$ to maximize the likelihood. Therefore, the MLE of θ is $\hat{\theta}_{MLE} = max_{i=1,\ldots,n} x_i$.

Multivariate Example: Consider estimation of both the mean and variance parameters of a normal population based on a random sample of size n denoted $X^n = (X_1, \ldots, X_n)^\top$. The likelihood is a function of two parameters (μ, σ^2):

$$L(\mu, \sigma^2; X^n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right).$$

The loglikelihood is easier to maximize so take the log and find

$$\ell(\mu,\sigma^2;X^n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2.$$

To find the MLEs we need to compute the gradient vector:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)\\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i-\mu)^2.\end{aligned}$$

Setting $\frac{\partial \ell}{\partial \mu}=0$ we immediately find $\hat{\mu}_{MLE} = \overline{X}_n$. Substituting this estimate for μ in the second equation we find $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, the version of the sample variance with denominator n instead of $n-1$. The following result says the MLE has variance approximately equal to the reciprocal of the Fisher information when the parameter is a scalar. In the multivariate setting the analogous quantity is the inverse Fisher information matrix, which is the inverse of the expectation of -1 times the matrix of second partial derivatives of the loglikelihood. Let's calculate this variance-covariance matrix next. First, we need to compute the matrix of second partial derivatives of the loglikelihood:

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \mu^2} &= -\frac{n}{\sigma^2}\\ \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4}\sum_{i=1}^n (X_i-\mu)\\ \frac{\partial^2 \ell}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (X_i-\mu)^2\end{aligned}$$

Next, take the expectation of these partial derivatives and multiply by -1:

$$\begin{aligned}-E(\frac{\partial^2 \ell}{\partial \mu^2}) &= \frac{n}{\sigma^2}\\ -E(\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2}) &= 0\\ -E(\frac{\partial^2 \ell}{\partial (\sigma^2)^2}) &= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4}.\end{aligned}$$

Finally, find the inverse of the matrix:

$$\begin{bmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}.\end{bmatrix}$$

Since this is a diagonal matrix, the inverse is simply the matrix of reciprocal values on the diagonal:

$$Cov(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}) \approx \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n}. \end{bmatrix}$$

according to the asymptotic results on MLEs given below. Since the unknown parameter σ^2 shows up in this covariance matrix, in practice we replace it with its estimate, yielding the estimated covariance matrix

$$\widehat{Cov}(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}) \approx \begin{bmatrix} \frac{\hat{\sigma}^2_{MLE}}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4_{MLE}}{n}. \end{bmatrix}$$

3.5 Properties of MLEs

Maximum likelihood estimation is a powerful technique because it produces estimators with good properties in a wide range of settings. These properties include consistency, asymptotic unbiasedness, asymptotic efficiency (attainment of the Cramer-Rao lower variance bound), and asymptotic normality; subject to the fulfilment of "regularity conditions". These conditions include 1. Indentifiability - the CDF $F(x;\theta)$ satisfies $\theta \neq \theta \Rightarrow F(x;\theta) \neq F(x;\theta')$ for all x 2. The sample sapce of the PDF does not depend on θ. 3. The true $\theta^\star$ is not on the boundary of the domain of θ. 4. The PDF $f(x;\theta)$ is three-times differentiable in θ, and for all θ in a neighborhood of $\theta^\star$ the function $|\frac{\partial^3}{\partial\theta^3} \log f(X;\theta)|$ is bounded by a function $M(X)$ with finite mean. 5. We can differentiate $\int f(x;\theta)dx$ twice wr.t. θ by exchanging the order of integration and differentiation (limits).

Some of these conditions can be weakened, depending on the property one is trying to prove, and in some cases by making more advanced arguments, but these are the conditions we will use to establish asymptotic normality below.

Proof sketch of asymptotic normality:

Define the loglikelihood function $\ell(\theta) := \log L(\theta; X_1, \dots, X_n)$ and expand its first derivative in Taylor series about the MLE $\hat{\theta}_n$ as follows:

$$\ell'(\hat{\theta}_n) = \ell'(\theta^\star) + (\hat{\theta}_n - \theta^\star)\ell''(\theta^\star) + \tfrac{1}{2}(\hat{\theta}_n - \theta^\star)^2\ell'''(\tilde{\theta}),$$

where $\tilde{\theta}$ is some value between $\hat{\theta}_n$ and $\theta^\star$ by Taylor's theorem. Rearranging terms we have

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) = \frac{n^{-1/2}\ell'(\theta^\star)}{-n^{-1}\ell''(\theta^\star) - (2n)^{-1}(\hat{\theta}_n - \theta^\star)\ell'''(\tilde{\theta})}.$$

The CLT implies the numerator on the RHS above converges in distribution to $N(0, I(\theta^\star))$. The weak LLN implies the first term in the denominator converges to $I(\theta^\star)$ in probability. By the fourth regularity condition and the weak LLN the term $\frac{1}{n}\ell'''(\tilde{\theta})$ in the denominator converges to $E(M(X))$ and hence, is *bounded in probability.* This term is multiplied by $(\hat{\theta}_n - \theta^\star)$ which converges to zero by consistency, and hence this last product term in the denominator converges to zero in probability. Together, these bounds imply

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) \to N(0, I(\theta^\star)^{-1}).$$

The formal support for this last claim is provided by Slutsky's Theorem which we'll not cover here. We also used consistency in this proof sketch which we haven't shown. This latter result provides the additional and stronger properties of asymptotic unbiasedness, efficiency, and normality.

Chapter 4

Sampling Distributions

4.1 Sample Mean

Let $X_1, X_2, \dots, X_n$ denote a random sample from a distribution/population with a finite mean μ and a finite variance σ^2. Denote the sample mean by $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. According to the Central Limit Theorem,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \overset{i.d.}{\to} N(0,1)$$

for large sample size, meaning the distribution of the centered and scaled (standardized) sample mean converges to standard normal. Therefore, for large sample sizes,

$$\overline{X}_n \mathrel{\dot\sim} N(\mu, \sigma^2/n)$$

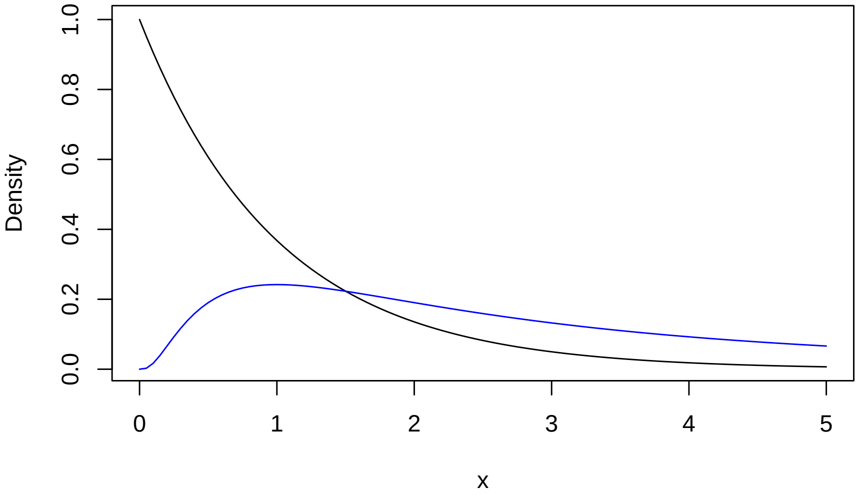
We say the *approximate/large-sample sampling distribution* of the sample mean is Normal.

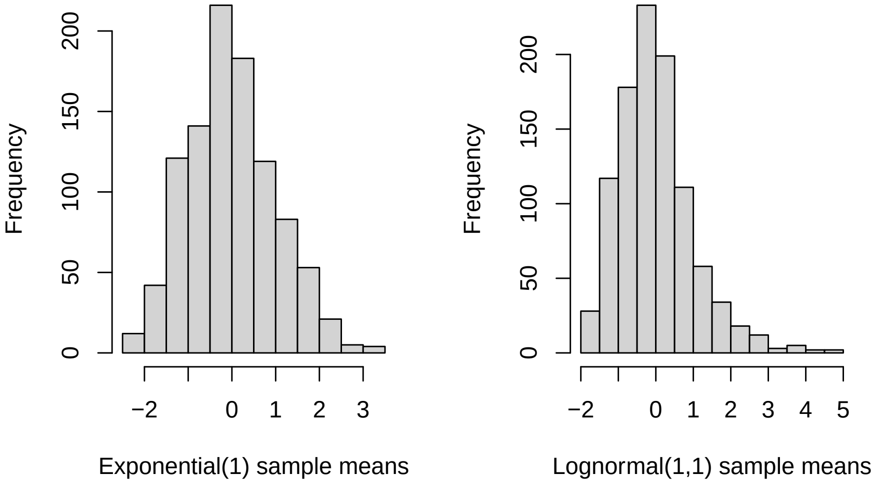
If $X_1, X_2, \dots, X_n$ constitutes a random sample from a Normal population, then the sampling distribution of the sample mean is *exactly* normal:

$$\overline{X}_n \sim N(\mu, \sigma^2/n) \quad \text{for all } n,$$

which can be verified using MGFs.

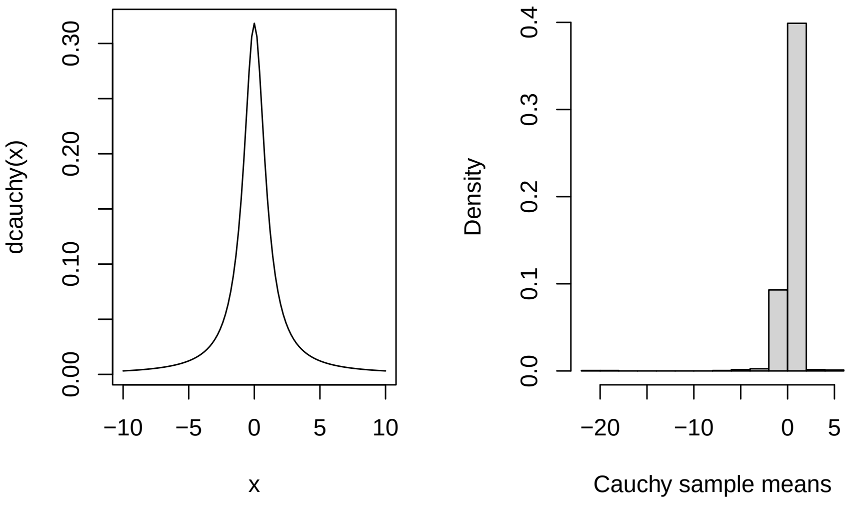
How large is large? For many disributions the CLT will “kick in” for modest sample size, even, say, $n = 50$ is often sufficient. But, ultimately, the number of samples needed in order for the sample mean to be nearly normally distributed depends on the underlying distribution. Very skewed or multimodal distributions may “resist” the CLT much longer than distributions that already are close to normal. Below is a quick Monte Carlo experiment illustrating the CLT, and sample mean sampling distribution, for two distributions: exponential and log-normal.





As a final reminder, note the CLT does not apply to "heavy-tailed" distributions that lack a mean. For example, the Cauchy distribution is very likely to produce extreme values, and, as a result, it has no finite mean. Therefore, the

CLT does not apply and it's unclear what will be the sampling distribution of it's sample mean. It turns out the sample mean is still Cauchy-distributed—taking the average of Cauchy samples does not reduce uncertainty at all. For an illustration, see the Monte-Carlo-approximated distribution of the sample mean of 50 samples from a Cauchy population:



4.2 Sample Variance

Let $X_1, X_2, ..., X_n$ denote a random sample from a Normal population with mean μ and variance σ^2. Denote the sample variance by $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Next, we will discover the sampling distribution of the (normal) sample variance.

First, we need to establish the fact that if $Z \sim N(0,1)$, then Z^2 has what is called a Chi-squared distribution, with one "degree of freedom"—this is the parameter of this distribution. This Chi-squared distribution has MGF $(1-2t)^{-k/2}$ where k is the degrees of freedom parameter. It's corresponding density function is

$$f(x) = 2^{-k/2}\left[\Gamma(k/2)\right]^{-1} x^{k/2-1} e^{-x/2}, \quad x > 0$$

and you might recognize this as a Gamma distribution with shape $k/2$ and rate $1/2$, so it's mean is $(k/2)/(1/2) = k$. Ww can establish this fact by the MGF technique. Evaluate the MGF of Z^2 as follows:

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

where the last inequality follows from the fact the integrand is a normal density function with mean zero and variance $\frac{1}{1-2t}$ (provided $t < 2$) so that the integral equals one. This shows Z^2 is Chi-squared distributed with 1 degree of freedom.

As we've many times, the MGF argument works particularly well with sums of independent random variables. If $Z_1, Z_2, \ldots, Z_n$ are iid standard normal, then $\sum_{i=1}^n Z_i^2$ is Chi-squared distributed with n degrees of freedom. To show this, simply show the MGF of $\sum_{i=1}^n Z_i^2$ equals $(1-2t)^{-n/2}$ using the previous result for $n = 1$, the definition of the MGF, and independence.

Next, we'll show a linear transformation of the the sample variance has a Chi-squared distribution with $n-1$ degrees of freedom.

Start by making the linear transformation

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^n\left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2.$$

Here's the main trick: add and subtract μ inside the square in the sum, then expand the square and simplify:

$$\begin{aligned}
\frac{(n-1)S_n^2}{\sigma^2} &= \sum_{i=1}^n\left(\frac{X_i-\mu+\mu-\overline{X}_n}{\sigma}\right)^2 \\
&= \sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2 + 2\sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{\mu-\overline{X}_n}{\sigma}\right) + \sum_{i=1}^n\left(\frac{\mu-\overline{X}_n}{\sigma}\right)^2 \\
&= \sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2 - 2n\left(\frac{\overline{X}_n-\mu}{\sigma}\right)^2 + n\left(\frac{\overline{X}_n-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2 - n\left(\frac{\overline{X}_n-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^n\left(\frac{X_i-\mu}{\sigma}\right)^2 - \left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2
\end{aligned}$$

So far, we have shown

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}_n}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2.$$

The first term on the right hand side is a summation of n squares of independent, standard normal r.v.'s $Z_i=\frac{X_i-\overline{X}_n}{\sigma}$. Baed on our previous result, this sum must be distributed as Chi-squared with n degrees of freedom. The second term on the right is also the square of a standard normal random variable, so it must be distributed as Chi-squared with one degree of freedom. Consequently, the sum on the left hand side must be a Chi-squared random variable with $n-1$ degrees of freedom.

4.2.1 Large-sample sampling distribution of sample variance

If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with at least 4 finite moments, then the sample variance S_n^2 is approximately normally distributed for large n. To see this, note

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\approx\frac{1}{n}\sum_{i=1}^{n}Z_i$$

where $Z_i=(X_i-\mu)^2$. The latter expression is a sample average of iid random variables, and, as such, the CLT implies (the centered and scaled transformation of) it has a limiting normal distribution. This reasoning can be made rigorous.

4.3 Sampling distribution of studentized sample mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2; and, let $\overline{X}_n$ and S_n^2 denote the sample mean and sample variance, respectively.

Let

$$T_n:=\frac{\overline{X}_n-\mu}{\sqrt{S_n^2/n}}$$

denote the "studentized sample mean". T_n is the standardized sample mean—$\overline{X}_n$ minus μ and divided by $\sqrt{\sigma^2/n}$—with the true variance replaced by its estimate, the sample variance.

The standardized sample mean has a standard normal distribution (even approximately so for non-normal random samples by the CLT), but the studentized sample mean is *not* normally-distributed. Rather, the studentized sample mean has a Student's t distribution with $(n-1)$ degrees of freedom.

A Student's t random variable is defined in the following way: If Z is standard normal and independent of $V \sim$ Chi-Squared $(n-1)$, then

$$T := \frac{Z}{\sqrt{V/(n-1)}}$$

is a Student's t random variable with $n-1$ degrees of freedom.

Rewriting the studentized sample mean, we have

$$T_n = \frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{(n-1)S_n^2/\{(n-1)\sigma^2\}}} = \frac{Z}{\sqrt{V/(n-1)}}.$$

It remains to verify $\overline{X}_n$ and S_n^2 are independent, and a sketch of this result is given next.

4.3.1 Part of Student's Theorem - Indepndence of $\overline{X}_n$ and S_n^2

Let $Z_1, \ldots, Z_n$ be a random sample of standard normal random variables. Without loss of generality let's sketch a proof that $\overline{Z}_n$ and $W = \sum_{i=1}^n (Z_i - \overline{Z}_n)^2$ are independent. Define the matrix O_n to be the $n \times n$ matrix with entries defined by:

$$For 1 \le i \le n-1, \quad o_{ij} = \begin{cases} \frac{1}{\sqrt{i(i+1)}}, & j \le i \\ -\frac{1}{\sqrt{i(i+1)}}, & j = i+1 \end{cases}$$

and zero otherwise; except that $o_{nj} = n^{-1/2}$. Then, it can be checked that O_n is *orthogonal*, i.e., $O_n^\top O_n = I_n$, the identity. Next, define the vectors $Z = (Z_1, \ldots, Z_n)^\top$ and $Y = O_n Z$. By construction of O_n we have

$$Y^\top Y = Z^\top O_n^\top O_n Z = Z^\top Z = \sum_{i=1}^n Z_i^2.$$

And, we also have

$$Y_n = \sum_{i=1}^n \frac{Z_i}{\sqrt{n}} = \sqrt{n}\overline{Z}_n.$$

Therefore,

$$\sum_{i=1}^{n-1} Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_n^2 = Z^\top Z - n\overline{Z}_n^2 = W.$$

So far, we have shown $\overline{Z}_n$ is a function of Y_n and W is a function of $Y_1, \ldots, Y_{n-1}$. But, $Y = O_n Z$ is normally distributed with covariance $O_n^\top O_n = I_n$. Therefore, Y_i, $i = 1, \ldots, n$, are independent, which implies $\overline{Z}_n$ and W are independent.

4.4 Differences of Sample Means

So far we have considered statistics baed on a single random sample. Now, we'll consider statistics based on separate random samples from two populations. First, consider a random sample $X_1, X_2, \dots, X_n$ from a normal distribution with mean and variance μ_X and σ_X^2 and another random sample $Y_1, Y_2, \dots, Y_m$ from a different normal distribution with mean and variance μ_Y and σ_Y^2.

4.4.1 Standarized difference

The standardized difference of means is given by

$$Z = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

and follows a standard normal distribution.

4.4.2 Studentized difference, equal variances

Suppose $\sigma_X^2 = \sigma_Y^2$, and define the *pooled sample variance* $S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$. The studentized difference of sample means is defined by

$$T = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}}$$

follows a Student's t distribution with $n + m - 2$ degrees of freedom.

4.4.3 Studentized difference, unequal variances

When the population variances are unequal a different version of the studentized difference of means is sometimes used:

$$T = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$

This difference does not have an exact t distribution, but can be approximated well by a t distribution with degrees of freedom given by Satterthwaite's formula:

$$df = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2}.$$

Since the degrees of freedom is an integer-valued parameter the result of this formula is *rounded down* to the nearest integer.

4.5 Ratios of Sample Variances

As above, consider a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with mean and variance μ_X and σ_X^2 and another random sample $Y_1, Y_2, \ldots, Y_m$ from a different normal distribution with mean and variance μ_Y and σ_Y^2.

If $U \sim$ Chi-Squared with df $n-1$ and $V \sim$ Chi-Squared with df $m-1$ and U and V are independent, then

$$F := \frac{U/(n-1)}{V/(m-1)}$$

has an F distribution with two degrees of freedom parameters $df1 = n-1$ and $df2 = m-1$, often called the numerator and denominator degrees of freedom.

If S_X^2 and S_Y^2 are sample variances, then

$$\frac{\frac{(n-1)S_X^2}{\sigma_X^2(n-1)}}{\frac{(m-1)S_Y^2}{\sigma_Y^2(m-1)}} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

has an $F(n-1, m-1)$ distribution.

Chapter 5

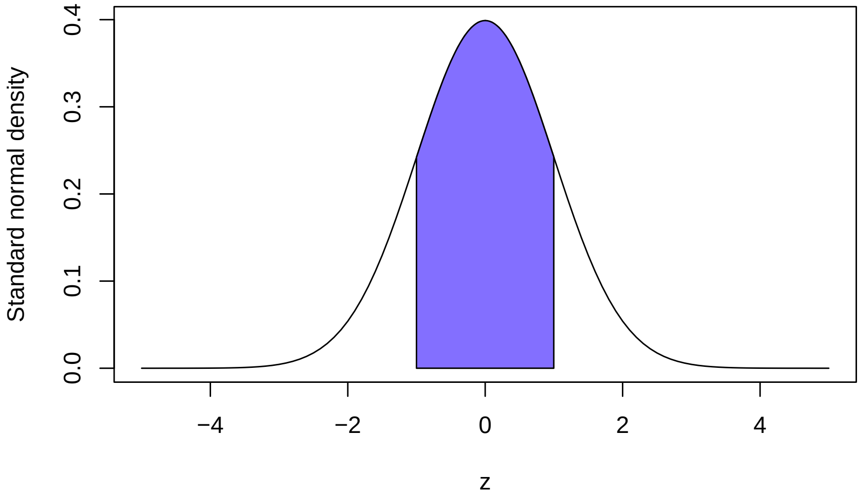
Confidence Intervals

5.1 Why want interval-valued estimates?

Point estimators, like the sample mean, certainly help to describe features of the population. But, the trouble with point estimators is that, on their own, they provide no sense of uncertainty. On one hand, a point estimator is just one number, and one number seems very precise. So, in a sort of psychological sense (and I'm being very loose with that term) people may feel a false sense of certainty about a point estimate. For example, if I say "the sample mean is 5" someone could very reasonably expect the population mean is quite close to 5, but, given a little intuition, that person's opinion may change quite a bit depending on whether the sample size is, say, 10 or 100. On the other hand, in a mathematical statistical sense, we know an estimator is a random variable with a distribution that, usually, has a finite variance. And, if known, we could use knowledge of that distribution to quantify the variability (read uncertainty) inherent in the estimator. So, besides the point estimate itself, it would be helpful to report some measure of variability of the estimate to give the user a more informed view of the estimate of the population mean (or other feature being estimated).

One way to include information of the uncertainty of a point estimate is to include an estimate of the estimator's standard deviation—this is usually called a "standard error". For example, the sample mean $\overline{X}_n$ estimates the population mean μ and if the population has a finite variance σ^2 then the standard deviation of $\overline{X}_n$ is $\sigma/\sqrt{n}$. Usually, σ^2 is unknown, so it is replaced by a point estimate, say, the usual sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ to get the standard error $\sqrt{S_n^2/n}$. A common practice is to report the estimate $\overline{x}_n$ along with the standard error $s_n/\sqrt{n}$.

One drawback of reporting an estimate and its standard deviation is that it's not clear how much variability one standard deviation really represents; it depends on the sampling distribution of the estimator. For example, if an estimator is normally distributed then an interval of one standard deviation about its mean contains about 68% of the distribution, and see the figure below. One the other hand, suppose an estimator has a Chi-squared distribution with 5 degrees of freedom. Then an interval of radius one standard deviation about its mean contains about 72% of the distribution. The point is that one standard deviation does not provide an objective summary of uncertainty across different distributions.



To remedy this issue of the vagueness of standard deviation/standard error we introduce the concept of a confidence interval. An interval-valued estimate is an interval or range of values representing a set that may plausibly contain the true parameter value. Let's denote this interval by (ℓ, u), its lower and upper bounds. Then, a reasonable way to define the endpoints is to select them so that the interval has the *confidence property*

$$P(\ell \leq \theta \leq u) = 1 - \alpha$$

where θ is the true parameter value and $\alpha \in (0, 1)$ is up to the user to decide—usually α is near zero. The probability statement is with respect to ℓ and u, meaning these endpoints are random variables—in particular, they should be statistics, i.e., depended on the data. Then, a *confidence interval* is a random interval with a prescribed chance of "catching" the true parameter.

5.2 Normal population mean example

Suppose $X_1, \dots, X_n$ is a random sample from $N(\mu, 1)$. Let's find a $100(1-\alpha)\%$ confidence interval (CI) for μ. Since $\overline{X}_n$ is a great estimator of μ it makes sense the endpoints of the interval would depend on $\overline{X}_n$. The confidence interval is supposed to be a range of plausible values of μ; and, $\overline{X}_n$ seems the most plausible, so the interval ought to look like $(\overline{X}_n - c_1, \overline{X}_n + c_2)$; in other words, the interval is the sample mean plus or minus something extra. We known $\overline{X}_n \sim N(\mu, 1/n)$, which is a symmetric distribution. So, it would make sense for our $\pm$ something extra to be symmetric, i.e. the interval should have the form $(\overline{X}_n - c, \overline{X}_n + c)$ for some $c > 0$. Now, it just remains to specify c such that the interval has the confidence property. We want

$$P(\overline{X}_n - c \le \mu \le \overline{X}_n + c) = 1 - \alpha.$$

Subtract $\overline{X}_n$ and standardize to get the standard normal probability

$$P\left(-\frac{c}{1/\sqrt{n}} \le \frac{\overline{X}_n - \mu}{1/\sqrt{n}} \le \frac{c}{1/\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-\frac{c}{1/\sqrt{n}} \le Z \le \frac{c}{1/\sqrt{n}}\right) = 1 - \alpha.$$

This means $P(Z \le \frac{c}{1/\sqrt{n}}) = 1 - \alpha/2$ by symmetry of the standard normal distribution. In other words, $\frac{c}{1/\sqrt{n}}$ is the value such that

$$1 - \alpha/2 = F_Z(\frac{c}{1/\sqrt{n}}) = \int_{-\infty}^{\frac{c}{1/\sqrt{n}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

This defines $\frac{c}{1/\sqrt{n}}$ as the $1 - \alpha/2$ standard normal quantile. Given a choice of α it's not easy to solve for c in the above (integral) equation, but we can use the built-in R function qnorm to find this value. Let $\alpha = 0.025$, for example, so that the probability our interval contains μ is 95%. Then, $\frac{c}{1/\sqrt{n}} = qnorm(0.975) = 1.96$ and $c = 1.96\frac{1}{\sqrt{n}}$. Conclude a 95% CI for μ is given by

$$\left(\overline{X}_n - 1.96\frac{1}{\sqrt{n}}, \overline{X}_n - 1.96\frac{1}{\sqrt{n}}\right).$$

More generally, if we denote the $1 - \alpha/2$ quantile of the standard normal by $z_{1-\alpha/2}$ then

$$\left(\overline{X}_n - z_{1-\alpha/2}\frac{1}{\sqrt{n}}, \overline{X}_n - z_{1-\alpha/2}\frac{1}{\sqrt{n}}\right)$$

defines a $100(1-\alpha)\%$ CI for μ.

Remark: It's not essential that the above CI (or any CI) is symmetric about a point estimator, but when the sampling distribution of the estimator is symmetric this is the best choice. To see this, note that

$$\left(\overline{X}_n - z_{1-\alpha/3}\frac{1}{\sqrt{n}}, \overline{X}_n - z_{1-2\alpha/3}\frac{1}{\sqrt{n}}\right)$$

is also a 95% CI for μ, but it is *always* larger (wider) than the symmetric interval. So, the symmetric interval, which being shorter is more precise, is preferable.

5.3 Other "Exact" CIs for normal population mean and variance parameters

5.3.1 Population mean, unknown variance

Let $X_1, \ldots, X_n$ be a random sample of size $n \geq 2$ from $N(\mu, \sigma^2)$ where both mean and variance parameters are unknown. We know the studentized mean follows a Student's t distribution with $n-1$ degrees of freedom:

$$T = \frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \sim t(n-1).$$

Mirroring the argument from above when σ^2 is a known value, we assume a $100(1-\alpha)\%$ CI for μ would be symmetric about $\overline{X}_n$ with radius $t_{1-\alpha/2}(n-1)\frac{S_n}{\sqrt{n}}$ where $t_{1-\alpha/2}(n-1)$ is the $1-\alpha/2$ quantile of the Student's t distribution with $n-1$ degrees of freedom:

$$P\left(\overline{X}_n - t_{1-\alpha/2}(n-1)\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{1-\alpha/2}(n-1)\frac{S_n}{\sqrt{n}}\right) = 1-\alpha.$$

5.3.2 Population variance, unknown mean

Let $X_1, \ldots, X_n$ be a random sample of size $n \geq 2$ from $N(\mu, \sigma^2)$ where both mean and variance parameters are unknown. We know the following transformation is Chi-squared distributed:

$$\frac{(n-1)S_n^2)}{\sigma^2} \sim \chi^2(n-1).$$

Now, the Chi-squared distribution is not symmetric—it's skewed. There are a number of ways to define a $100(1-\alpha)\%$ CI for σ^2 based on the above sampling distribution. Since $\sigma^2 > 0$ we can define a one-sided interval starting from zero:

$$\left(0, \frac{(n-1)S_n^2}{\chi_\alpha^2(n-1)}\right)$$

where $\chi_{1-\alpha}^2(n-1)$ is the $100(1-\alpha)$ quantile of a Chi-squared r.v. with $(n-1)$ degrees of freedom. Then,

$$\begin{aligned} P(0 \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi_\alpha^2(n-1)}) &= P(0 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{1-\alpha}^2(n-1)) \\ &= P(\chi^2(n-1) \geq \chi_\alpha^2(n-1)) \\ &= 1-\alpha \end{aligned}$$

by definition of the quantile.

We can define two-sided intervals as well. The "equi-tailed" interval is not the shortest possible interval, but it has a simple form:

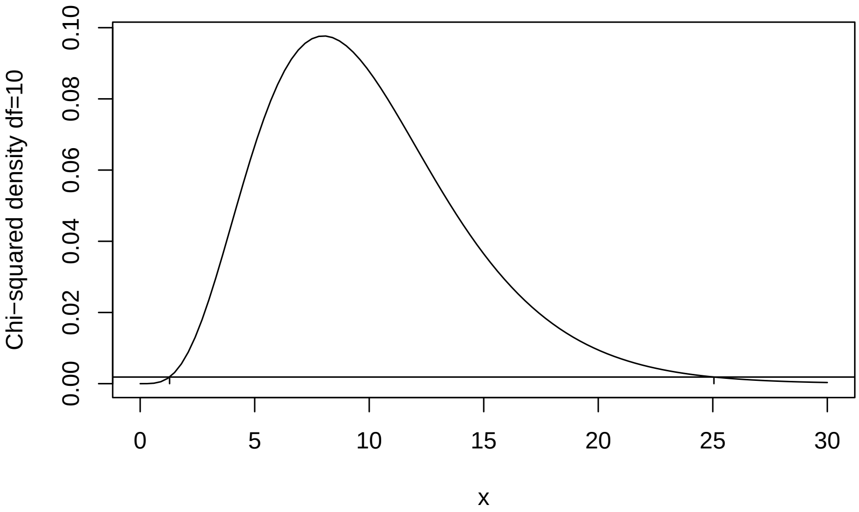
$$\left(\frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)S_n^2}{\chi^2_{\alpha/2}(n-1)}\right).$$

In order to find the shortest possible $100(1-\alpha)\%$ CI one could draw a horizontal line slicing across the Chi-squared density such that the area under both the line and the density equals α. Then, the intersection points of the line and the density would give the lower and upper Chi-squared quantiles to use in the two-sided interval. Since the density is not symmetric this interval generally would not be based on the $1-\alpha/2$ and $\alpha/2$ quantiles like the equi-tailed interval. For example, the shortest 95% CI when $df = 10$ is based on quantile values of about 1.284 and 25.05. The following plot illustrates this.

```
## 0.04998856 with absolute error < 8.5e-05
```

```
## [1] 0.001862438
```

```
## [1] 0.001863508
```



5.3.3 Two normal samples, comparing means

Let's tackle the next topic by way of an example. Suppose infant walking time — the age in months at which babies can walk on their own — is normally distributed. Generally, parents are excited for their infants to begin walking and some parents "take steps" to encourage their infants to walk. We'll assume it is reasonable to model the population of infant walking times among those infants whose parents no not intervene in walking skills development as a random sample from a normal distribution with mean μ_x and variance $sigma_x^2$; likewise we model the corresponding population of infant walking times influenced by parent behavior by a normal distribution with mean μ_y and variance $sigma_y^2$. Assume this parental influence factor represents a more-or-less standard "treatment"; for example, parents might spend 60 minutes a day encouraging their infant to crawl and stand with assistance or play in a bounce seat to build lower body strength. Child development researchers would be interested to know whether such parental interventions influence infant walking times.

Given observations from an experiment comparing these two groups of infants, how would we answer the researchers' question? One answer comes in the form of a CI for the difference of means $\mu_x - \mu_y$. In this case a negative difference indicates some impact of parental involvement. And, if a 95% CI does not contain zero, then the CI reflects strong evidence in the data that the true means are different.

Based on our knowledge of sampling distributions we know that if $\sigma_x^2 = \sigma_y^2$ the following statistic has a Student's t distribution with $n+m-2$ df where n and m are the sample sizes in each group:

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

where S_p^2 denotes the pooled (sample-size weighted average) sample variance. As a result, the following probability statement is exact:

$$\begin{aligned}
1-\alpha &= P\left(t_{\alpha/2,n+m-2} \le \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \le t_{1-\alpha/2,n+m-2}\right) \\
&= P\left(t_{\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} \le \overline{X} - \overline{Y} - (\mu_x - \mu_y) \le t_{1-\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}\right) \\
&= P\Bigg(-(\overline{X} - \overline{Y}) + t_{\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} \le -(\mu_x - \mu_y) \le -(\overline{X} - \overline{Y}) + t_{1-\alpha/2,n+m-2}\sqrt{S_p^2\Bigg(\frac{1}{n}} \\
&= P\Bigg((\overline{X} - \overline{Y}) - t_{1-\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} \le (\mu_x - \mu_y) \le (\overline{X} - \overline{Y}) + t_{1-\alpha/2,n+m-2}\sqrt{S_p^2\Bigg(\frac{1}{n} +}
\end{aligned}$$

where the last line follows from multiplying by -1 and noting $t_{1-\alpha/2,n+m-2} = -t_{\alpha/2,n+m-2}$ by symmetry. Therefore, a $100(1-\alpha)\%$ CI for the difference of means $\mu_x - \mu_y$ is given by

$$\left((\overline{X}-\overline{Y}) - t_{1-\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)},\, (\overline{X}-\overline{Y}) + t_{1-\alpha/2,n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}\right).$$

Suppose we conduct the infant walking time experiment and observe the following data:

$X:$ 13.2 16.6 16.7 16.8 17.0 17.1 17.3 17.4 17.4 17.7 17.9

$Y:$ 16.8 17.1 17.6 17.7 18.8 18.8 19.3 19.6 21.0 21.2 21.5 22.1 22.8

Computing the sample statistics we find $\overline{x} = 16.83$, $\overline{y} = 19.56$, $s_x^2 = 1.61$, and $s_y^2 = 3.97$. The pooled sample variance is $(10s_x^2 + 12s_y^2)/22 = 2.90$. The 0.975 quantile of the Student's t distribution with 22 df can be found using “qt(0.975, 22)” and equals about 2.07. Plugging all these values into the CI formula, the 95% CI for the difference in mean infant walking times is

$$(-4.18,\ -1.29)$$

We conclude the mean infant walking times in the group with parental intervention were between about 1.3 and 4.2 months shorter than in the non-intervention group. (Note this is made-up “data” only for illustration purposes.)

The careful reader may be concerned that our CI method used above relies on the two normal populations having identical variances but our sample variances are substantially different. This equal-variance CI procedure is often referred to as Students's two-sample CI, in contrast to an alternative method that can be used when variances are unequal called Welch's two-sample CI. Recall the studentized difference of sample means when variances are unequal approximately (not exactly) follows a Student's t distribution with Satterthwaite's choice of degrees of freedom:

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \dot{\sim} t(\nu)$$

where

$$\nu = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2}.$$

Following the same argument as above, this implies an *approximate* $100(1-\alpha)\%$ CI for the difference of means is given by

$$\left((\overline{X}-\overline{Y}) - t_{1-\alpha/2,\nu}\sqrt{\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)},\, (\overline{X}-\overline{Y}) + t_{1-\alpha/2,\nu}\sqrt{S_p^2\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)}\right).$$

Plugging in the relevant sample statistics, we get $\nu = 20$ so that we use the quantile $t_{0.975}(20) = 2.086$. Welch's two-sample approximate 95% interval is

$$(-4.14,\ -1.33)$$

which is hardly different than Student's interval. We'll investigate the comparative performances of these two intervals further in lab.

5.3.4 Two normal samples, comparing variances

In some applications it is either more important or at least of equal importance to compare population variances rather than means. For example, consider one or more manufacturing processes for producing optical lenses—these might be for glasses or contacts. Lenses have many dimensional quantities that determine their properties, for example, the refraction index of the lens material and the thickness of the lens, which both contribute to the power of the lens. Suppose a manufacturing process is supposed to produce a lens of a given thickness. A small tolerance is allowable but any lenses produced with thickness outside the tolerance must be rejected. We can think of the difference between the specified, target lens thickness and the population mean thickness of lenses produced by a specific process as *bias*. Then, our preference between two or more production processes is a question about bias-variance tradeoff. A production process with no bias but a high variance in lens thickness will result in many lenses being rejected—perhaps even more than a manufacturing process that produces biased lenses with very low variability in thickness. Clearly, for choosing between processes we need to evaluate both process bias and process variance.

For comparing two process variances (where the underlying popualtions are normal) we use a ratio statistic. The sampling distribution of the following ratio of sample variances has an *F distribution* when the population variances are equal:

$$\frac{S_x^2}{S_y^2} \sim F(n-1, m-1).$$

Recall that $(n-1)S_x^2/\sigma_x^2 \sim$ Chi-squared$(n-1)$ and $(m-1)S_y^2/\sigma_y^2 \sim$ Chi-squared$(m-1)$. The ratio of two independent, Chi-squared r.v.'s divided by the dfs defines an F random variable with two parameters which are the degrees of freedom of the numerator and denominator Chi-squared r.v.'s. When $\sigma_x^2 = \sigma_y^2$, we have

$$\frac{\frac{(n-1)S_x^2}{\sigma_x^2(n-1)}}{\frac{(m-1)S_y^2}{\sigma_y^2(m-1)}} = \frac{\frac{(n-1)S_x^2}{(n-1)}}{\frac{(m-1)S_y^2}{(m-1)}} = \frac{\text{Chi-squared(n-1)/(n-1)}}{\text{Chi-squared(m-1)/(m-1)}} = \frac{S_x^2}{S_y^2} = F(n-1, m-1).$$

And, when the variances are not equal, we simply have

$$\frac{S_x^2\sigma_y^2}{S_y^2\sigma_x^2} \sim F(n-1, m-1).$$

We can invert this relationship to find a CI for σ_y^2/σ_x^2 using the following probability computation:

$$\begin{aligned}1-\alpha &= P(F_{\alpha/2}(n-1,m-1) \leq \frac{S_x^2\sigma_y^2}{S_y^2\sigma_x^2} \leq F_{1-\alpha/2}(n-1,m-1))\\ &= P(\frac{S_y^2}{S_x^2}F_{\alpha/2}(n-1,m-1) \leq \sigma_y^2/\sigma_x^2 \leq \frac{S_y^2}{S_x^2}F_{1-\alpha/2}(n-1,m-1))\\ &= P(\frac{S_y^2}{S_x^2}\frac{1}{F_{1-\alpha/2}(m-1,n-1)} \leq \sigma_y^2/\sigma_x^2 \leq \frac{S_y^2}{S_x^2}F_{1-\alpha/2}(n-1,m-1)),\end{aligned}$$

where the last line uses a reciprocal-equality property of F distributions with swapped degrees of freedom, namely $F_\alpha(df1, df2) = 1/(F_{1-\alpha}(df2, df1))$. This is not essential if you are using R to compute F quantiles but is helpful for using F tables, which typically only list upper tail quantiles. Then, a $100(1-\alpha)\%$ CI for the ratio σ_y^2/σ_x^2 is given by

$$\left(\frac{S_y^2}{S_x^2}\frac{1}{F_{1-\alpha/2}(m-1,n-1)},\ \frac{S_y^2}{S_x^2}F_{1-\alpha/2}(n-1,m-1)\right).$$

Given the following observations (in nanometers) of lens thickness using two different production methods, compute 95% CIs for the difference in means (use Welch's method) and the ratio of variances:

$$X : 2.9859\ \ 3.0042\ \ 3.0055\ \ 3.0083\ \ 3.0113\ \ 3.0124\ \ 3.0150\ \ 3.0214$$

$$Y : 2.9964\ \ 3.0011\ \ 3.0099\ \ 3.0110\ \ 3.0124\ \ 3.0130\ \ 3.0140\ \ 3.0205$$

The 95% CI for the difference in population mean thicknesses is

$$(-0.0116,\ 0.0080).$$

The 95% CI for the ratio of population variances σ_y^2/σ_x^2 is

$$(0.1058,\ 2.6394).$$

5.4 CIs for proportions

5.4.1 A single Bernoulli proportion

Using the DeMoivre-Laplace CLT we know the standardized sample proportion is approximately normally distributed:

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

Then, again denoting lower standard normal quantiles by z_α, we have

$$1-\alpha \approx P\left(z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{1-\alpha/2}\right).$$

As in other cases covered above, we can derive an approximate CI for p by algebraic manipulations within the probability statement. A complicating factor is that the true proportion p shows up in both the numerator and denominator of the standardized sample proportion. A common technique to get around this is to replace the standard deviation of $\hat{p}$ by the estimated standard deviation $\sqrt{\hat{p}(1-\hat{p})/n}$. Using the estimated standard deviation, an approximate $100(1-\alpha)\%$ CI for p is given by

$$\left(\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

It is possible to derive a CI for p without resorting to using the estimated standard deviation. In that case, we have to solve for p in the quadratic equation

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = z_{1-\alpha/2}.$$

The endpoints of the CI based on the true standard deviation of $\hat{p}$ are given by

$$\frac{n\hat{p}+0.5z^2_{\alpha/2} \pm z_{1-\alpha/2}\sqrt{n\hat{p}(1-\hat{p})+0.25z^2_{\alpha/2}}}{n+z^2_{\alpha/2}}.$$

Example: A survey is conducted to determine the level of support for the construction of a new nuclear power plant in a community. 140 of 400 randomly sampled voters favor the construction project. Find a 99% approximate CI for the true proportion of voters favoring the project. The sample proportion is $\hat{p} = 0.35$ and the estimated standard deviation of $\hat{p}$ is $\sqrt{0.35 \cdot 0.65/400} = 0.02384848$. Given $z_{0.995} = 2.575$ an approximate 99% CI for p is given by

$$(28.86\%, 41.14\%).$$

Alternatively, using the formula for the CI based on the true standard deviation of $\hat{p}$ we find the interval

$$(29.15\%, 41.34\%).$$

5.4.2 Difference of two Bernoulli proportions

The most common experiments compare two or more populations. For two Bernoulli populations an approximate CI for the difference of population proportions can be computed based on the DeMoivre-Laplace CLT similarly to the single population interval above based on the estimated standard deviation of the sample proportion. Our CI for the difference of proportions is given by

$$\left(\hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\quad \hat{p}_1 - \hat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right).$$

Example: Suppose a poll of Illinois voters finds 132 of 200 male voters and 90 of 150 female voters favor a certain candidate. A 99% CI for the difference of population favorability proportions between the sexes is

$$(-7.4\%,\ 19.4\%)$$

indicating there is, plausibly, no difference in the candidate's favorability between the sexes.

5.5 Approximate CIs based on MLEs

So far we have derived CIs for parameters based on sampling distributions of statistics used to estimate those parameters. We can use the same strategy to derive CIs for parameters estimated using maximum likelihood whenever asymptotic normality of the corresponding MLE is justified. Recall there are a few cases where asymptotic normality may not hold, such as estimating the lower or upper bounds of a uniform distribution—that case violates the condition that the sample space does not depend on the parameter. However, in many situations asymptotic normality of the MLE does hold, and we can use it to derive approximate CIs.

Recall that, under some regularity conditions, the MLE $\hat{\theta}$ for the scalar parameter θ is approximately normally distributed with mean θ and variance $[nI(\theta)]^{-1}$ where $I(\theta)$ is the Fisher information for one sample. If we replace the variance by the estimated variance by plugging in $\hat{\theta}$ then we obtain the CI

$$\left(\hat{\theta} \pm z_{1-\alpha/2}[nI(\hat{\theta})]^{-1/2}\right),$$

where you'll recall that $I(\theta) = E\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$ or, equivalently (most of the time), $I(\theta) = -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right)$ where ℓ denotes the loglikelihood function.

Example: The MLE of the exponential rate parameter is $\hat{\theta} = \overline{X}$. We have previously found $[nI(\theta)]^{-1/2} = \sqrt{(\overline{X})^2/n}$ so that an MLE-based approximate

CI for the exponential rate θ is given by

$$\left(\overline{X} \pm z_{1-\alpha/2}\frac{\overline{X}}{\sqrt{n}}\right).$$

5.5.1 Multivariate case

When θ is a vector parameter the MLE $\hat{\theta}$ is approximately distributed as a multivariate normal random vector with mean vector θ and (estimated) covariance matrix $[nI(\hat{\theta})]^{-1}$. We can compute approximate CIs for each element of θ using the marginal approximate normal distributions of the elements of $\hat{\theta}$ as in the scalar case. However, we can also compute a joint confidence region for two or more elements of θ. The elliptical region

$$\{\theta : (\hat{\theta} - \theta)[nI(\hat{\theta})]^{-1}(\hat{\theta} - \theta)^{\top} \le z^2_{1-\alpha/2}\}$$

is an approximate $100(1-\alpha)\%$ confidence region for θ. It is also smaller than the set product of marginal confidence intervals, which is a hyper-rectangle containing this ellipse.

Example: The MLEs for a normal population mean and variance have approximate normal distribution

$$(\overline{X}, \hat{\sigma}^2) \dot{\sim} N\left((\mu, \sigma^2), \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}\right).$$

An approximate $100(1-\alpha)\%$ elliptical confidence region for (μ, σ^2) is given by

$$\{(\mu, \sigma^2) : (\overline{X} - \mu, \hat{\sigma}^2 - \sigma^2)\begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix}^{-1}(\overline{X} - \mu, \hat{\sigma}^2 - \sigma^2)^{\top} \le z^2_{1-\alpha/2}\}.$$

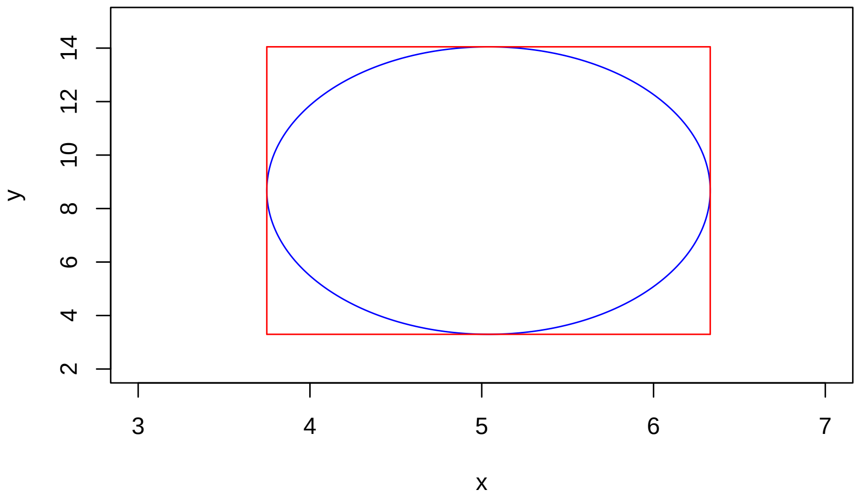
```
set.seed(56482)
data <- rnorm(20, 5, 3)
xbar <- mean(data)
s2 <- var(data)*(19)/20

xc <- xbar # center x_c or h
yc <- s2 # y_c or k
a <- 1.96*sqrt(s2/20) # major axis length
b <-  1.96*sqrt(2*(s2^2)/20) # minor axis length
phi <- 0 # angle of major axis with x axis phi or tau

t <- seq(0, 2*pi, 0.01)
x <- xc + a*cos(t)*cos(phi) - b*sin(t)*sin(phi)
y <- yc + a*cos(t)*sin(phi) + b*sin(t)*cos(phi)
plot(x,y, col='blue',  type = 'l', xlim = c(3,7), ylim = c(2,15))
```

```
CI.mu <- c(xbar - 1.96*sqrt(s2/20), xbar + 1.96*sqrt(s2/20))
CI.sig2 <- c(s2 - 1.96*sqrt(2*(s2^2)/20), s2 + 1.96*sqrt(2*(s2^2)/20))

lines(c(CI.mu[1], CI.mu[2]),c(CI.sig2[1], CI.sig2[1]), col = 'red')
lines(c(CI.mu[1], CI.mu[2]),c(CI.sig2[2], CI.sig2[2]), col = 'red')
lines(c(CI.mu[1], CI.mu[1]),c(CI.sig2[1], CI.sig2[2]), col = 'red')
lines(c(CI.mu[2], CI.mu[2]),c(CI.sig2[1], CI.sig2[2]), col = 'red')
```



5.5.2 The Delta method

The delta theorem says that asymptotic normality of the MLE is preserved under smooth (differentiable) transformations. Let $g(\cdot)$ be a differentiable function. Then,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \dot{\sim} N\left(0, [g'(\theta)]^\top I(\theta)^{-1} g'(\theta)\right).$$

The power of the delta theorem is its implied method for constructing CIs for any "nice" function of population parameters.

Example: For a normal population experiment a common function of parameters that is of interest to researchers is the so-called "signal-to-noise ratio" given by μ/σ. In this case, $g(\theta) = g(\mu, \sigma^2) = \mu/\sqrt{\sigma^2}$. And, $g'(\theta) = (1/\sqrt{\sigma^2}, -0.5\mu/((\sigma^2)^{1.5}))$. Using the delta theorem result, the MLE of

μ/σ is given by $\overline{X}/\sqrt{\hat{\sigma}^2}$ and has approximate variance

$$(1/\sqrt{\hat{\sigma}^2}, -0.5\overline{X}/((\hat{\sigma}^2)^{1.5})) \begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix} (1/\sqrt{\hat{\sigma}^2}, -0.5\overline{X}/((\hat{\sigma}^2)^{1.5}))^\top = \frac{1}{n} + \frac{\overline{X}^2}{2n\hat{\sigma}^2}.$$

Then, an approximate CI for μ/σ is given by

$$\left(\overline{X}/\sqrt{\hat{\sigma}^2} \pm z_{1-\alpha/2}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{2n\hat{\sigma}^2}} \right).$$

Chapter 6

Hypothesis Testing

So far our discussion of statistical inference has been limited to “estimation”, which is essentially any method that determines which parameter values match/agree with observed data. We have discussed both parametric and non-parametric estimation, although we have not used those terms. Non-parametric estimation refers to estimation of population parameters without specifying any probability model for the data. For example, the sample mean is a reasonable estimator of the population mean (should it exist) regardless of the sampling distribution of the data (the population distribution). Parametric estimators are tied to assumed sampling distributions. For example, maximum likelihood estimators totally rely on an assumed sampling distribution of the data (population distribution). Parametric estimation can be viewed as selecting which particular probability model within a given set (the assumed distribution) best fits the data.

We’ll next expand our discussion of statistical inference to ask a different, but related question. Instead of asking which parameters/model distributions agree with a given data set, we ask “how much evidence does a given data set provide for a particular parameter value or set of parameter values?” The particular parameter values are referred to as the hypothesis or “null hypothesis”. And, a procedure that determines whether the data provides sufficient evidence against the null hypothesis so as to make it implausible is called a “hypothesis test”.

6.1 Notation

Consider a generic parameter θ taking values in some set Θ. The null hypothesis θ takes exactly the value θ_0 is written $H_0 : \theta = \theta_0$. Its complement is the alternative hypothesis $H_a : \theta \neq \theta_0$, which is always the set complement of the null hypothesis. In particular, $H_0 : \theta = \theta_0$ is called a “point-null hypothesis” because the set in the null hypothesis consists of only a single point. In other cases, the null hypothesis is "composite", e.g., $H_0 : \theta \leq \theta_0$ versus $H_a : \theta > \theta_0$.

Example: Suppose an experiment consist of collecting a random sample from a normal population with unknown mean μ and unknown variance σ^2. A point null hypothesis that $\mu = 0$ is phrased $H_0 : \mu = 0$ versus the alternative hypothesis $\mu \neq 0$. A point null hypothesis that $\sigma^2 = 1$ is phrased $H_0 : \sigma^2 = 1$ versus $H_a : \sigma^2 \neq 1$, where the alternative hypothesis tacitly assumes $\sigma^2 > 0$.

6.2 Hypothesis testing outcomes

A hypothesis testing procedure begins with a pair of null and alternative hypotheses. Next, data relevant to the hypotheses is collected, and a decision is made as to whether the data provides sufficient evidence against the null hypothesis such that it may be rejected as implausible. Otherwise, the data are viewed as sufficiently consistent with the null hypothesis as to retain it.

As such, there are four possible outcomes of this procedure. In two cases, the correct decision is made: either the null hypothesis is true and it is retained, or the null hypothesis is false and it is reject. There are two corresponding errors: 1) A Type 1 error is committed when the null hypothesis is true but it is rejected; and, 2) A Type 2 error is committed when the null hypothesis is false and it is retained.

Since the decision to reject or retain the null hypothesis is based on a random sample of data, the decision is itself random; so, each of these outcomes has a certain probability of occuring. Given the null hypothesis is true, the chance of a Type 1 error occurring is denoted α and is also called the Type 1 error rate. Given the null hypothesis is false the chance of a Type 2 error occurring is denoted β, and its complement, the chance of rejecting H_0 is denoted $1 - \beta$ and referred to as the "power of the test".

Of course, it would be nice if one could eliminate any chance of an error occurring, but that is not possible. To see this, consider the extreme case of a non-random hypothesis test that always rejects the null hypothesis, no matter the data. Such a rule minimizes the chance of a Type 2 error (in fact there is such chance) but maximizes the chance of a Type 1 error. And, the opposite extreme, in which the null hypothesis is always retained similarly minimizes Type 1 error at the cost of maximizing Type 2 errors. Any testing procedure in between necessarily has some positive chance of each type of error.

As we will see, it is possible to construct hypothesis tests with an explicit cap (upper bound) on the Type 1 error rate α. This means that, whenever possible, it makes sense to design experiments and hypotheses such that Type 1 errors are more serious than Type 2 errors, so that they may be expressly controlled. A hypothesis test that limits α to a prespecified level is called a "level-α test".

6.3 Tests based on a normal population

6.3.1 Test for a normal mean when the variance is known

Suppose a given population is known to follow a normal distribution with an unknown mean and a known variance σ^2. We wish to test the point null hypothesis $H_0 : \mu = \mu_0$ for some given number μ_0 versus the alternative $H_a : \mu \neq \mu_0$. And, we wish our test to limit the Type 1 error rate to a given level α, say, 5%.

We obtain a random sample of size n from the population. How should we proceed? If H_0 is true, then our data should look like it came from $N(\mu_0, \sigma^2)$; specifically, the sample mean should be "close" to μ_0. If $\overline{X}$ is close to μ_0, then H_0 seems plausible, otherwise not. Our intuition says our test should look like: "reject H_0 if $|\overline{X} - \mu_0| > c$" for some cutoff $c > 0$. The constraint on Type 1 error rate to be no more than α is enough to determine c:

$$\begin{aligned}1 - \alpha &= P(-c \leq \overline{X} - \mu_0 \leq c)\\ &= P(-c/\sqrt{\sigma^2/n} \leq Z \leq c/\sqrt{\sigma^2/n})\end{aligned}$$

This implies we should reject H_0 if $|Z| > z_{1-\alpha/2}$ where

$$Z = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}}$$

is the standardized sample mean assuming H_0 is true (often it is said, "under H_0").

Next, we investigate the power of this test. Recall the power is the probability the test rejects the null hypothesis when the null is false. For most tests, like the current one, the null hypothesis may be false in many ways. If H_0 is false then $\mu \neq \mu_0$, so μ is some other number, lets say y. Intuitively, for y values close to μ_0 we would not necessarily expect to notice H_0 is false, but for y values far from μ_0 we would expect $\overline{X}$ to be far from μ_0 and for the test to reject μ_0. So, we expect the power to be a function of y and to be increasing as a function of $|y - \mu_0|$. Below we compute this power function:

$$\begin{aligned}P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2} | \mu = y\right) &= P\left(\frac{\overline{X} - y}{\sigma/\sqrt{n}} < z_{\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}}\right)\\ &= P\left(Z < z_{\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}}\right).\end{aligned}$$

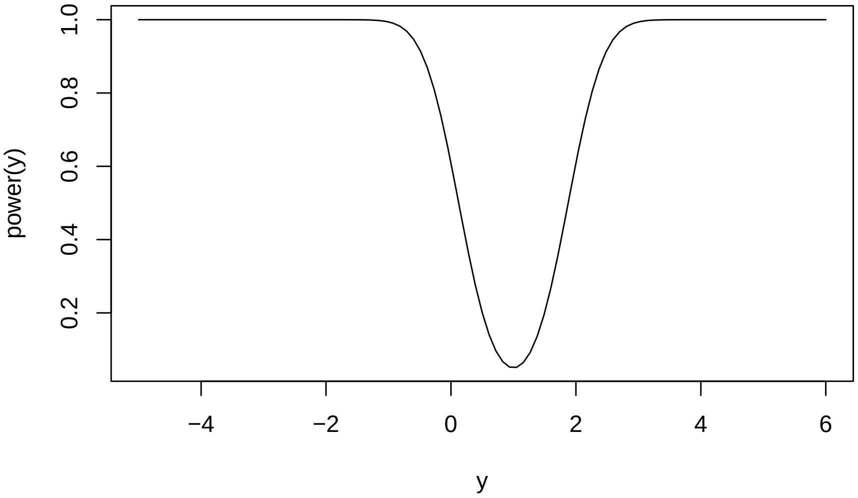
$$\begin{aligned}P\left(z_{1-\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} | \mu = y\right) &= P\left(z_{1-\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}} < \frac{\overline{X} - y}{\sigma/\sqrt{n}}\right)\\ &= P\left(z_{1-\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}} < Z\right)\\ &= 1 - P\left(Z < z_{1-\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}}\right).\end{aligned}$$

Therefore, the power as a function of y is

$$(1-\beta) = P\left(Z < z_{\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}}\right) + 1 - P\left(Z < z_{1-\alpha/2} + \frac{\mu_0 - y}{\sigma/\sqrt{n}}\right).$$

We illustrate this power function below for $H_0 : \mu = 1$ where $\sigma^2 = 2$, $n = 10$, and $\alpha = 0.05$.

```
alpha <- 0.05
mu0 <- 1
sigma2 <- 2
n <- 10
power <- function(y) pnorm(qnorm(alpha/2)+(1-y)/sqrt(sigma2/n))+1-pnorm(qnorm(1-alpha/2
curve(power, from = -5, to = 6, xlab = 'y', ylab = 'power(y)')
```



The power is minimized (and equal to 0.05) at $y = \mu_0$ where the test only rejects at rate α by design. The power increases smoothly and symetrically as a function of $|y - \mu_0|$ until the test has nearly a 100% chance of rejecting H_0 for true mean sufficiently far from μ_0.

6.3.2 Test for a normal mean when the variance is unknown

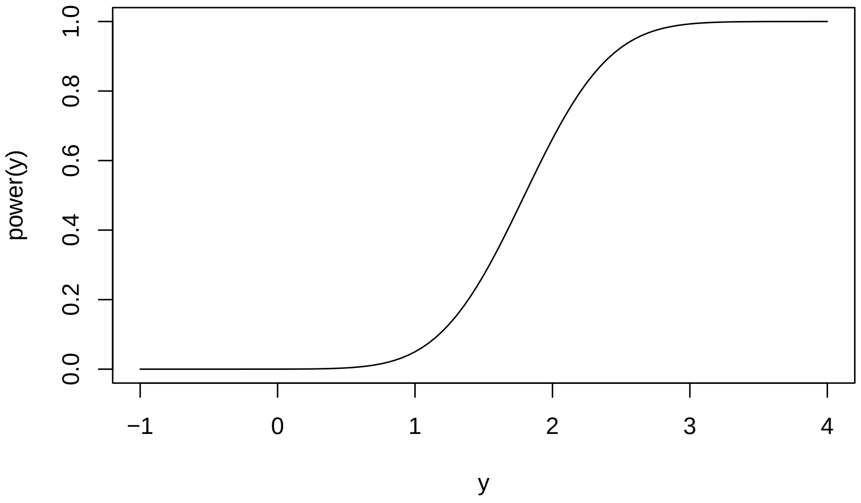
As we have discussed many times when the population variance is unknown the Studentized sample mean has a Student's t distribution with $n - 1$ degrees of freedom. A Student's $t-$based test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ is essentially the same as the above test but based on the T statistic, $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ rather than the Z statistic $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$. A level-$\alpha$ test rejects H_0 if $T > t_{1-\alpha/2}(n-1)$ or $T < t_{\alpha/2}(n-1)$, where $t_\alpha(df)$ is the Student's t lower α quantile with df degrees of freedom.

For a different angle, consider the test of the one-sided (composite) hypotheses $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$. Naturally, values of $\overline{X}$ smaller than μ_0 lend support to the null hypothesis, so a level α test of H_0 only rejects the null if $T > t_{1-\alpha}(n-1)$. Consider the power of this one-sided test. Since the test only rejects for large values of $\overline{X}$, the power should increase as T increases. Mimicking (half of) our power calculation for the Z test above we have

$$1 - \beta = P\left(\frac{\overline{X}-y}{S/\sqrt{n}} + \frac{y-\mu_0}{S/\sqrt{n}} > t_{1-\alpha}(n-1)\right).$$

The random variable $\frac{\overline{X}-y}{S/\sqrt{n}} + \frac{y-\mu_0}{S/\sqrt{n}}$ has a Student's t distribution with $n-1$ degrees of freedom and **non-centrality parameter** $(y-\mu_0)/(\sigma/\sqrt{n})$.

```r
alpha <- 0.05
mu0 <- 1
sigma2 <- 2
n <- 10
power <- function(y)  1-pt(qt(1-alpha, n-1), df = n-1, ncp = (y - mu0)/(sqrt(sigma2/n)))
curve(power, from = -1, to = 4, xlab = 'y', ylab = 'power(y)')
```



```
power(2)
```

```
## [1] 0.6623762
```

Here's a quick simulation to illustrate the above power function: we'll sample data sets of $n = 10$ samples from $N(\mu = 2, \sigma^2 = 2)$ a total of 10000 times. Each time we'll compute the p-value of the test $H_0 : \mu \leq 1$ versus $H_a : \mu > 1$. According to the above power calculation, we should reject H_0 about two thirds of the time if $\alpha = 0.05$.

```
res <- rep(NA,10000)
for(i in 1:10000){
  x <- rnorm(n,2,sqrt(sigma2))
  res[i] = t.test(x,mu = mu0,alternative = "greater")$p.value
}
sum(res<0.05)/10000
```

```
## [1] 0.6626
```

6.3.3 Test for a normal population variance

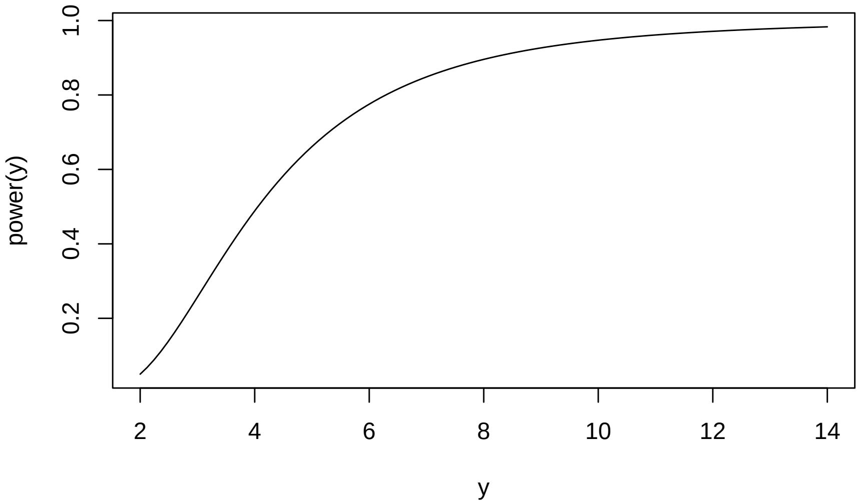
Consider a test that rejects the null hypothesis $H_0 : \sigma^2 \leq \sigma_0^2$ if $\frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{1-\alpha}(n-1)$ where $\chi^2_\alpha(df)$ is the Chi-squared lower α quantile for a Chi-squared distribution with df degrees of freedom. Such a rejection rule defines a level-α test:

$$P\left(\frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{1-\alpha}(n-1) | \sigma^2 = \sigma_0^2\right) = P\left(\chi^2(n-1) \geq \chi^2_{1-\alpha}(n-1)\right) = \alpha,$$

by definition. We can compute the power curve of such a test:

$$\begin{aligned} P\left(\frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{1-\alpha}(n-1) | \sigma^2 = y\right) &= P\left(\frac{(n-1)S^2}{y} \geq \frac{\sigma_0^2}{y}\chi^2_{1-\alpha}(n-1)\right) \\ &= P\left(\chi^2(n-1) \geq \frac{\sigma_0^2}{y}\chi^2_{1-\alpha}(n-1)\right) \end{aligned}$$

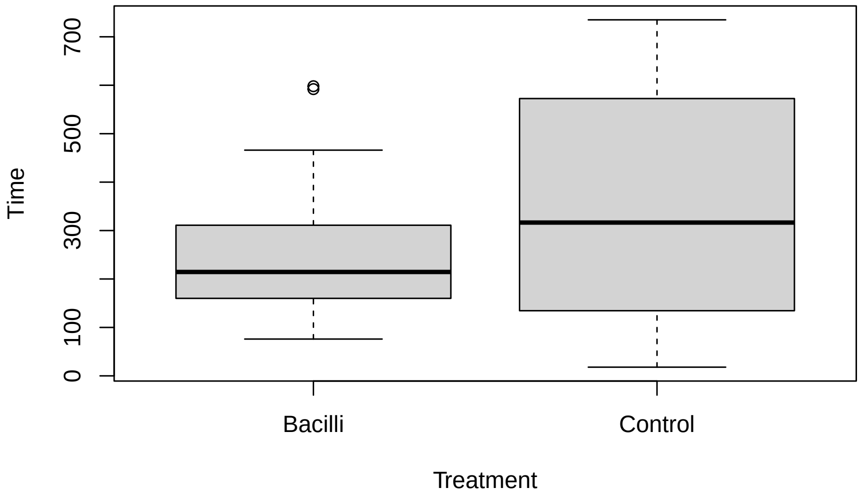
```
alpha <- 0.05
mu0 <- 1
sigma2 <- 2
n <- 10
power <- function(y)  1-pchisq((2/y)*qchisq(1-alpha, n-1),n-1)
curve(power, from = 2, to = 14, xlab = 'y', ylab = 'power(y)')
```



6.3.4 Example: The lifetimes of infected Guinea pigs

58 Guinea pigs were infected with a bacilli (bacteria) to determine its effect on lifespan. The control group of non-infected Guinea pigs lived on average 325 days. We'll test whether the infected pigs live, on average, less than 325 days. A comparative boxplot does not show any indications of non-normality in the lifetimes, so a t-test is reasonable. Based on the extreme test statistic (-5.33) and the tiny p-value (<0.0001) we conclude the infected pigs live, on average, less than 325 days. As this was a randomized experiment, it is reasonable to conclude the treatment/infection with bacilli is the cause of the reduced average lifespan. And, provided there is no reason to suspect these Guinea pigs are somehow different than Guinea pigs in general, we could assume the same results would hold, generally, in the (hypothetical) population of all Guinea pigs.

```
pigs <-
read.csv('E:\\Teaching\\STAT500\\MyMaterials\\Homework\\guinea_pigs.csv')
boxplot(Time~Treatment, data=pigs)
```



```
mu0 <- 325
m_C <- mean(pigs$Time[pigs$Treatment == "Control"])
m_C
```

```
## [1] 345.2188
```

```
m_B <- mean(pigs$Time[pigs$Treatment == "Bacilli"])
m_B
```

```
## [1] 242.5345
```

```
s_B <- sd(pigs$Time[pigs$Treatment == "Bacilli"])
s_B
```

```
## [1] 117.9309
```

```
n_B <- sum(pigs$Treatment == "Bacilli")
n_B
```

```
## [1] 58
```

```
c(m_B + qt(0.05, n_B - 1)*s_B/sqrt(n_B), m_B + qt(0.95, n_B -
1)*s_B/sqrt(n_B))
```

```
## [1] 216.643 268.426
```

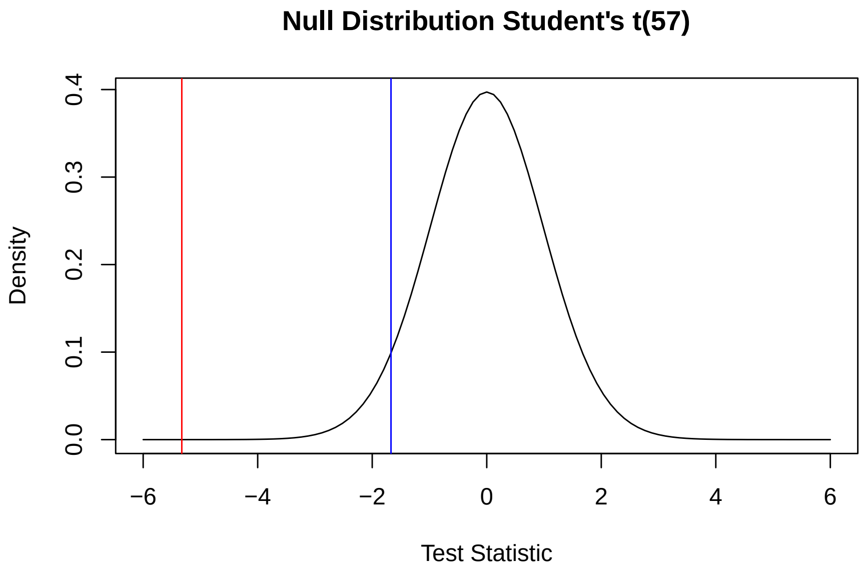
```
t.stat <- (m_B - mu0)/(s_B/sqrt(n_B))
t.stat
```

```
## [1] -5.325481
```

```
pt(t.stat, n_B-1)
```

```
## [1] 8.875078e-07
```

```
dt1 <- function(x) dt(x, n_B - 1)
curve(dt1, -6,6, xlab = "Test Statistic", ylab = "Density", main = "Null Distribution Student's t
abline(v = t.stat, col = 'red')
abline(v = qt(0.05, n_B-1), col = 'blue')
```



Next, we note the standard deviation of lifetime in the control group is about 222 days. One might expect the infected group to have a lover standard deviation of lifetime given many will die prematurely. Let's test $H_0 : \sigma \geq 222$ versus

$H_0 : \sigma < 222$ using a the pivotal quantity $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$—this null hypothesis test statistic has a Chi-squared distribution with $n-1$ degrees of freedom. Based on the very small p-value, we reject the null hypothesis. For reference, the figure below shows the test statistic (red) and $\alpha = 0.05$ cutoff (blue) in relation to the null distribution of the test statistic.

```
s_B <- sd(pigs$Time[pigs$Treatment == "Bacilli"])
s_B
```

```
## [1] 117.9309
```

```
n_B <- sum(pigs$Treatment == "Bacilli")
n_B
```

```
## [1] 58
```

```
c((n_B - 1)*(s_B^2)/qchisq(0.975, n_B - 1), (n_B - 1)*(s_B^2)/qchisq(0.025, n_B - 1))
```

```
## [1]  9940.021 20846.868
```

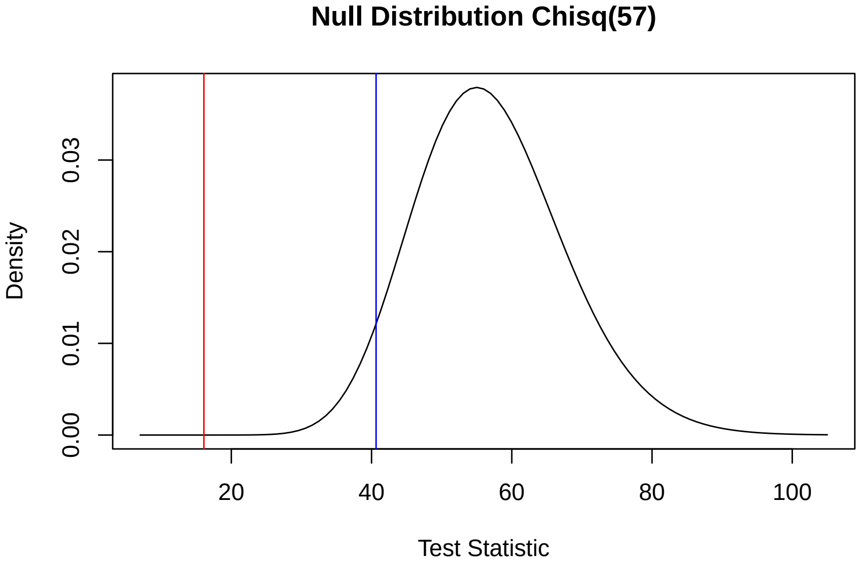
```
chi.stat <- (n_B - 1)*(s_B^2)/(222^2)
chi.stat
```

```
## [1] 16.08511
```

```
pchisq(chi.stat, n_B-1)
```

```
## [1] 1.713066e-08
```

```
dchisq1 <- function(x) dchisq(x, n_B - 1)
curve(dchisq1, 7,105, xlab = "Test Statistic", ylab = "Density", main = "Null Distribu
abline(v = chi.stat, col = 'red')
abline(v = qchisq(0.05, n_B-1), col = 'blue')
```



6.3.5 Tests for a difference of normal population means

Suppose $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ are random samples from two normal populations with unknown means and variances μ_x, μ_y, σ_x^2, and σ_y^2. Consider testing the point-null hypothesis $H_0 : \mu_x - \mu_y = \Delta$ versus $H_a : \mu_x - \mu_y \neq \Delta$ for a given constant Δ. Following our previous strategies, the test decision should be based on one of the two statistics

$$T = \frac{\overline{X} - \overline{Y} - \Delta}{\sqrt{S_x^2/n + S_y^2/m}} \quad \text{or} \quad T_p = \frac{\overline{X} - \overline{Y} - \Delta}{\sqrt{S_p^2\,(1/n + 1/m)}}.$$

Under the null hypothesis, T_p is exactly t distributed with $n + m - 2$ df if $\sigma_x^2 = \sigma_y^2$, and under the null hypothesis T is approximately t distributed with the Welch-Satterthwaite choice of df. These two statistics, T_p and T, are the basis for the pooled Student's t test and Welch's unpooled test for $\mu_x - \mu_y$. For either test, we reject H_0 if the test statistic is too large in absolute value, i.e., $|T_p| > t_{1-\alpha/2}(n+m-2)$. The Welch test is usually the better choice because, as we have previously seen, Student's pooled CIs perform may poorly when $n \neq m$ and $\sigma_x^2 \neq \sigma_y^2$, and the Student's t test experiences the same problems.

These two-sample t-tests may be performed using R's built-in t.test function:

```
set.seed(12345)
X <- rnorm(13, 5, 2)
```

```
Y <- rnorm(19, 4, 1)
# testing a point null with Delta = 1, alpha = 0.05
t.test(X,Y,alternative = 'two.sided',mu=1,var.equal=FALSE,conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  X and Y
## t = -0.23467, df = 17.877, p-value = 0.8171
## alternative hypothesis: true difference in means is not equal to 1
## 95 percent confidence interval:
##  -0.2897301  2.0306764
## sample estimates:
## mean of x mean of y
##  5.114192  4.243719
```

```
t.test(X,Y,alternative = 'two.sided',mu=1,var.equal=TRUE,conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  X and Y
## t = -0.25733, df = 30, p-value = 0.7987
## alternative hypothesis: true difference in means is not equal to 1
## 95 percent confidence interval:
##  -0.1574864  1.8984327
## sample estimates:
## mean of x mean of y
##  5.114192  4.243719
```

The t.test function returns a "p-value" which is defined as the null-hypothesis-true probability of observing the observed test statistic or a more extreme value. In other words, it is equal to the equivalent α for which the observed test statistic lies on the boundary of the reject/retain criterion. For example, Welch's test gives a p-value of 0.8171. The test statistic is

$$T = \frac{5.114192 - 4.243719 - 1}{\sqrt{3.18206/13 + 1.137897/19}} = -0.2346661.$$

And, the Welch-Satterthwaite df is 17.877 according to t.test. For a test statistic value of -0.2346661 to be borderline for rejection means the upper $t(17.877)$ quantile for rejection would equal 0.2346661. Using the *pt* function we have $(1 - pt(0.2346661, 17.877)) = 0.4085673$ which constitutes the probability on one half of the symmetric "rejection region", so that the p-value is twice this number, or 0.8171347, as given by t.test.

```
(1-pt(0.2346661,17.877))
```

```
## [1] 0.4085673
```

6.3.6 Test for equality of normal population variances

Let $X_1, \dots, X_n$ be iid $N(\mu_x, \sigma_x^2)$ and $Y_1, \dots, Y_m$ be iid $N(\mu_y, \sigma_y^2)$ and consider testing $H_0 : \sigma_x^2/\sigma_y^2 = 1$ versus $H_a : \sigma_x^2/\sigma_y^2 \neq 1$. As we have seen previously

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F(n-1, m-1)$$

and, under the null hypothesis, we have

$$\frac{S_x^2}{S_y^2} \overset{H_0}{\sim} F(n-1, m-1),$$

so that a level α test rejects the null hypothesis if $F = \frac{S_x^2}{S_y^2} > F_{1-\alpha/2}(n-1, m-1)$ or $F < F_{\alpha/2}(n-1, m-1)$. Example: Suppose we obtain the following two random samples of X and Y values:

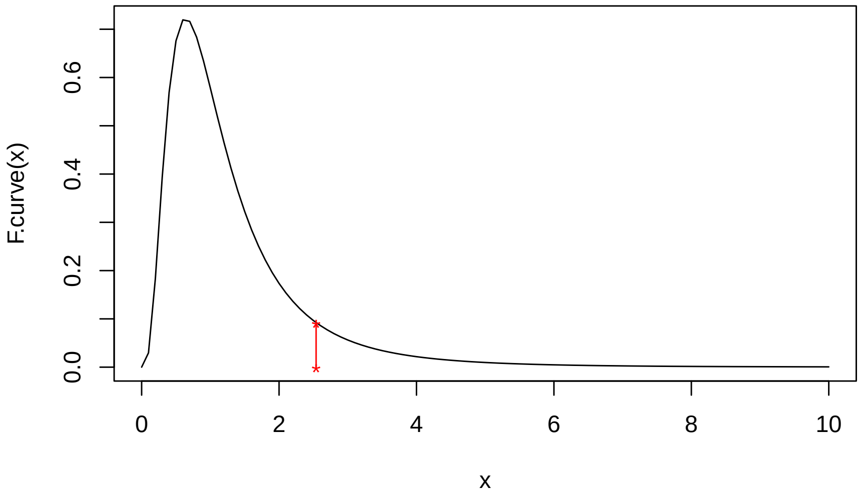
$$x : 25.20,\ \ 23.53,\ \ 18.02,\ \ 18.96,\ \ 15.70,\ \ 10.35,\ \ 24.07,\ \ 17.10,\ \ 21.51,\ \ 7.48,\ \ 12.76$$

$$y : 19.16,\ \ 20.05,\ \ 16.99,\ \ 19.83,\ \ 21.07,\ \ 21.47,\ \ 25.44,\ \ 11.96,\ \ 19.45$$

```
x <- c(25.20, 23.53, 18.02, 18.96, 15.70, 10.35, 24.07, 17.10, 21.51,  7.48, 12.76)
y <- c(19.16, 20.05, 16.99, 19.83, 21.07, 21.47, 25.44, 11.96, 19.45)
F <- var(x)/var(y)
F
```

```
## [1] 2.539221
```

```
F.curve <- function(x) df(x, 10, 8)
curve(F.curve, 0, 10)
points(F, F.curve(F), pch = '*', col = 'red')
points(F, 0, pch = '*', col = 'red')
lines(c(F,F), c(0,F.curve(F)), col = 'red')
```



```
1-pf(F, 10, 8)
```

```
## [1] 0.09989054
```

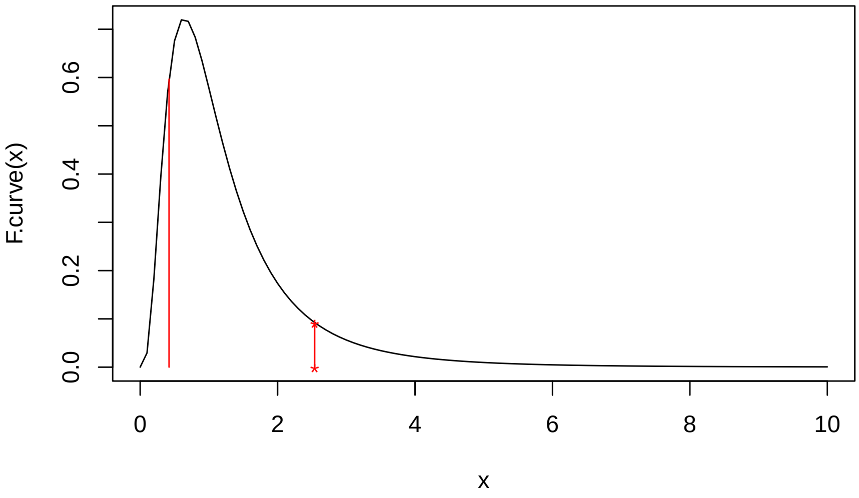
```
2*(1-pf(F, 10, 8))
```

```
## [1] 0.1997811
```

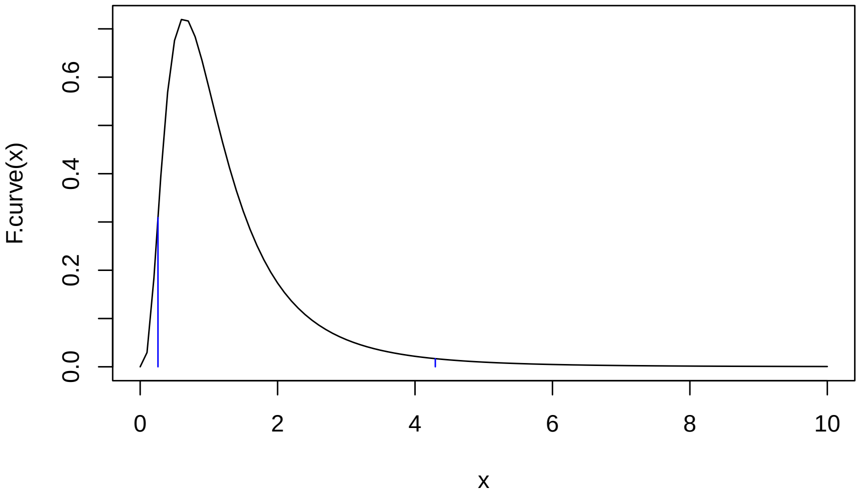
```
qf(0.09989054, 10,8)
```

```
## [1] 0.4204884
```

```
curve(F.curve, 0, 10)
points(F, F.curve(F), pch = '*', col = 'red')
points(F, 0, pch = '*', col = 'red')
lines(c(F,F), c(0,F.curve(F)), col = 'red')
lines(c(qf(0.09989054, 10,8),qf(0.09989054, 10,8)), c(0,F.curve(qf(0.09989054, 10,8)))
```



```
curve(F.curve, 0, 10)
lines(c(qf(0.025, 10,8),qf(0.025, 10,8)), c(0,F.curve(qf(0.025, 10,8))), col = 'blue')
lines(c(qf(0.975, 10,8),qf(0.975, 10,8)), c(0,F.curve(qf(0.975, 10,8))), col = 'blue')
```



6.4 Likelihood-based Tests

In our discussions of point and interval estimation we found likelihood-based methods provided a general means of deriving estimators with (at least) good large-sample performance. The same is true with respect to testing. However, there's more than one way to use the likelihood function to define a test, and we'll discuss two strategies.

6.4.1 Wald type tests

Wald-type tests follow the strategy we've been using. The test is based on a point estimator and its distribution. The testing rule is essentially to reject the null if the point estimator sufficiently disagrees with the null where "sufficiently" is determined by the estimator's distribution. In the case of likelihood-based Wald tests the point estimator is the MLE and the distribution is the (asymptotic/approximate) normal distribution of the MLE. Therefore, all Wald-type likelihood-based tests of $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ are of the form "Reject H_0 if:"

$$\theta_0 \notin (\hat\theta \pm z_{1-\alpha/2}[nI(\theta_0)]^{-1/2}),$$

because

$$\frac{\hat\theta - \theta_0}{[nI(\theta_0)]^{-1/2}} \overset{H_0}{\sim} N(0,1)$$

where θ is a scalar. Note: for point null tests $H_0 : \theta = \theta_0$ for $p \times 1$ vector parameters $(\hat\theta - \theta_0)[nI(\theta_0)]^{-1/2}$ is a p-variate standard normal and the quadratic form $(\hat\theta - \theta_0)^\top[nI(\theta_0)]^{-1}(\hat\theta - \theta_0)$ has a Chi-squared distribution with p degrees of freedom. And, the Wald test is based on quantiles of this Chi-squared distribution.

Example: Suppose $X_1, \ldots, X_n$ is a random sample from an Exponential distribution with scale parameter λ and consider testing $H_0 : \lambda \le \lambda_0$ versus $H_a : \lambda > \lambda_0$. The sample mean is the MLE of λ, and $n\overline{X} = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$. Therefore, an exact test is based on the Gamma distribution, while a likelihood-based Wald test is based on $\overline{X} \overset{H_0}{\sim} N(\lambda_0, \lambda_0^2/n)$, approximately. Suppose $\lambda_0 = 5$ and we observe the following random sample of size 12:

$$x : 12.57,\ 1.83,\ 2.10,\ 3.00,\ 0.65,\ 2.74,\ 14.21,\ 1.72,\ 3.13,\ 1.63,\ 2.42,\ 9.33$$

```
x = c(12.57,  1.83,  2.10,  3.00,  0.65,  2.74, 14.21,  1.72,  3.13,  1.63,  2.42,  9.3
xbar <- mean(x)
xbar
```

```
## [1] 4.610833
```

```
c(xbar - 1.96*5/sqrt(12), xbar + 1.96*5/sqrt(12))
```

```
## [1] 1.781817 7.439850
```

```
5
```

```
## [1] 5
```

```
c(qgamma(0.025, shape = 12, scale = 5), qgamma(0.975, shape = 12, scale = 5))
```

```
## [1] 31.00288 98.41019
```

```
xbar * 12
```

```
## [1] 55.33
```

```
c(qgamma(0.025, shape = 12, scale = 5), qgamma(0.975, shape = 12, scale = 5))/12
```

```
## [1] 2.583573 8.200849
```

```
xbar
```

```
## [1] 4.610833
```

Simulation of Type 1 error rates:

```
reps <- 10000
errors <- matrix(0, reps, 2)
for(r in 1:reps){
  x <- rexp(12,rate = 1/5)
  xbar <- mean(x)
  rr.wald <- c(xbar - 1.96*5/sqrt(12), xbar + 1.96*5/sqrt(12))
  rr.exact <- c(qgamma(0.025, shape = 12, scale = 5), qgamma(0.975, shape = 12, scale = 5))/12
  errors[r, ] <- c(ifelse(5>rr.wald[1] & 5<rr.wald[2],0,1), ifelse(xbar>rr.exact[1] & xbar<rr.exa
}
colMeans(errors)
```

```
## [1] 0.0487 0.0536
```

6.4.2 Likelihood ratio tests

Likelihood ratio tests (LRTs) operate a little differently compared to Wald tests. The LRT idea is that parameter values with corresponding likelihood values near the maximum are plausible, while those with low likelihood values are implausible. For testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ this means if $\frac{L(\theta_0)}{L(\hat{\theta})}$ is close to 1 the null is supported while if the ratio is close to zero the alternative is supported. How "close" is close again depends on the null-hypothesis distribution of the likelihood ratio. A general asymptotic result holds that for scalar parameters

$$-2[\ell(\theta_0) - \ell(\hat{\theta})] \overset{H_0}{\sim} Chi - squared(1)$$

approximately. This result also holds for general composite hypotheses $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta \cap \Theta_0^c$ so that

$$-2[\sup_{\theta \in \Theta_0} \{\ell(\theta)\} - \ell(\hat{\theta})] \overset{H_0}{\sim} Chi - squared(df)$$

where df is equal to the difference in dimensions of the parameter space.

Example: Consider a test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ for a normal population with unknown variance. In that case Θ is the two-dimensional parameter space for (μ, σ^2) which is $\mathbb{R} \times \mathbb{R}^+$ and Θ_0 is the one-dimensional subspace $\{\mu_0\} \times \mathbb{R}^+$. The unrestricted MLEs are $\overline{x}$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i - \overline{x})^2$. It's not hard to show the restriced MLE for σ^2 under the null hypothesis is $\hat{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (x_i - \mu_0)^2$. To test H_0 we compute the test statistic

$$\Lambda := -2[\ell(\mu_0, \hat{\sigma}_0^2) - \ell(\overline{x}, \hat{\sigma}^2)]$$

and reject the null hypothesis if $\Lambda > \chi^2_{1-\alpha}(1)$, the $1-\alpha$ lower quantile of the Chi squared distribution with 1 degree of freedom. Let's perform the test given the following sample:

```
x <- rnorm(22, 8, 6)
x
```

```
##  [1]  4.0634008  9.1198213 13.6167056  7.1889954 13.9577182 10.2795735
##  [7]  7.8406222  6.1487191  0.2090839  0.7775581 10.5912890 10.3346847
## [13] 19.6868480  0.8674458 12.9657995  8.1965838  2.9815080  7.5730705
## [19]  8.2187202  2.9247010  4.4469792  3.5998290
```

```
xbar <- mean(x)
xbar
```

```
## [1] 7.526803
```

```
mu0 <- 7
mle <- (1/22)*sum((x-xbar)^2)
mle
```

```
## [1] 23.11427
```

```
mle0 <- (1/22)*sum((x - mu0)^2)
mle0
```

```
## [1] 23.39179
```

```
loglhood <- function(mu, sigma2)  sum(dnorm(x,mu,sqrt(sigma2),log = TRUE))
Lambda <- (-2)*(loglhood(mu0,mle0) - loglhood(xbar,mle))
Lambda
```

```
## [1] 0.2625694
```

```
qchisq(0.95, 1)
```

```
## [1] 3.841459
```

Example: Consider the following data on cell phone battery charge times

$$8.0,\ \ 8.1,\ \ 8.9,\ \ 9.1,\ \ 9.2,\ \ 9.3,\ \ 9.4,\ \ 9.4$$

Assuming these times constitute a random sample from a normal population test the hypothesis $H_0 : \sigma^2 \le 0.2$ versus $H_a : \sigma^2 > 0.2$ using an LRT at $\alpha = 0.05$. The loglikelihood, again, is

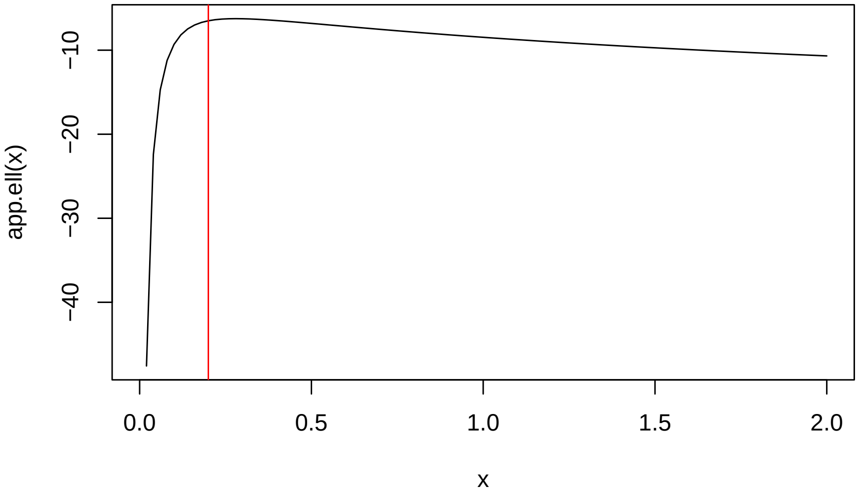
$$\ell(\mu, \sigma^2; x) = -(n/2)\log[2\pi\sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The unrestricted MLEs are $\overline{x} = 8.925$ and $\hat{\sigma}^2 = 0.28$. Under the null hypothesis the MLE of μ remains $\overline{x} = 8.925$ because the derivative of the loglikelihood w.r.t. μ is proportional to σ^2; i.e., it equals $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$. Then, setting equal to zero we see the MLE for μ is still $\overline{x}$. To find the restricted MLE of σ^2 under H_0, plot the loglikelihood as a function of σ^2 with $\mu = \overline{x}$ below and note that it is maximized at 0.2 over the set $\sigma^2 \le 0.2$. The test statistic is

$$-2(\ell(8.925, 0.2; x) - \ell(8.925, 0.28; x)) = 0.501$$

which is below the cutoff of 3.84, the 95% quantile of the Chi-squared distribution with 1 degree of freedom (the difference in two versus one free parameter in the alternative versus null hypothesis).

```
x <- c(8.0,8.1,8.9,9.1,9.2,9.3,9.4,9.4)
ell <- function(sig2) sum(dnorm(x,mean(x),sqrt(sig2),log = TRUE))
app.ell <- function(sig2) apply(matrix(sig2,length(sig2),1),1,ell)
curve(app.ell,0,2)
abline(v = 0.2, col = 'red')
```



```
test.stat <- -2*(ell(0.2)-ell(0.28))
test.stat
```

```
## [1] 0.5010792
```

```
qchisq(0.95,1)
```

```
## [1] 3.841459
```

6.5 Chi Squared tests for tabulated data

To this point most of our experiments have dealt with continuous data. But, a common type of data encountered in experiments is tabulated or cross-classified data. For example, consider the following data on phsychologists' opinions concerning the causes of schizophrenia:

	Origin			
School	Biogenic	Environment	Combination	
Eclectic	90	12	78	180
Medical	13	1	6	20
Psychoanalytic	19	13	50	82
	122	26	134	282

A test of interest is whether the psychologists' schools of thought are related to their responses. Statistically, this can be interpreted as a test of independence between the variables "school" and "cause". A test of this hypothesis of independence may be based on a multinomial sampling distribution where the 282 respondents are assumed to be a random sample from a multinomial distribution with 9 cross-classified categories based on "school" and "cause". This distribution has 8 free parameters for the category memberships. Under the null hypothesis, the joint probabilities are products of the marginal "school" and "cause" probabilities, so that there are only 4 free parameters.

A Chi-squared test for independence is based on the following test statistic:

$$\chi^2 = \sum_{i=1}^{I} (O_i - E_i)^2 / E_i$$

where O_i is the observed category count and E_i is the expected category count under H_0. Under the hypothesis of independence we expect the category counts to be $n = 282$ times the products of marginal category probabilities. For example, we expect to find $282 * (180/282) * (122/282) = 77.87$ counts in the Biogenic-Eclectic category; we observe 90. Similarly, we expect $26 * 20/282 = 1.84$ counts in the Medical-Environment category; we observe 1. The first summand in the test statistic is $(77.87 - 90)^2/77.87 = 1.88952$. Following along this way, we compute the test statistic to be $\chi^2 = 22.37769$. The chi^2 test statistic is approximately Chi-squared distributed with df equal to the difference in the number of free parameters in H_0 and H_a. In this case, that is $8 - 4 = 4$. Since the 95% quantile of the Chi-squared distribution with 4 df is 9.487729, we reject the null hypothesis of independence. We could also perform an LRT of the independence hypothesis based on the multinomial distribution of the cell counts. The MLEs of the cell proportions are simply the observed cell proportions. The multinomial loglikelihood is given by

$$\ell(p; x_1, \dots, x_n) = \log \frac{n!}{x_1! \times \dots \times x_9!} + \sum_{j=1}^{9} x_j \log p_j.$$

In R, we can simply use "dmultinom" with option log = TRUE to compute the loglikleihood.

```
test.stat <- -2*(dmultinom(c(90,12,78,13,1,6,19,13,50),prob = c(180*122, 180*26, 180*13
test.stat
```

```
## [1] 23.03619
```

```
qchisq(0.95,4)
```

```
## [1] 9.487729
```

For completeness, let's show the MLEs of multinomial cell probabilities are the corresponding sample proportions. The loglikelihood is equal to

$$\ell(p;x) = \log\left(\frac{n!}{x_1! \times x_2! \times \cdots \times x_9!} + x_1\log(p_1) \times \cdots \times x_8\log(p_8) + (n - x_1 - x_2 - \cdots - x_8)\log(1 - p_1 - \cdots$$

Take the derivative of the loglikelihood w.r.t. p_1, set equal to zero, and find

$$\frac{1 - p_1 - p_2 - \cdots - p_8}{p_1} - \frac{n - x_1 - x_2 - \cdots - x_8}{x_1}$$

Divide the right-hand-side by n in both the numerator and denominator to find

$$\frac{1 - p_1 - p_2 - \cdots - p_8}{p_1} = \frac{1 - \frac{x_1}{n} - \frac{x_2}{n} - \cdots - \frac{x_8}{n}}{\frac{x_1}{n}}.$$

That this holds simultaneously for derivatives with respect to p_2 through p_8 implies the MLEs of the cell probabilitites are exactly the sample cell proportions. A similar argument can be made under H_0 to show the MLEs under H_0 are the sample marginal cell proportions.

6.5.1 Goodness of fit tests

The same Chi-squared test used above comparing observed awith expected cell counts is used to test the "goodness of fit" of a model to a data set. This test is most often used with discrete data, but can be used with continuous data by "binning" or discretizing continuous data in the same fashion as a histogram. For tabulated data the goodness of fit test statistic has df $k - p - 1$ where k is the number of cells in the table and p is the number of estimated parameters in the null hypothesis.

Example: Suppose an experiment records the number of customers entering a salon each hour during a weekday for the purpose of determining staffing requirements. The observed data are as follows:

8am	9am	10am	11am	12pm	1pm	2pm	3pm	4pm	5pm	6pm
3	6	14	11	6	4	4	3	8	11	3

We decide to perform a goodness of fit test to determine whether the data fit a Poisson distribution. It the data appear to be a Poisson random sample, then that implies there is not a strong temporal pattern in the data. On the other hand if the data do not fit a Poisson distribution the lack of fit may be caused by a temporal pattern (so that the data are not iid). Based on the computation below, we see we reject the null hypothesis that the data are a Poisson random sample at α level 5%.

Below we bin the data into categories of 0-3 counts, 4-7 counts, 8-11, and 12+. A good rule of thumb is to choose bins such that the expected counts are not less than 5.

```
x <- c(3,6,14,11,6,4,4,3,8,11,3)
n <- 11
bins <- c(3,4,3,1)
eis <- n*c(sum(dpois(0:3, mean(x))),sum(dpois(4:7, mean(x))),sum(dpois(8:11, mean(x))),   1-ppois
sum(((bins - eis)^2)/eis)
```

```
## [1] 4.610001
```

```
qchisq(0.95,4-1-1)
```

```
## [1] 5.991465
```

```
1-pchisq(sum(((bins - eis)^2)/eis),4-1-1)
```

```
## [1] 0.09975877
```

Chapter 7

ANOVA

One of the most important experiments we've discussed in the course involves the comparison of two populations in terms of their means. This kind of comparison motivated our study of the two-sample t-test and CI for the difference of two population means. These are very useful statistical methods, but why stop at comparing two populations? Often, we would like to compare three or more populations. For example, suppose a company marketing laundry detergent wants to compare 4 product designs (the detergent container and its labeling) in a pilot study. Using t-tests or CIs we could make pairwise comparisons (there are 4 choose 2, or, 6 pf these) but this causes some issues. Specifically, suppose there are no difference in the product designs and suppose we conduct 6 pairwise t tests at $\alpha = 0.05$. If the 6 tests are independent (the test statistics are mutually independen r.v.'s) then the chance we make at least 1 Type 1 error is $1 - 0.95^6 = 0.265$! Even if there are no deifferences, there's a good chance we'll find at least one difference by chance alone. In order to avoid the Type 1 error inflation caused by *multiple testing* one strategy—called Bonferroni correction—requires that we conduct each test at level $\alpha/6 = 0.00833$ so that the *family-wise* Type 1 error rate is no more than $1 - (1 - 0.00833)^6 = 0.041 \leq 0.05$. This fixes our Type 1 error inflation problem, but it also means our power to detect fale null hypotheses (detect differences in response for different product designs) is severely hampered. Can we do better? Fortunately, the answer is yes! The alternative, improved strategy for comparing multiple population means is called ANOVA (analysis of variance) and its the focus of this section.

7.1 Analysis of variance

The general setup is as follows: we conduct an experiment in which we observe a random sample of size n from each of J populations so that the total sample size is $N = nJ$. Denote the samples by X_{ij} for individual $i = 1, \ldots, n$ in the sample from population $j = 1, \ldots, J$. (It's possible, of course, to consider random samples of different sizes, but for simplicity we consider only the *balanced* case). Our goal in this experiemnt is to evaluate the null hypothesis $H_0 : \mu_j = \mu$ for all $j = 1, \ldots, J$, meaning that all the populations have the same mean. Further, we assume populations are normal with equal variance; this means that the populations are, essentially, equivalent under H_0. The alternative hypothesis is that at least one population mean differs from the others.

Rather than comparing sample means to evaluate H_0 we (somewhat counter-intuitively) compare two estimators of the population variance. Consider the following decomposition:

$$\sum_{i=1}^{n}\sum_{j=1}^{J}(X_{ij} - \overline{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{J}(X_{ij} - \overline{X}_j)^2 + n\sum_{j=1}^{J}(\overline{X}_j - \overline{X})^2.$$

The left hand side of the equals sign in the above display is called the total sum of squares TSS and it decomposes into the sum of the error sum of squares ESS (or residual sum of squares) and the treatment sum of squares TrSS. Under the null hypothesis all three of these sums of squares can be used to estimate σ^2, the common variance among the J populations. Further, the error sum of squares and treatment sum of squares are independent and satisfy

$$\frac{TrSS}{ESS} \overset{H_0}{\sim} F(J-1, J(n-1)).$$

The ANOVA F-test rejects the null hypothesis of equality of group means if $F = TrSS/ESS$ is more extreme than the upper $1-\alpha$ quantile of the F distribution with df $J-1$ and $J(n-1)$.

7.2 Crop data in-depth example

The embedded data comes from an experiment examining the effects of three types of fertilizer on crop yield. According to the data below there are 96 observations of yield, 32 each from each fertilizer type. And, the mean yields for the three fertilizer types are all very close to 177.

```
crops <- read.csv("crop.data.csv", header = TRUE, colClasses = c("factor", "factor", "
m <- mean(crops$yield)
m1 <- mean(crops$yield[crops$fertilizer == 1])
m2 <- mean(crops$yield[crops$fertilizer == 2])
m3 <- mean(crops$yield[crops$fertilizer == 3])

m
```

```
## [1] 177.0155
```

```
m1
```

```
## [1] 176.757
```

```
m2
```

```
## [1] 176.9332
```

```
m3
```

```
## [1] 177.3562
```

```
length(crops$yield)
```

```
## [1] 96
```

```
sum(crops$fertilizer == 1)
```

```
## [1] 32
```

```
sum(crops$fertilizer == 2)
```

```
## [1] 32
```

```
sum(crops$fertilizer == 3)
```

```
## [1] 32
```

We can compute the ANOVA F-test "by hand" in R by computing the component sums of squares and comparing to the appropriate lower and upper F quantiles using the qf function or by computing the p-value.

```
# SSE

SSE <- sum((crops$yield[crops$fertilizer == 1] - m1)^2)+sum((crops$yield[crops$fertilizer == 2] -
SSE
```

```
## [1] 35.88619
```

```
# SSTr
n <- sum(crops$fertilizer == 1)
n
```

```
## [1] 32
```

```
SSTr <- n*((m1-m)^2)+n*((m2-m)^2)+n*((m3-m)^2)
SSTr
```

```
## [1] 6.068047
```

```
# F test
J <- 3
F <- (SSTr / (J-1)) / (SSE / (J*(n-1)))
F
```

```
## [1] 7.862752
```

```
MSE <- (SSE / (J*(n-1)))

p.value <- 1-pf(abs(F), J-1, J*(n-1))
p.value
```

```
## [1] 0.0006999158
```

We can also use built-in R functions to compute the ANOVA F-test. Note the matching p-values.

```
summary(aov(yield ~ fertilizer, data = crops))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fertilizer   2   6.07  3.0340   7.863  7e-04 ***
## Residuals   93  35.89  0.3859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

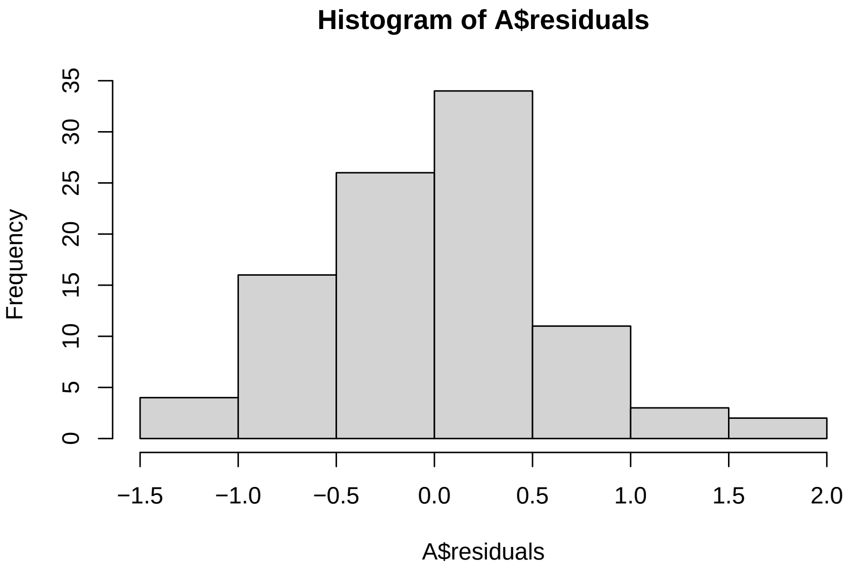
```
A <- aov(yield ~ fertilizer, data = crops)
A
```

```
## Call:
##    aov(formula = yield ~ fertilizer, data = crops)
```

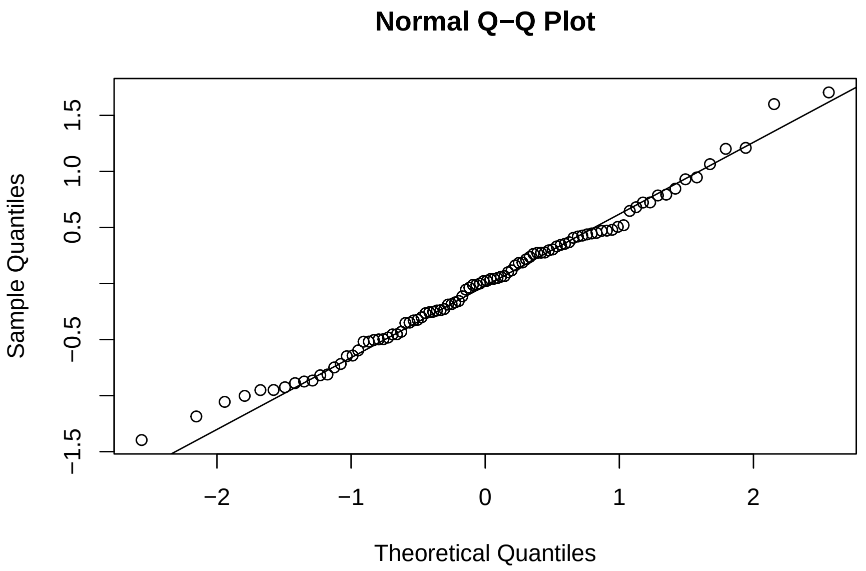
```
##
## Terms:
##                 fertilizer Residuals
## Sum of Squares     6.06805  35.88619
## Deg. of Freedom          2        93
##
## Residual standard error: 0.6211867
## Estimated effects may be unbalanced
```

An important part of applying ANOVA is checking the normalitay and equal variance assumptions are reasonable. If they are not, then the F-test may be misleading because the test statistic may not have an F distribution under the null. The best way to check for normality is to examine the *residuals*—the values $x_{ij} - \overline{x}_j$; under the null and assuming equal variance these N values are approximately normally distributed with variance σ^2. We can also examine the Studentized resisduals by normalizing the residuals by the sample standard deviation—these should look approximately standard normal for a large sample (otherwise approximately Student's t with n-1 df).

```
hist(A$residuals)
```



```
qqnorm(A$residuals)
qqline(A$residuals)
```



7.2.1 Checking the equal variance assumption

We can also check the group-wise variances for equality. A formal method for checking these is Levene's test, which amounts to performing ANOVA on the absolute values of the residuals themselves. The null hypothesis of Levene's ANOVA is equality of variances over groups/populations.

```
sd(crops$yield[crops$fertilizer==1])
```

```
## [1] 0.6849233
```

```
sd(crops$yield[crops$fertilizer==2])
```

```
## [1] 0.5740668
```

```
sd(crops$yield[crops$fertilizer==3])
```

```
## [1] 0.5991214
```

```
crops$resid <- c(abs(crops$yield[crops$fertilizer==1]-mean(crops$yield[crops$fertilizer==1])),
abs(crops$yield[crops$fertilizer==2]-mean(crops$yield[crops$fertilizer==2])),
abs(crops$yield[crops$fertilizer==3]-mean(crops$yield[crops$fertilizer==3])))

summary(aov(resid ~ fertilizer, data = crops))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fertilizer   2   0.24  0.1197    0.88  0.418
## Residuals   93  12.65  0.1360
```

```
ml <- mean(crops$resid)
m1l <- mean(crops$resid[crops$fertilizer == 1])
m2l <- mean(crops$resid[crops$fertilizer == 2])
m3l <- mean(crops$resid[crops$fertilizer == 3])

ml
```

```
## [1] 0.4894152
```

```
m1l
```

```
## [1] 0.5596244
```

```
m2l
```

```
## [1] 0.4475853
```

```
m3l
```

```
## [1] 0.4610358
```

```
# SSE

SSE.L <- sum((crops$resid[crops$fertilizer == 1] - m1l)^2)+sum((crops$resid[crops$fertilizer == 2
SSE.L
```

```
## [1] 12.65207
```

```
# SSTr
n <- sum(crops$fertilizer == 1)
n
```

```
## [1] 32
```

```
SSTr.L <- n*((m1l-ml)^2)+n*((m2l-ml)^2)+n*((m3l-ml)^2)
SSTr.L
```

```
## [1] 0.2395029
```

```
# Levene F test
J <- 3
F.L <- (SSTr.L / (J-1)) / (SSE.L / (J*(n-1)))
F.L
```

```
## [1] 0.880242
```

```
qf(.95, J-1, J*(n-1))
```

```
## [1] 3.094337
```

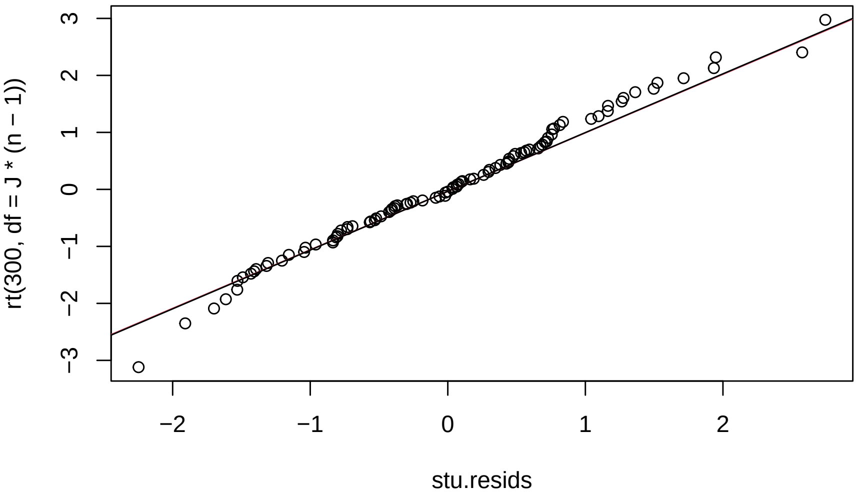
7.2.2 Normality and alternatives to the F test

As we have done in the past, we can investigate the assumption of normality by constructing plots of residuals. The observed, Studentized residuals are given by

$$e_{ij} = \frac{y_{ij} - \bar{y}_{i\cdot}}{\sqrt{MSW}} \dot{\sim} t_{N-I}$$

The qq-plot reveals no obvious signs of non-normality of residuals.

```
resids <- A$residuals
stu.resids <- resids/sqrt(MSE)
qqplot(stu.resids, rt(300, df = J*(n-1)))
qqline(stu.resids, distribution = function(p) qt(p, df = J*(n-1)),probs = c(0.25, 0.75)
qqline(stu.resids)
```



In case the qq-plot suggests normality does NOT hold, then an alternative approximate test may be used. The Kruskal-Wallis test is similar (but not exactly the same) as performing ANOVA on the ranks r_{ij} of the responses y_{ij}. Ranking means sorting the responses from least to greatest and associating each with its order in that sequence. If there are any tied values, ranks are averaged. For example, the sequence of responses $24.3, 27.9, 27.9, 29.3, 30, 1$ has ranks $1, 2.5, 2.5, 4, 5$.

Let r_{ij} be the ranks of the y_{ij} and let $\bar{r}_{i\cdot}$ and $\bar{r}_{\cdot\cdot}$ be the averages of ranks in each population sample and overall, respectively. Then, the Kruskal-Wallis test statistic is the ratio of the SSB (or called SSTr) to SST of the ranks, multiplied by $N-1$:

$$\chi^2 = (N-1) \cdot \frac{\sum_{i=1}^{I} n_i(\bar{r}_{i\cdot} - \bar{r}_{\cdot\cdot})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(r_{ij} - \bar{r}_{\cdot\cdot})^2}.$$

The null hypothesis of the Kruskall-Wallis test is that the I populations are really all the same population (the I distributions are equal). In the special case in which we assume all populations are normal with the same mean, this null hypothesis is equivalent to the assertion all the means are equal. Under the Kruskal-Wallis null hypothesis, the test statistic χ^2 is *approximately* distributed as a Chi-squared r.v. with $I-1$ degrees of freedom.

Next, we will compute the K-W test statistic and associated p-value by hand, and then compare our results to the built-in R function “kruskal.test”.

```
rij <- rank(crops$yield)

r1 <- mean(rij[crops$fertilizer == 1])
r2 <- mean(rij[crops$fertilizer == 2])
r3 <- mean(rij[crops$fertilizer == 3])

# SSE

SSE <- sum((rij[crops$fertilizer == 1] - r1)^2)+sum((rij[crops$fertilizer == 2] - r2)^2
SSE
```

```
## [1] 63603.81
```

```
SST <- sum((rij - mean(rij))^2)
SST
```

```
## [1] 73720
```

```
# SSTr
SSTr <- SST-SSE
SSTr
```

```
## [1] 10116.19
```

```
test.stat <- (n*J - 1)*(SSTr / SST)
test.stat
```

```
## [1] 13.03632
```

```
1-pchisq(test.stat, J-1)
```

```
## [1] 0.00147638
```

```
kruskal.test(yield ~ fertilizer, data = crops)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  yield by fertilizer
## Kruskal-Wallis chi-squared = 13.036, df = 2, p-value = 0.001476
```

Note that the Kruskal-Wallis test rejects the null hypothesis that the three populations of crop yields are the same, which agrees with the ANOVA F test.

Chapter 8

Simple Linear Regression

In this chapter we consider estimation and inference of a random response, conditional on another observed variable—the independent variable, also called predictor or covariate.

8.1 Simple linear regression model

The simple linear regression model says that a random response Y_i has conditional mean $\beta_0 + \beta_1 x_i$ given $X_i = x_i$ where X_i may be a random variable. Moreover, the responses Y_i, $i = 1, \dots, n$, are independent, with common variance σ^2, and are normally distributed. This is commonly written using the following statistical notation:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

where ϵ_i is the "random residual"—what is left over after subtracting the mean from the response.

The model is "linear" because the mean $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$ is a linear function (a line in two dimensions y and x). The model is "simple" because it is the simplest line (two dimensions). Later we will expand the model by adding covariates, i.e., $x_{1i}, x_{2i}, ...$, to the linear conditional mean function and call it *multiple linear regression.*

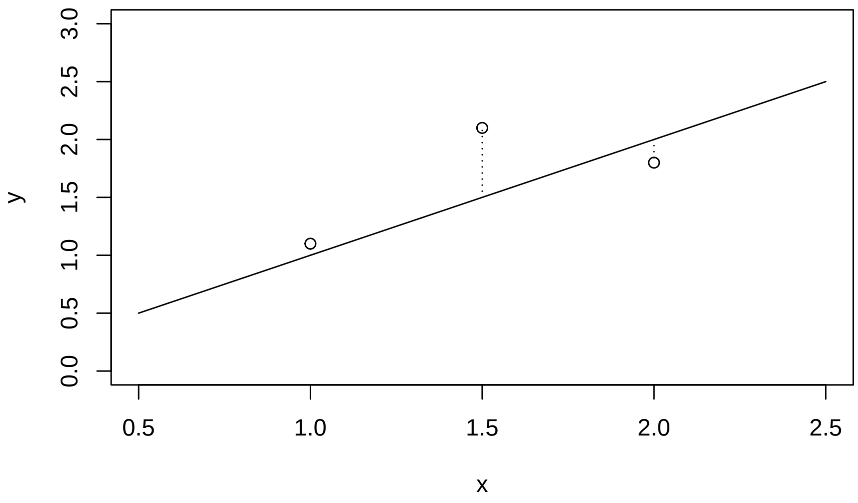
8.1.1 Estimation

We estimate (β_0, β_1) simultaneously using the method of "least squares". The method of least squares defines the line $y = \beta_0 + \beta_1 x$ that is "closest" to the points $(y_i, x_i), i = 1, \ldots, n$ is the one minimizing the sum of square vertical distances (residuals) from the points to the line:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

The plot below shows these residuals for three points (1.0,1.1), (1.5,2.1), and (2.0,1.8), compared to the line $y = x$.

```
plot(c(1,1.5,2),c(1.1,2.1,1.8), xlab = 'x', ylab = 'y', main = '', xlim = c(0.5,2.5), y
lines(c(0.5,2.5),c(0.5,2.5))
lines(c(1,1), c(1,1.1), lty = 3)
lines(c(1.5,1.5), c(1.5,2.1), lty = 3)
lines(c(2,2), c(2,1.8), lty = 3)
```



We can determine the estimators $(\hat{\beta}_0, \hat{\beta}_1)$ by minimizing the sum of squared residuals using calculus:

$$\begin{aligned}
\frac{\partial}{\partial \beta_0}\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2 & \\
&= -2\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)\\
\text{set } 0 &= -2\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)\\
&\Rightarrow \hat{\beta}_0 = n^{-1}\sum_{i=1}^n (y_i-\beta_1x_i)\\
&= \bar{y}-\beta_1\bar{x}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_1}\sum_{i=1}^n (y_i-\beta_0-\beta_1x_i)^2 & \\
&= -2\sum_{i=1}^n x_i(y_i-\beta_0-\beta_1x_i)\\
\text{set } 0 &= -2\sum_{i=1}^n x_i(y_i-\beta_0-\beta_1x_i)\\
\Rightarrow 0 &= \sum_{i=1}^n (y_ix_i - x_i(\bar{y}-\beta_1\bar{x})-\beta_1x_i^2) \quad \text{by substituting } \beta_0\\
\beta_1\sum(x_i^2-x_i\bar{x}) &= \sum_{i=1}^n (y_ix_i-\bar{y}x_i)\\
\Rightarrow \hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_ix_i)-n\bar{y}\bar{x}}{\sum_{i=1}^n (x_i^2)-n\bar{x}^2} = \frac{\sum_{i=1}^n [(y_i-\bar{y})(x_i-\bar{x})]}{\sum_{i=1}^n (x_i-\bar{x})^2}
\end{aligned}$$

### LSEs are unbiased

A nice property of the least squares method is that it produces unbiased estimators. Consider the expectation $E(\hat{\beta}_1)$. Since the x_i's are non-random, treat these as constants, and find $\hat{\beta}_1$ is unbiased:

$$\begin{aligned}
E(\hat{\beta}_1) &= \frac{\sum_{i=1}^n E((y_i-\bar{y}))(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}\\
&= \frac{\sum_{i=1}^n (\beta_0+\beta_1x_i - n^{-1}\sum_{i=1}^n [\beta_0+\beta_1x_i])(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}\\
&= \frac{\beta_1\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}\\
&= \beta_1.
\end{aligned}$$

Similarly, the estimator of the intercept is unbiased:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= n^{-1} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\ &= \beta_0 \end{aligned}$$

8.1.2 Estimation of the common variance

Estimating σ^2 in the regression model is similar to the method used in ANOVA. Define the observed/fitted residuals $\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Since $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased, we have $E(\hat{e}_i) = E(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$. Therefore, $V(\hat{e}_i) = E(\hat{e}_i^2)$. And, the method of moments suggests we estimate the variance(in this case second moment) by the sample variance:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

where we divide by $n-2$ so that the resulting estimator is unbiased. (I'll leave that as a challenging exercise for the reader).

8.1.3 Inference for regression slope parameter

There are two main inference (and prediction) problems of interest in regression. The first is inference on β_1, and, in particular, testing $\beta_1 = 0$. If $\beta_1 = 0$ then there is no linear relationship between the covariate x and the response Y, and Y has a constant mean (so is iid, rather than independent). Testing $\beta_1 = 0$ is similar to the ANOVA F test for categorical x, rather than continuous x in regression.

For inference on β_1 we need the sampling distribution of $\hat{\beta}_1$. Recognize that we can write this estimator as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

because $\sum_{i=1}^{n} \bar{Y}(x_i - \bar{x}) = 0$. Then, we see that $\hat{\beta}_1$ is a linear combination of Y_i, i.e., $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$ for non-random c_i. By a MGF argument we have used before, linear combinations of normal random variables are also normally-distributed. This means,

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma_2 \sum_{i=1}^{n} c_i^2\right).$$

Furthermore, by essentially the same argument as in the proof of Student's Theorem,

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

so that the studentized slope estimator has a Student's t distribution:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}}} \sim t_{n-2}$$

A test of $H_0 : \beta_1 = b$ versus $H_a : \beta_1 \neq b$ rejects the null if

$$\frac{|\hat{\beta}_1 - b|}{\sqrt{\hat{\sigma}^2 \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}}} > t_{1-\alpha/2,n-2}$$

Similarly, a $100(1-\alpha)\%$ CI for β_1 is given by

$$\left(\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}}\right).$$

8.1.4 Inference on the conditional mean response

Besides testing for $H_0 : \beta_1 = 0$, which means that the covariate has no effect on the mean response, experimenters often want to make inferences about the conditional mean response at a particular covariate value x. As we showed above, $\hat{\beta}_0 + \hat{\beta}_1 x$ is unbiased for $E(Y|x) = \beta_0 + \beta_1 x$. Furthermore, the variance of the estimator of the conditional mean is as follows:

$$\begin{aligned}
V(\hat{\beta}_0 + \hat{\beta}_1 x) &= V(\overline{Y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x) \\
&= V\left(\overline{Y} + (x - \overline{x})\frac{\sum Y_i (x_i - \overline{x})}{\sum (x_i - \overline{x})^2}\right) \\
&= V\left(\sum Y_i \left(n^{-1} + (x - \overline{x})\frac{(x_i - \overline{x})}{\sum (x_i - \overline{x})^2}\right)\right) \\
&= \sum_{i=1}^{n} V\left(Y_i \left(n^{-1} + (x - \overline{x})\frac{(x_i - \overline{x})}{\sum (x_i - \overline{x})^2}\right)\right) \\
&= \sigma^2 \sum_{i=1}^{n} \frac{1}{n^2} + \frac{(x - \overline{x})^2 (x_i - \overline{x})^2}{\left[\sum (x_i - \overline{x})^2\right]^2} \\
&= \sigma^2 \left(n^{-1} + \frac{(x - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)
\end{aligned}$$

Replacing the unknown variance σ^2 by its estimator (the MSE) we obtain a Student's t random variable with $n-2$ degrees of freedom:

$$t = \frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x)}{\sqrt{MSE\left(n^{-1} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}\right)}} \sim t_{n-2}.$$

This may be used to define a confidence interval for the conditional mean response at a given covariate value x:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{1-\alpha/2, n-2}\sqrt{MSE\left(n^{-1} + \frac{(x - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)}\right)$$

8.1.5 Prediction of responses at given x values

Let $Y^\star$ denote a future response for a given covariate value x. We want to predict $Y^\star \sim N(\beta_0 + \beta_1 x, \sigma^2)$. Now, we can represent $Y^\star$ using the regression model as

$$Y^\star = \beta_0 + \beta_1 x + \epsilon^\star$$

for a future random residual $\epsilon^\star$. An unbiased point prediction of $Y^\star$ is given by the estimate of its mean $\hat{\beta}_0 + \hat{\beta}_1 x$. The variance of the point prediction is

$$V(\widehat{Y^\star}) = V(\hat{\beta}_0 + \hat{\beta}_1 x + \epsilon^\star) = \sigma^2 \left(1 + n^{-1} + \frac{(x - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right).$$

Therefore, a $100(1-\alpha)\%$ prediction interval for $Y^\star$ is given by

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{1-\alpha/2, n-2}\sqrt{MSE\left(1 + n^{-1} + \frac{(x - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)}\right).$$

Prediction intervals have interpretations similar to confidence intervals. If the model is true, a 95% prediction interval for $Y^\star$ at a given x has the following property: suppose we repeat a regression experiment many, many times. Each time we compute a 95% interval and then observe a new response $Y^\star$. then, about 95% of the time, our computed prediction interval contains that realized $Y^\star$. This is easiest to see given a simulation experiment.

```
x <- 1:10  # covariate values are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
cover <- rep(0,1000) # record whether the prediction interval contains the new response Y*
for(r in 1:1000){  # loop 1000 "experiments"
Y <- rnorm(10,1+2*x)  # randomly sample the responses
my.lm <- summary(lm(Y~x))  # compute the point estimates of beta0, beta1 and sigma^2
MSE <- sum(my.lm$residuals^2)/8 # point estimate of sigma^2
beta0.hat <- my.lm$coefficients[1] # storing the estimates of beta0 and beta 1
beta1.hat <- my.lm$coefficients[2]
mean.hat <- beta0.hat + beta1.hat*4 # the estimated conditional mean response
se.term <- (1+1/10 + ((4-mean(x))^2) / (sum((x-mean(x))^2))) # the 1 + 1/n + (x-xbar)^2/sum(xi -
my.PI <- c(mean.hat + qt(0.025,8)*sqrt(MSE*se.term), mean.hat + qt(0.975,8)*sqrt(MSE*se.term)) #
Y.star <- rnorm(1,1+2*4) # sample a new response Y*
cover[r] <- ifelse(my.PI[1] < Y.star & my.PI[2] > Y.star, 1, 0) # check if interval contains ("co
}
mean(cover)  # realized coverage proportion, should be about 95%
```

```
## [1] 0.953
```

8.1.6 Simultaneous confidence bands for the regression line

If experimenters want to quantify uncertainty about the entire regression line, then we may compute simultaneous confidence bands. The interpretation of these bands is similar to the interpretation of a confidence interval for a ream response at a given x, except the object we are claiming to capture by the bands is the line itself, rather than a point on the line.

We may construct such bands based on simultaneous confidence intervals for the two parameters the line depends on, which are (β_0, β_1). By *simultaneous*, we mean a confidence set (in two dimensions) that has probability $100(1-\alpha)\%$ to capture the point (β_0, β_1) with respect to repeated experimentation/random sampling.

The Scheffe method may be used to construct simultaneous CIs for (β_0, β_1). The Scheffe simultaneous CIs are given by

$$\left(\hat\beta_0 \pm \sqrt{\frac{2F_{1-\alpha,2,n-2}MSEn^{-1}\sum x_i^2}{\sum(x_i-\overline{x})^2}}\right),$$

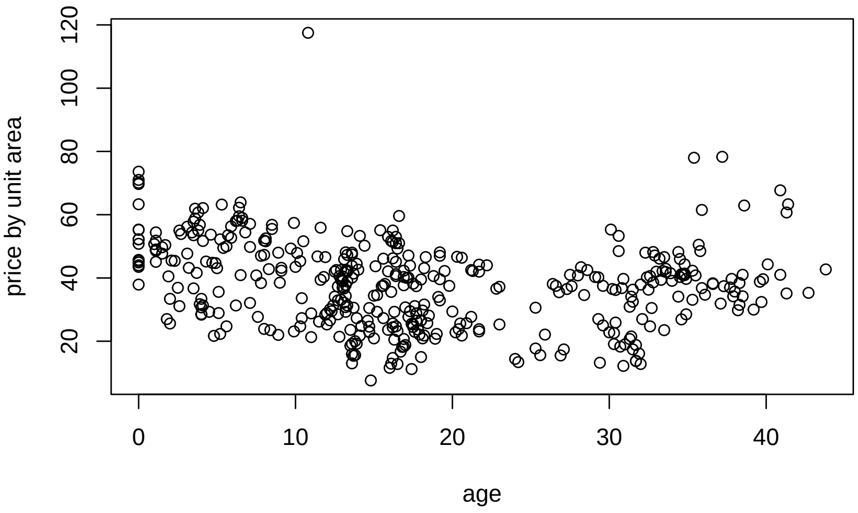
and

$$\left(\hat{\beta}_1 \pm \sqrt{\frac{2F_{1-\alpha,2,n-2}MSE}{\sum(x_i-\overline{x})^2}}\right).$$

A simultaneous confidence band for $\beta_0 + \beta_1 x$ over $x \in (a, b)$ has lower bound $b_0 + b_1 x$ and upper bound $B_0 + B_1 x$ where b_0 and b_1 are the lower bounds of the simultaneous CIs of β_0 and β_1 and B_0 and B_1 are the upper bounds.

8.2 Example: House prices as function of house age

```
houses <- read.csv("Real estate.csv", header = TRUE)
plot(houses$X2.house.age, houses$Y.house.price.of.unit.area, xlab = 'age', ylab = 'pric
```



8.2.1 Fitting the regression

From the output of the summary of teh lm fit we see that the fitted regression line is

$$\text{Price by unit area } = 42.4347 - 0.2515 \text{ age}.$$

The estimated variance is $MSE = \hat{\sigma}^2 = 13.32^2$.

We have the observed test statistics:

$$\frac{\hat{\beta}_0}{\sqrt{\frac{MSEn^{-1}\sum x_i^2}{\sum(x_i-\overline{x})^2}}} = 35.042$$

and

$$\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum(x_i-\overline{x})^2}}} = -4.372$$

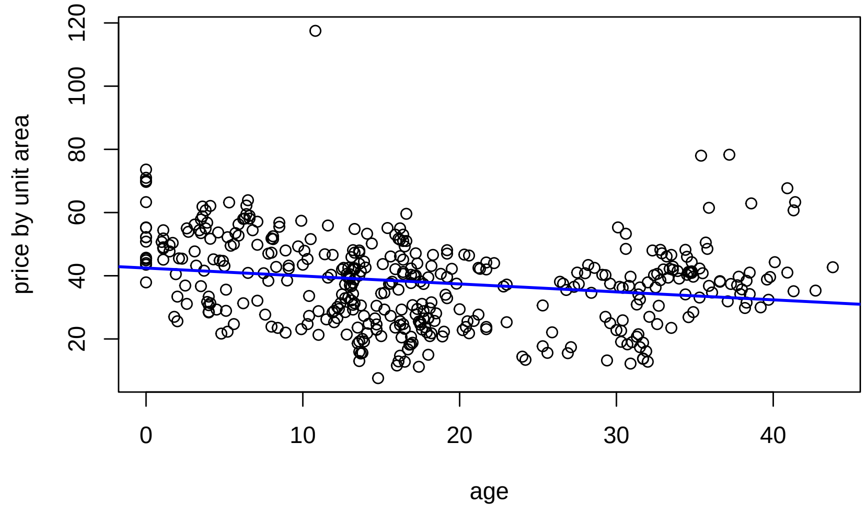
which are both significant, meaning the intercept and slope are not zero.

```r
my.lm <- lm(Y.house.price.of.unit.area~X2.house.age, data = houses)
summary(my.lm)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ X2.house.age, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.113 -10.738   1.626   8.199  77.781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.43470    1.21098  35.042  < 2e-16 ***
## X2.house.age -0.25149    0.05752  -4.372 1.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.32 on 412 degrees of freedom
## Multiple R-squared:  0.04434,    Adjusted R-squared:  0.04202
## F-statistic: 19.11 on 1 and 412 DF,  p-value: 1.56e-05
```

The fitted line, overlaid on the data:

```r
houses <- read.csv("Real estate.csv", header = TRUE)
plot(houses$X2.house.age, houses$Y.house.price.of.unit.area, xlab = 'age', ylab = 'price by unit
abline(a = 42.43470, b = -.25149, col = 'blue', lwd = 2)
```

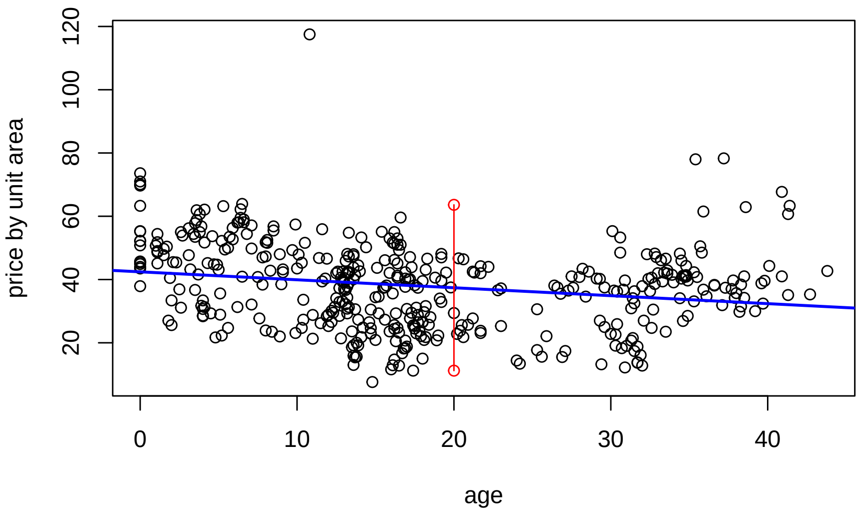


Predicting the home price of a house that is 20 years old:

```
my.summ <- summary(my.lm)
MSE <- my.summ$sigma^2
n<- length(houses$Y.house.price.of.unit.area)
se.term <- 1+ 1/n + ((20-mean(houses$X2.house.age))^2) / sum((houses$X2.house.age - mea
PI <- c(my.summ$coefficients[1] + my.summ$coefficients[2]*20 + qt(0.025, my.summ$df[2]]
PI
```

```
## [1] 11.19322 63.61663
```

```
houses <- read.csv("Real estate.csv", header = TRUE)
plot(houses$X2.house.age, houses$Y.house.price.of.unit.area, xlab = 'age', ylab = 'pric
abline(a = 42.43470, b = -.25149, col = 'blue', lwd = 2)
lines(c(20,20), PI, col = 'red', type = 'o')
```

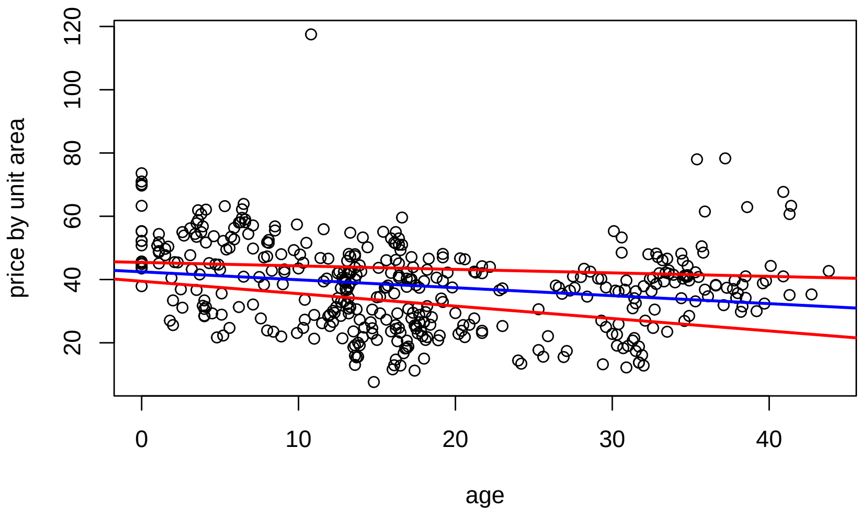


Simultaneous Confidence bands:

```
my.summ <- summary(my.lm)
MSE <- my.summ$sigma^2
n<- length(houses$Y.house.price.of.unit.area)
se.term0 <- (1/n)*sum(houses$X2.house.age^2)/sum((houses$X2.house.age - mean(houses$X2.house.age)
se.term1 <- 1/sum((houses$X2.house.age - mean(houses$X2.house.age))^2)

beta0.intv <- c(my.summ$coefficients[1] - sqrt(2*qf(0.95,2,my.summ$df[2])*MSE*se.term0), my.summ$
beta1.intv <- c(my.summ$coefficients[2] - sqrt(2*qf(0.95,2,my.summ$df[2])*MSE*se.term1), my.summ$
```

```
houses <- read.csv("Real estate.csv", header = TRUE)
plot(houses$X2.house.age, houses$Y.house.price.of.unit.area, xlab = 'age', ylab = 'price by unit
abline(a = 42.43470, b = -.25149, col = 'blue', lwd = 2)
abline(a = beta0.intv[1], b = beta1.intv[1], col = 'red', lwd = 2)
abline(a = beta0.intv[2], b = beta1.intv[2], col = 'red', lwd = 2)
```



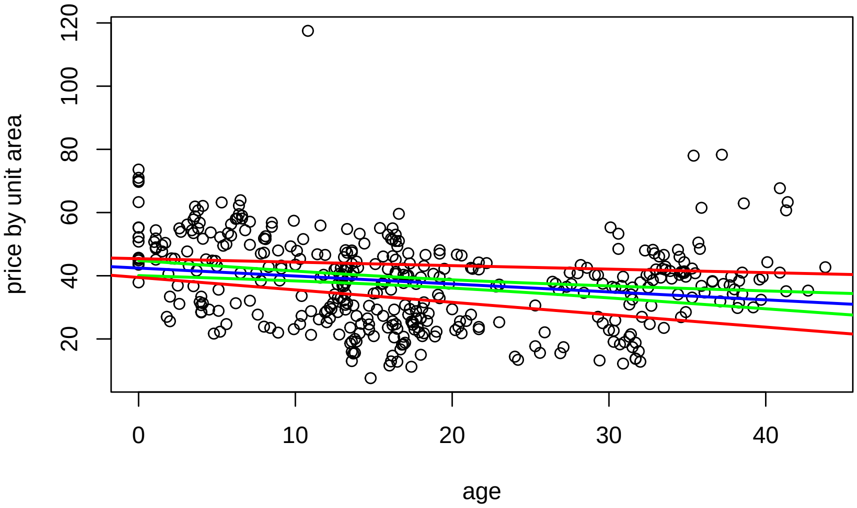
Pointwise confidence bands using CIs for conditional mean responses:

```
my.summ <- summary(my.lm)
MSE <- my.summ$sigma^2
n<- length(houses$Y.house.price.of.unit.area)
CI.fun <- function(x){
  se.term <- 1/n + ((x-mean(houses$X2.house.age))^2) / sum((houses$X2.house.age - mean
  ci <- c(my.summ$coefficients[1] + my.summ$coefficients[2]*x + qt(0.025, my.summ$df[2
  return(ci)
}
```

Pointwise intervals in green:

```
houses <- read.csv("Real estate.csv", header = TRUE)
plot(houses$X2.house.age, houses$Y.house.price.of.unit.area, xlab = 'age', ylab = 'pric
abline(a = 42.43470, b = -.25149, col = 'blue', lwd = 2)

x.seq <- seq(from = 0, to = 50, length.out = 400)
app.fun <- apply(matrix(x.seq, 400,1),1,CI.fun)
lines(x.seq, app.fun[1,], col = 'green', lwd = 2)
lines(x.seq, app.fun[2,], col = 'green', lwd = 2)
abline(a = beta0.intv[1], b = beta1.intv[1], col = 'red', lwd = 2)
abline(a = beta0.intv[2], b = beta1.intv[2], col = 'red', lwd = 2)
```


price by unit area
20
40
60
80
100
120
0
10
20
30
40
age

Chapter 9

Multiple Linear Regression

In this chapter we extend the simple linear regression model to accomodate multiple covariates $X_1, \dots, X_p$ to better predict the response Y.

9.1 The Multiple Linear Regression Model

Multiple linear regression is a model for predicting the behavior of a response variable Y using several predictors (also called covariates or explanatory variables) $X_1, \dots, X_p$ assumed to have some influence on the response. In particular, the model is linear:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + \epsilon_i$$

for $i = 1, \dots, n$ responses. We assume the random residuals ϵ_i are a normal random sample with mean zero and constant variance σ^2.

It is convenient to write the multiple linear regression model in matrix-vector form:

$$Y = X\beta + \epsilon$$

where Y is the $n \times 1$ vector of response variables, X is the $n \times (p+1)$ matrix consisting of an $n \times 1$ column vector of 1's and the p columns of predictor variables, where β is a $(p+1) \times 1$ vector of unknown coefficients, and where ϵ is the $n \times 1$ vector of random residuals.

9.2 Point estimation

Just as in simple linear regression, we estimate β using the method of least squares. Define the observed residuals

$$e(\hat{\beta}) = Y_i - X_i^\top \hat{\beta}$$

where $\hat{\beta}$ is some particular value of the vector β and X_i is the i^{th} row of the "design matrix" X defined above. Then, the sum of squared observed residuals (often called sum of squares errors) is

$$SSE = \sum_{i=1}^{n} \left(Y_i - X_i^\top \hat{\beta} \right)^2 .$$

In matrix-vector notation, the SSE is an inner product (dot product):

$$SSE = (Y - X\hat{\beta})^\top (Y - X\hat{\beta}).$$

The method of least squares says that we should choose the estimator $\hat{\beta}$ to be the value minimizing the SSE. And, just as in simple linear regression, we can compute this minimizer using calculus. In multiple linear regression, the calculus is a bit more complicated, because we are dealing with vector rather than scalar quantities. To compute the least squares estimator, begin by expanding SSE:

$$\begin{aligned} SSE &= (Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\ &= Y^\top Y - Y^\top X\beta - \beta^\top X^\top Y + \hat{\beta}^\top X^\top X\beta \end{aligned}$$

Next, differentiate term by term, using the following facts from calculus: for $n \times 1$ vectors a and x

$$\frac{\partial}{\partial x} a^\top x = a^\top;$$

and, for an $n \times n$ matrix A:

$$\frac{\partial}{\partial x} x^\top A x = (A + A^\top)x.$$

Applying these rules, we obtain

$$\begin{aligned} \frac{\partial}{\partial \beta} SSE &= -Y^\top X - (Y^\top X)^\top + (X^\top X + X^\top X)\beta \\ &= 2X^\top X\beta - 2X^\top Y \end{aligned}$$

If $X^\top X$ is invertible, we may set this equation equal to the $(p+1) \times 1$ zero vector and solve, obtaining

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Just as in simple linear regression, our least squares estimator is unbiased:

$$E(\hat{\beta}) = E((X^\top X)^{-1} X^\top Y) = (X^\top X)^{-1} X^\top X\beta = \beta.$$

To estimate σ^2, we use (an adjustment to) the method of moments,

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)} SSE = \frac{1}{n - (p+1)} (Y - X\hat{\beta})^\top (Y - X\hat{\beta}).$$

This estimator is also unbiased, i.e., $E(\hat{\sigma}^2) = \sigma^2$.

9.3 Sampling Distributions

The estimated regression coefficients have a multivariate normal distribution due to the fact they are equal to linear combinations of the normally-distributed responses:

$$\hat{\beta} \sim N_{p+1}\left(\beta, \sigma^2 (X^\top X)^{-1}\right).$$

This means the least squares estimator is normally distributed, unbiased, and has covariance matrix $\sigma^2(X^\top X)^{-1}$. For example, $V(\hat{\beta}_j) = \sigma^2(X^\top X)^{-1}_{j,j}$, which is the variance parameter σ^2 multiplies by the j^{th} diagonal entry of the matrix $(X^\top X)^{-1}$.

The variance estimator $\hat{\sigma}^2$ (properly scaled) has a Chi-Squared distribution:

$$\frac{[n-(p+1)]\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1).$$

Furthermore, the least squares estimator is independent of the variance estimator, which implies the Studentized estimated coefficients are $t-$distributed:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{j,j}}} \sim t_{n-p-1}.$$

9.4 Inference on regression coefficients

The Student's t sampling distribution noted above enables us to derive exact confidence intervals and hypothesis tests for individual regression coefficients. A $100(1-\alpha)\%$ CI for β_j is given by

$$(\hat{\beta}_j \pm t_{1-\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{j,j}}).$$

To test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ we reject the null hypothesis at level α if $|t| > t_{1-\alpha.2,n-p-1}$ where

$$t = \frac{\hat{\beta}_j - 0}{\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{j,j}}}.$$

The interpretation of the above null hypothesis is that including covariate X_j in the model does not significantly improve accuracy of predicting Y.

We can also test for significance/insignificance of several covariates at a time using *partial F tests*. Let X^R denote a subset of the columns of the design matrix X where we have removed columns corresponding to a subset of r of the coefficients in β. Fit the model $Y = X^R\beta^R + \epsilon$ and compute the sum of squared residuals for this **reduced** model, call it, SSE^R. Let SSE^F denote the sum of squared residuals for the **full** model with all p covariates. Define

$$F = \frac{(SSE^R - SSE^F)/r}{SSE^F/(n-p-1)}.$$

Then, under H_0 : all of these r covariates have coefficients equal to zero we have $F \sim F_{r,n-p-1}$, that is, the test statistic F has an F distribution with numerator and denominator degrees of freedom r and $n-p-1$. We reject the null hypothesis at level α if $F > F_{1-\alpha,r,n-p-1}$.

9.4.1 Example: Housing price data

```
houses <- read.csv("Real estate.csv", header = TRUE)
colnames(houses) <- c('date', 'age', 'distance_to_transit', 'num_stores', 'lat', 'long
```

Let's fit a multiple linear regression model to predict price per unit area using a house's age, its distance to the nearest transit station, and the number of nearby convenience stores. Intuitively, age affects a home's value, while the other two variables have to do with the value of the location and neighborhood.

```
my.lm <- lm(price_per_unit_area~age+distance_to_transit+num_stores, data = houses)
summary(my.lm)
```

```
##
## Call:
## lm(formula = price_per_unit_area ~ age + distance_to_transit +
##     num_stores, data = houses)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.013025 -0.003845 -0.000584  0.002303  0.052341
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.206e+02  3.203e+00   37.65   <2e-16 ***
## age                  4.619e-04  1.591e-03    0.29    0.772
## distance_to_transit -3.774e-05  3.932e-05   -0.96    0.338
## num_stores          -9.802e-06  3.555e-07  -27.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0091 on 410 degrees of freedom
## Multiple R-squared:  0.651,  Adjusted R-squared:  0.6484
## F-statistic: 254.9 on 3 and 410 DF,  p-value: < 2.2e-16
```

The fitted regression line is

$$\text{price per unit area} = 120.6+0.0004619\times\text{age}-0.00003774\times\text{distance to transit}-0.000009802\times\text{number of stores}.$$

However, t-tests suggest that neither the inclusion of age nor distance to transit in the model are improving the price predictions (p-values are 0.772 and 0.338).

```
X <- model.matrix(price_per_unit_area~age+distance_to_transit+num_stores, data = houses)
hat.sigma2 <- summary(my.lm)$sigma^2
cov.beta <- hat.sigma2*solve(t(X)%*%X)
cov.beta[2,3]/sqrt(cov.beta[2,2]*cov.beta[3,3])
```

```
## [1] -0.01602387
```

```
cov.beta[2,4]/sqrt(cov.beta[2,2]*cov.beta[4,4])
```

```
## [1] -0.06045947
```

```
cov.beta[3,4]/sqrt(cov.beta[3,3]*cov.beta[4,4])
```

```
## [1] -0.0246031
```

Above we computed the estimated correlation between the coefficients of age, distance to transit, and number of stores. These estimated correlations are all close to zero, meaning the estimated coefficients are not correlated. When estimated coefficients are strongly correlated, it is challenging to interpret the multiple linear regression model. For example, the estimate of $\widehat{\beta}_1 = 0.0004619$ means that for a one unit increase in age, the home price by unit area increases by 0.0004619 units **provided all other covariates are held constant**. If the covariates are not correlated, this explanation makes sense. However, if two covariates are strongly, say, positively correlated, then when one goes up it must be that the other goes up as well, so it is not reasonable to assume one can be changed while the other is held constant. Correlation of the covariates is called *multicollinearity* and it is important to check for multicollinearity before attempting to interpret coefficients as above.

Note that the summary of the lm function call includes an F test result in the last line. It says "F-statistic: 254.9 on 3 and 410 DF, p-value: $<$ 2.2e-16". This is called the *model F test* and it is a partial F test of the hypothesis that only the intercept coefficient is non zero, i.e. $\beta_0 \neq 0$ but $\beta_j = 0$ for all $j > 0$. We can match this F test result by fitting the "intercept-only" model and comparing SSEs:

```
my.lm.int <- lm(price_per_unit_area~1, data = houses)
SSE.R <- sum(my.lm.int$residuals^2)
SSE.F <- sum(my.lm$residuals^2)
n <- length(houses$price_per_unit_area)

F <- ((SSE.R - SSE.F)/3)/(SSE.F / (n-4))
F
```

```
## [1] 254.923
```

```
1-pf(F,3,n-4)
```

```
## [1] 0
```

Chapter 10

Likelihood-Based Inferences

Previously, we discussed approximate sampling distribution theory for maximum likelihood estimators. Recall that if $\hat{\theta}$ is the MLE of θ, the parameter of a distribution family P_θ, then (under some technical conditions)

$$\hat{\theta} \dot{\sim} N(\theta, I(\theta)^{-1})$$

for large n where $I(\theta)$ is the Fisher information

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta}\ell(\theta; X_1, \dots, X_n)\right]$$

for a random sample $X_1, \dots, X_n$ from P_θ.

In this chapter we discuss how to use this (and other) approximate (large sample) sampling distributions to make inferences about population parameters.

10.1 Wald tests and CIs

A $100(1-\alpha)\%$ Wald approximate CI for a scalar parameter θ has the form

$$\left(\hat{\theta} \pm z_{1-\alpha/2}\sqrt{I(\hat{\theta})^{-1}}\right)$$

This CI is straightforward to obtain by the same argument we have used before. Start with the identity

$$1-\alpha \approx P\left(z_{\alpha/2} < \frac{\hat{\theta}-\theta}{\sqrt{I(\theta)^{-1}}} < z_{1-\alpha/2}\right)$$

which follows from the approximate pivot discussed above, and in Chapter 2. Then, simply multiply by the standard error $\sqrt{I(\theta)^{-1}}$, subtract $\hat\theta$ from both sides, and multiply by -1 to get

$$1-\alpha \approx P\left(\hat\theta - z_{1-\alpha/2}\sqrt{I(\theta)^{-1}} < \theta < \hat\theta + z_{1-\alpha/2}\sqrt{I(\theta)^{-1}}\right)$$

Finally, since θ is unknown, replace θ by $\hat\theta$ in the standard error to yield the so-called Wald type CI

$$\left(\hat\theta \pm z_{1-\alpha/2}\sqrt{I(\hat\theta)^{-1}}\right).$$

If $\theta = (\theta_1, \dots, \theta_p)$ is a vector, then, the marginal Wald approximate CI for θ_j is given by

$$\left(\hat\theta_j \pm z_{1-\alpha/2}\sqrt{I(\hat\theta)^{-1}_{j,j}}\right).$$

where $I(\theta)$ is the Fisher information *matrix* and $I(\theta)_{j,j}$ is the j^{th} diagonal entry of the Fisher information matrix.

Besides CIs, the approximate pivot may be used for hypothesis tests as well. To test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ for a scalar parameter θ, reject the null hypothesis if $|z| > z_{1-\alpha/2}$ where

$$z = \frac{\hat\theta - \theta_0}{\sqrt{I(\theta_0)^{-1}}}.$$

(Alternatively, one may use $\hat\theta$ in the standard error rather than θ_0. Under the null hypothesis, there is little difference.)

For a vector parameter θ let

$$W = (\hat\theta - \theta_0)^\top I(\theta_0)^{-1}(\hat\theta - \theta_0).$$

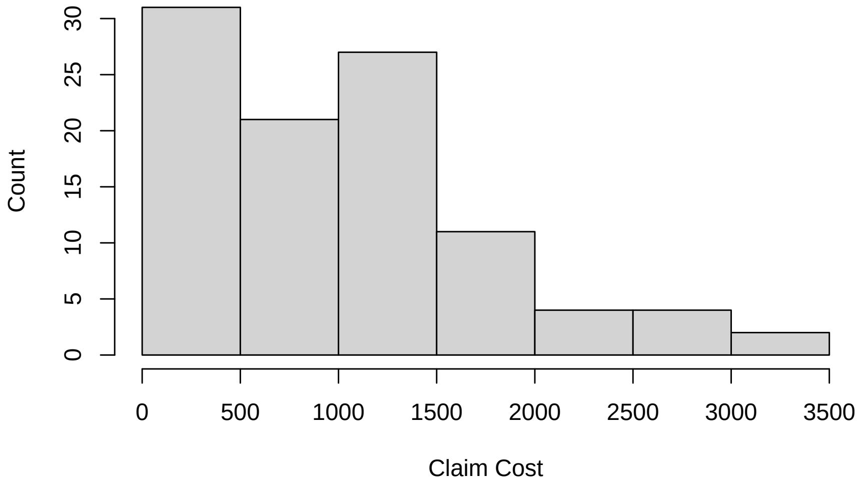
Under $H_0 : \theta = \theta_0$, $W \dot\sim \chi^2(p)$. And, if we use the estimated covariance:

$$V = (\hat\theta - \theta_0)^\top I(\hat\theta)^{-1}(\hat\theta - \theta_0),$$

then, $V \dot\sim F(p, n-p)$. So, we get either a Chi-Squared or F-based test for a vector parameter.

10.1.1 Example: Auto insurance claims

We have the following data on 100 auto insurance claims:



Given the auto claims are positively skewed, we will assume a Gamma distribution reasonably models the population and fit a Gamma to the data by maximum likelihood.

The Gamma density has form

$$f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}, \quad x, \alpha, \beta > 0$$

where (α, β) are shape and rate parameters. The likelihood for a random sample of size n is

$$L(\alpha, \beta; X_1, \dots, X_n) = \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^n \prod_{i=1}^{n} (x_i)^{\alpha-1} e^{-\beta \sum_{i=1}^{n} x_i}.$$

Taking the log, we have the loglikelihood function (suppressing dependence on data)

$$\ell(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \beta \sum_{i=1}^{n} x_i.$$

Next, we need the gradient (the first derivatives) w.r.t. the parameters (α, β):

$$\frac{\partial \ell}{\partial \alpha} = n \log \beta - \text{DiGamma}(\alpha) + \sum_{i=1}^{n} \log(x_i)$$

where the function DiGamma(α) is the logarithmic derivative of the Gamma function. Next, we have

$$\frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i.$$

We also need the second derivatives in order to compute the Fisher information:

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -\text{TriGamma}(\alpha)$$

where the TriGamma function is (you guessed it!) the second logarithmic derivative fo the Gamma function $\Gamma(\alpha)$.

$$\frac{\partial^2 \ell}{\partial \alpha \partial \beta} = n/\beta.$$

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\frac{n\alpha}{\beta^2}.$$

You may have guessed we will not be able to compute the MLEs by hand by solving $\frac{\partial \ell}{\partial \alpha} = 0$ and $\frac{\partial \ell}{\partial \beta} = 0$. Instead, we need an iterative solver (an algorithm). In R we can use *optim* to solve for the MLEs. Or, we can code our own version of Newton's method...

```
n <- length(data)
loglik <- function(param){
  sum(dgamma(data, shape = param[1], rate = param[2], log = TRUE))
}
grad <- function(param){
  g1 <- n*log(param[2]) - n*digamma(param[1]) + sum(log(data))
  g2 <- n*param[1]/param[2] - sum(data)
  return(c(g1,g2))
}
H <- function(param){
  h1 <- -n*trigamma(param[1])
  h2 <- n/param[2]
  h3 <- -n*param[1]/(param[2]^2)
  return(matrix(c(h1,h2,h2,h3),2,2))
}

m <- mean(data)
v <- var(data)

beta0 <- m/v
alpha0 <- m*beta0

delta <- 0.00001
```

```
diff.step <- 1
par.old <- c(alpha0, beta0)
steps <- 0
while((diff.step > delta) & (steps < 100)){
  par.new <- par.old - solve(H(par.old))%*%matrix(grad(par.old),2,1)
  diff.step <- max(abs(par.new - par.old))
  steps <- steps + 1
  par.old <- par.new
}
par.old
```

```
##             [,1]
## [1,] 1.966640491
## [2,] 0.001948425
```

```
loglik(par.old)
```

```
## [1] -780.5969
```

```
H(par.old)
```

```
##             [,1]         [,2]
## [1,]   -65.86955     51323.52
## [2,] 51323.51588 -51803341.74
```

```
solve(H(par.old))
```

```
##               [,1]          [,2]
## [1,] -6.657160e-02 -6.595499e-05
## [2,] -6.595499e-05 -8.464786e-08
```

```
mle.optim <- optim(c(alpha0, beta0), loglik, gr = grad, control = list(fnscale = -1), method = 'B
```

```
## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced
## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced

## Warning in dgamma(data, shape = param[1], rate = param[2], log = TRUE): NaNs
## produced
```

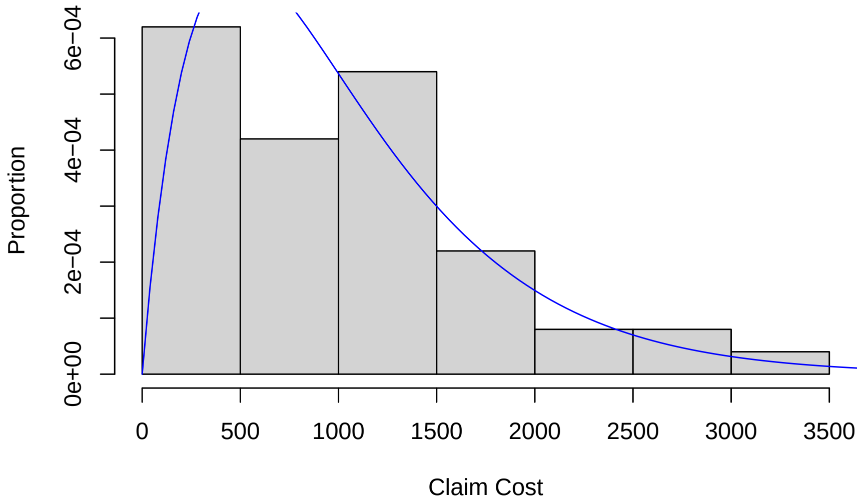
```
mle.optim
```

```
## $par
## [1] 1.966641732 0.001948426
##
## $value
## [1] -780.5969
##
## $counts
## function gradient
##       34        4
##
## $convergence
## [1] 0
```

```
##
## $message
## NULL
##
## $hessian
##               [,1]          [,2]
## [1,]   -65.86951     54017.32
## [2,] 54017.31581 -70328540.92
```

```
solve(mle.optim$hessian)
```

```
##                 [,1]           [,2]
## [1,] -4.101660e-02 -3.150366e-05
## [2,] -3.150366e-05 -3.841603e-08
```



Chapter 11

Bootstrap

11.1 Intro to the bootstrap

For hypothesis testing and confidence intervals, there is a “statistic” whose sampling distribution is required

For example, to test $H_0 : \mu = \mu_0$ using a normal random sample $X_1, ..., X_n$ with $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom given the null hypothesis is true.

Only in very simple problems can we determine an “exact” test statistic. For example, in the two-sample t-test with unknown, unequal population variances, the T statistic is only approximate.

We often rely on assumptions (like normality) or large sample sizes (CLT) in order to derive approximate (or asymptotic) distributions of statistics.

The bootstrap (in a manner similar to Monte Carlo) tries to approximate the distribution of a test statistic using simulations.

For a distribution with cdf F, we are interested in a parameter $\theta = \phi(F)$, and functional of F.

Examples: 1. Mean - $\phi(F) = \int x dF(x)$ 2. Anything else that is a function of moments, like variance 3. Median - $\phi(F) = \inf\{x : F(x) \geq 0.5\}$…

Given data $X_1, ..., X_n$ the empirical CDF (think of this as an estimate of F) is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}.$$

A natural estimate of θ is the "plug-in" estimator $\hat{\theta} = \phi(\hat{F})$.

Given the previous setup of CDF and empirical CDF, suppose we want to estimate θ. If we could somehow simulate from F, then our knowledge of Monte Carlo suggests the following: 1. Simulate many data sets, say M, of size n: $X_{1,1}, ..., X_{n,1}$, ..., $X_{1,M}, ..., X_{n,M}$. 2. From the M data sets, compute $\hat{\theta}_i$ for $i = 1, ..., M$. 3. Then, our many estimates should given us an idea of the mean and the variance of $\hat{\theta}$.

This actually works when we know the distribution of the test statistic. In the one-sample t-test for a normal mean, the test statistic has a $t(n-1)$ distribution. So, if we sample from this distribution many times, we will be able to find the rejection region for the hypothesis test by computing Monte Carlo estimates of the $t(n-1)$ quantiles, but this is rather trivial...

```
qt(0.025,9)
```

```
## [1] -2.262157
```

```
qt(0.975,9)
```

```
## [1] 2.262157
```

```
quantile(rt(100000,9), 0.975)
```

```
##    97.5%
## 2.272845
```

Of course, we (usually) cannot simulate from F because it is unknown (depends on unknown parameters).

Bootstrap idea: Replace simulation from F with simulation from $\hat{F}$!

Note: it is easy to simulate from $\hat{F}$, it's just simulation from a discrete uniform distribution with probabilities $1/n$ on each data point.

Suppose we are given the following two samples:

```
set.seed(3214354)
x <- rexp(10,1)
y <- rexp(9,0.8)
round(x,2)
```

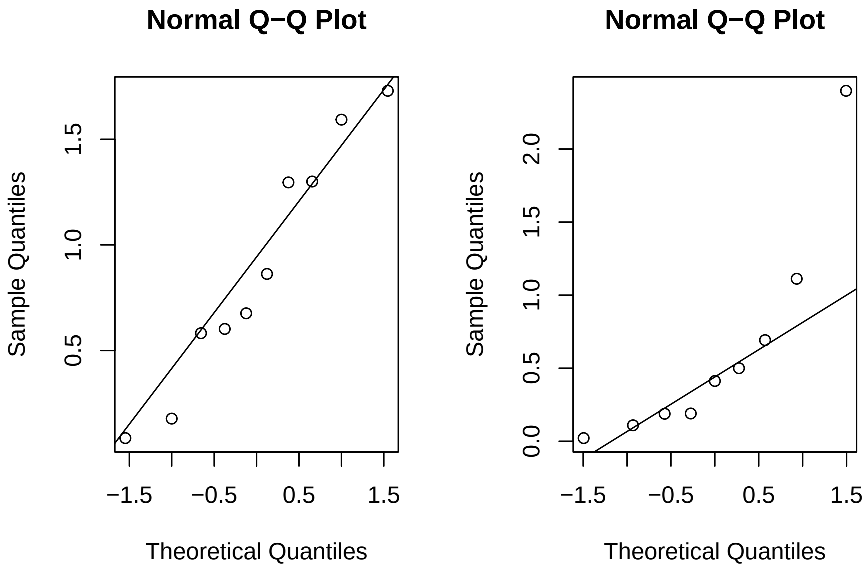
```
##  [1] 0.18 1.30 0.09 1.30 1.73 0.86 1.59 0.60 0.58 0.68
```

```
round(y,2)
```

```
## [1] 0.02 0.50 0.41 0.19 0.19 0.69 2.40 1.11 0.11
```

We could use the t-test, but wait…

```
par(mfrow = c(1,2))
qqnorm(x)
qqline(x)
qqnorm(y)
qqline(y)
```



```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.94146, p-value = 0.5694
```

```
shapiro.test(y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y
## W = 0.77085, p-value = 0.009449
```

What “test statistic” should I use to test $H_0 : \mu_X = \mu_Y$? Why not $\bar{X} - \bar{Y}$?

```
test.statistic <- mean(x) - mean(y)
boot.test.statistic <- rep(NA, 10000)
for(i in 1:10000){
  boot.x <- sample(x,replace = T)
  boot.y <- sample(y, replace = T)
  boot.test.statistic[i] <- mean(boot.x)-mean(boot.y)
}
test.statistic
```

```
## [1] 0.2657391
```

```
quantile(boot.test.statistic,0.025)
```

```
##      2.5%
## -0.345881
```

```
quantile(boot.test.statistic,0.975)
```

```
##    97.5%
## 0.792484
```

```
m <- median(boot.test.statistic)

approx.p.value = ifelse(test.statistic > m, sum((boot.test.statistic-m) > (test.statis
approx.p.value
```

```
## [1] 0.9726
```

11.2 Bootstrap Confidence Intervals

A variety of methods are used to construct confidence intervals using the bootstrap distribution of an estimator. We'll discuss the percentile, basic, and Studentized methods, although several others are available, including the bias-corrected and accelerated (BCa) intervals computed by default in R's boot package.

Suppose the bootstrap samples $\hat{\theta}_1, \dots, \hat{\theta}_B$ are available. The simplest bootstrap interval estimate is given by the percentile method, which defines the $100(1-\alpha)\%$ CI to be the $\alpha/2$ to $(1-\alpha/2)$ sample quantiles of $\hat{\theta}_1, \dots, \hat{\theta}_B$. There are good arguments against the percentile method, but we won't wade into the details presently. Those arguments tend to favor the basic bootstrap method, which defines the $100(1-\alpha)\%$ CI to be the interval $(2\overline{\theta} - \hat{\theta}_{1-\alpha/2},\, 2\overline{\theta} - \hat{\theta}_{\alpha/2})$, which is twise the sample mean of the bootstrapped estimates minus the $(1-\alpha/2)$ and $\alpha/2$ sample quantiles of those bootstrapped estimates $\hat{\theta}_1, \dots, \hat{\theta}_B$.

The last method we'll discuss is the Studentized bootstrap. The idea of the method is to mimic the usual z- or t-test pivotal quantity $\frac{\hat{\theta}-\theta}{se(\hat{\theta})}$. First, compute $\hat{\theta}$ for teh original data. Then, perform the following nested bootstrap, for b in 1 to B do: 1.Resample the data with replacement to generate the b^{th} bootstrap data set a. for this b^{th} data set compute the point estimate $\tilde{\theta}_b$ b. Resample the resamples! Generate K nested bootstrap resamples sample of data. For each of them compute $\tilde{\theta}_{b,k}$ for $k = 1, \dots, K$. c. Estimate $se(\tilde{\theta})$ as the standard deviation of $\tilde{\theta}_{b,k}$ for $k = 1, \dots, K$. d. Keep the studentized value $q_b = \frac{\tilde{\theta}_b - \hat{\theta}}{se(\tilde{\theta}_b)}$ 2. Let $se(\hat{\theta})$ be estimated by the sample standard deviation of $\tilde{\theta}_1, \dots, \tilde{\theta}_B$. 3. Define the Studentized interval as

$$(\hat{\theta} - q_{1-\alpha/2} se(\hat{\theta}),\, \hat{\theta} - q_{\alpha/2} se(\hat{\theta}))$$

where q_α is the α sample quantile of the Studentized q values.

```r
set.seed(3214354)
x <- rexp(10,1)
y <- rexp(9,0.8)
round(x,2)
```

```
##  [1] 0.18 1.30 0.09 1.30 1.73 0.86 1.59 0.60 0.58 0.68
```

```r
round(y,2)
```

```
## [1] 0.02 0.50 0.41 0.19 0.19 0.69 2.40 1.11 0.11
```

```
my.boot <- function(x,y, B, K, alpha){
  hattheta <- mean(x) - mean(y)
  tildetheta <- rep(NA,B)
  q <- rep(NA,B)
  for(b in 1:B){
  indices1 <- sample(10,10,replace = TRUE)
  indices2 <- sample(9,9,replace = TRUE)
  xstar <- x[indices1]
  ystar <- y[indices2]
  tildetheta[b] <- mean(xstar) - mean(ystar)
  tilde2 <- rep(NA,K)
    for(k in 1:K){
        indices1 <- sample(10,10,replace = TRUE)
        indices2 <- sample(9,9,replace = TRUE)
        xstar2 <- xstar[indices1]
        ystar2 <- ystar[indices2]
        tilde2[k] <- mean(xstar2)-mean(ystar2)
    }
  q[b] <- (tildetheta[b] - hattheta)/sd(tilde2)
  }
  intv <- hattheta - sd(tildetheta)*quantile(q,c(1-alpha/2,alpha/2))
  return(as.numeric(intv))
}
my.boot(x,y,1000,50,0.05)
```

```
## [1] -0.6147748  0.8576096
```

11.3 Bootstrapping linear models

When we have only one sample of iid data, bootstrapping is straightforward: simply sample with replacement n times from the original data to obtain a single bootstrap-resampled data set. Repeat many times, recording the estimate corresponding to each resampled set. When there is more “structure” present in the data we need to think carefully about how that structure should be treated by the resampling procedure.

In linear models, i.e., the Gauss-Markov model, for each response Y_i there is a corresponding covariate vector X_i. How should these pairs (Y_i, X_i) be treated by the resampling procedure? One method for bootstrapping a linear model is called row-resampling. Imagine binding the $n \times 1$ column of responses Y to the $n \times p$ matrix of covariates X and then resampling rows with replacement. We can do this by sampling the integers $1, \ldots, n$ with replacement, each time taking the corresponding row of (Y, X) as the bootstrap resample. Row resampling treats the pairs (Y_i, X_i) as random, but this is not always reasonable. For example, in a two-sample experiment in which we apply a treatment to 10 of 20 subjects at random and keep the remaining 10 as a control group, row resampling will produce resampled data sets with more than ten in one of the two groups and fewer in the other, which is not how the experiment was performed. When the *design* is fixed, e.g., we have precisely ten subjects in each group, a different bootstrap method, called residual boostrap, makes more sense. The residual bootstrap starts with the residuals $e = y - \hat{y} = y - X_i\hat{\beta}$. We resample the residuals to obtain $e^\star = (e_1^\star, \dots, e_n^\star)^\top$. Next, we compute the bootstrapped responses, $y_i^\star = X_i\hat{\beta} + e_i^\star$. Finally, we compute $\hat{\beta}^\star$ using X and $y^\star$. This way, the design matrix X remains the same throughout the bootstrap procedure, preserving the structure of the experiment. Example: Consider the following experiment to evaluate the effect of a bacterial infection on the lifespan of Guinea pigs as measure in days. First, we apply the ro-resampling method. Admittedly, this method makes little sense if we intend for the experiment to have 64 pigs in the control group and 58 in the treatment group.

```
pigs <- read.csv('guinea_pigs.csv')
my.lm <- lm(Time~Treatment, data = pigs)
my.lm
```

```
##
## Call:
## lm(formula = Time ~ Treatment, data = pigs)
##
## Coefficients:
##      (Intercept)  TreatmentControl
##            242.5             102.7
```

```
my.lm$coefficients
```

```
##      (Intercept) TreatmentControl
##         242.5345         102.6843
```

```
n <- length(pigs$Pig)

r.r.coefs <- matrix(NA,1000,2)
row.resample.boot <- function(indices){
  data.star <- pigs[indices,]
  my.lm <- lm(Time~Treatment, data = data.star)
  return(as.numeric(my.lm$coefficients))
}
for(i in 1:1000){
  r.r.coefs[i,] <- row.resample.boot(sample.int(n,n,TRUE))
}
```

```
r.r.means <- colMeans(r.r.coefs)
percentiles1 <- quantile(r.r.coefs[,1], c(0.05,0.95))
percentiles1
```

```
##       5%      95%
## 218.0000 268.3603
```

```
percentiles2 <- quantile(r.r.coefs[,2], c(0.05,0.95))
percentiles2
```

```
##        5%       95%
##  49.40079 154.59927
```

```
c(2*r.r.means[1] - percentiles1[2], 2*r.r.means[1] - percentiles1[1])
```

```
##      95%       5%
## 216.4051 266.7653
```

```
c(2*r.r.means[2] - percentiles2[2], 2*r.r.means[2] - percentiles2[1])
```

```
##       95%        5%
##  50.48009 155.67856
```

Next, we implement the residual-bootstrap method. Note the two methods give very similar 90% basic bootstrap CIs for β_2 but disagree a bit for β_1.

```
pigs <-
read.csv('guinea_pigs.csv')
my.lm <- lm(Time~Treatment, data = pigs)
my.lm
```

```
##
## Call:
## lm(formula = Time ~ Treatment, data = pigs)
##
## Coefficients:
##      (Intercept)  TreatmentControl
##            242.5             102.7
```

```
my.lm$coefficients
```

```
##      (Intercept) TreatmentControl
##         242.5345         102.6843
```

```
n <- length(pigs$Pig)
resids <- my.lm$residuals

r.r.coefs <- matrix(NA,1000,2)
resid.boot <- function(indices){
  resids.star <- resids[indices]
  y.star <- cbind(rep(1,n), ifelse(pigs$Treatment=="Control",1,0))%*%matrix(as.numeric(my.lm$coef
  data.star <- pigs
  data.star$y.star <- y.star
  my.lm <- lm(y.star~Treatment, data = data.star)
  return(as.numeric(my.lm$coefficients))
}
for(i in 1:1000){
  r.r.coefs[i,] <- resid.boot(sample.int(n,n,TRUE))
}
r.r.means <- colMeans(r.r.coefs)
percentiles1 <- quantile(r.r.coefs[,1], c(0.05,0.95))
percentiles1
```

```
##       5%      95%
## 205.1938 281.0952
```

```
percentiles2 <- quantile(r.r.coefs[,2], c(0.05,0.95))
percentiles2
```

```
##        5%       95%
##  50.20677 154.69151
```

```
c(2*r.r.means[1] - percentiles1[2], 2*r.r.means[1] - percentiles1[1])
```

```
##      95%       5%
## 203.5823 279.4838
```

```
c(2*r.r.means[2] - percentiles2[2], 2*r.r.means[2] - percentiles2[1])
```

```
##       95%        5%
##  52.21648 156.70123
```

Chapter 12

Bayesian Statistics

12.1 Introduction

All of the statistical methods we have discussed so far could be rightly described as "frequentist" in nature. Frequentist statistics, broadly speaking, is concerned with the random behavior of statistics (functions of randomly sampled data) over repeated sampling from the population under study. In frequentist statistics probability is used to describe random behavior induced by random sampling. And, by and large, the effectiveness of frequentist methods—like tests and confidence intervals—is judged by their random behaviors over repeated sampling; e.g., by the type 1 error and power of tests, and the coverage probability of interval estimators.

In contrast, Bayesian statistics uses probability in two distinct ways: to describe random sampling variability in much the same way as frequentist statistics, and to characterize the uncertainty in unknown parameters. This second use of probability is novel, compared to frequentist statistics (although Bayesian statistics came first chronologically). In frequentist statistics unknown parameters—like the unknown mean μ of a normal population—are treated as unknown constant variables. In Bayesian statistics, unknown parameters are (mostly) treated as random variables—but, users of Bayesian statistics do not necessarily believe them to be random variables.

There are some interesting consequences of treating parameters as a random variables. For one, information or assumptions about the parameters apart from the data can be used to assign a *prior probability distribution* to the parameters, effectively describing their random behavior, or, the beliefs about likely values of those parameters before any data from the population is collected. Second, treating the parameters as random variables allows one to assign probabilities to assertions about the parameters, e.g., $P(\mu \leq 2)$ where μ is the unknown mean of a normal population under study. In frequentist statistics, no such (non-trivial) probability makes sense. However, the interpretation of such *Bayesian* probabilities is somewhat unclear—they are not frequentist probabilities describing behavior of the parameter with respect to repeated sampling because the parameter has no random sampling distribution.

12.2 Bayesian methods in general

Bayesian statistics is similar to frequentist likelihood-based inference in that both start with a likelihood describing the sampling distribution of the data, say, $f(x;\theta)$. For iid data we have $L(\theta) = \prod_{i=1}^n f(x_i;\theta)$ as usual. Bayesian statistics, however, requires an additional ingredient, the prior distribution mentioned above, denoted $\pi(\theta)$. Recall that the likelihood is the joint distribution of the data given the parameter value θ. But, now regard the parameter as a random variable, and see that the likelihood is the conditional distribution of the data given the random parameter takes on a certain value. Since the joint distribution of random variables (denoted g below) is equal to the product of their conditional and marginal distributions, we have

$$g(X_1,\dots,X_n,\theta) = f(X_1,\dots,X_n|\theta)\pi(\theta) = L(\theta)\pi(\theta).$$

All Bayesian inference (and prediction) is based on the *posterior distribution* $\Pi_n(\theta)$, which is the conditional distribution of the parameter θ given the observed data, and is derived using Bayes' Rule:

$$\Pi_n(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\vartheta)\pi(\vartheta)d\vartheta}$$

where the denominator is the marginal density of the data, $f(x_1,\dots,x_n) = \int f(x_1,\dots,x_n,\vartheta)d\vartheta$, obtained by integration in the usual manner.

If the prior distribution represents what is known (or assumed) about the parameter **before** data is observed, then the posterior distribution represents the up-to-date knowledge (or beliefs) about the parameter **after** data is observed.

The posterior distribution for a parameter can usually be used to recreate methods analogous to frequentist ones. For example, suppose θ is a scalar parameter so that $\Pi_n(\theta)$ is a univariate density. We could define a $100(1-\alpha)\%$ **credible interval** for θ as the interval of values between the $\alpha/2$ and $1-\alpha/2$ quantiles of $\Pi_n(\theta)$. One might expect that since the Bayesian posterior is based on the likelihood that such an interval would be similar to a confidence interval constructed using a frequentist likelihood-based technique, like a Wald or likelihood ratio-based interval. Usually, that is the case. Hypothesis testing is less straightforward in Bayesian statistics. For example, consider the point null hypothesis $H_0 : \mu = \mu_0$ about a normal population mean. The posterior will be a continuous distribution for μ, and as such, the posterior probability of the null hypothesis will be $P(\mu = \mu_0) = 0$. Therefore, it wouldn't make much sense to **accept** null hypotheses with large posterior probabilities, because point nulls will always have zero probability under continuous posteriors. Alternatively, one might define a testing rule in reference to the posterior credible interval above, e.g., reject the null if μ_0 is not in the $100(1-\alpha)\%$ credible interval for some specific choice of α. In many cases, such a test behave similarly to frequentist-based hypothesis tests.

12.3 Bayesian methods for a single normal population

Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, which represents a normal population with an unknown mean and a known variance. Also, suppose the prior distribution $\pi(\mu)$ is chosen to be $N(\mu_0, \sigma_0^2)$ for known values (μ_0, σ_0^2)—these are often called **hyperparameters**.

The posterior distribution for μ is proportional to $L(\mu)\pi(\mu)$, treating the denominator as a proportionality constant. Multiplying the likelihood and prior, we have

$$\begin{aligned}
\Pi_n(\mu) &\propto L(\mu)\pi(\mu) \\
&= (2\pi\sigma^2)^{-n/2}(2\pi\sigma_0^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma^2}(\mu-\mu_0)^2\right\}
\end{aligned}$$

Look only at the expression in the exponent, and complete the square in μ to get

$$\begin{aligned}
&-\frac{1}{2\sigma^2}\sum_{x_i-\mu}^{2} - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 \\
&\quad = \mu^2\left[-\frac{n}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right] + \mu\left[\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right] - \frac{1}{2\sigma^2}\sum x_i^2 - \frac{\mu_0^2}{2\sigma_0^2} \\
&\quad = -\frac{1}{2\left[\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right]}\left(\mu^2 - 2\mu\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} - \left[\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right]^2\right) - \frac{1}{2\sigma^2}\sum x_i^2 - \frac{\mu_0^2}{2\sigma_0^2} - \frac{1}{2}\frac{\left[\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right]^2}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} \\
&\quad = -\frac{1}{2\left[\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right]}\left(\mu - \frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)^2 + (\text{constants in } \mu).
\end{aligned}$$

The above calculation shows that the part of likelihood×prior depending on μ appears to be a normal density function. Therefore, since the posterior must be a proper density function (integrating to 1) it follows that

$$\Pi_n(\mu) \text{ is } N\left(\text{mean } = \frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}, \text{ variance } = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right).$$

Two remarks about the posterior distribution are quite important. First, notice that the posterior mean may be rewritten as the following weighted average:

$$\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} = \bar{x}\frac{\frac{\sigma^2}{n\sigma_0^2}}{1+\frac{\sigma^2}{n\sigma_0^2}} + \mu_0\frac{1}{1+\frac{\sigma^2}{n\sigma_0^2}}.$$

And, when n is large, the posterior mean is approximately $\bar{x}$ because the weight on $\bar{x}$ approaches 1 while the weight on μ_0 approaches 0. The posterior variance may be rewritten as

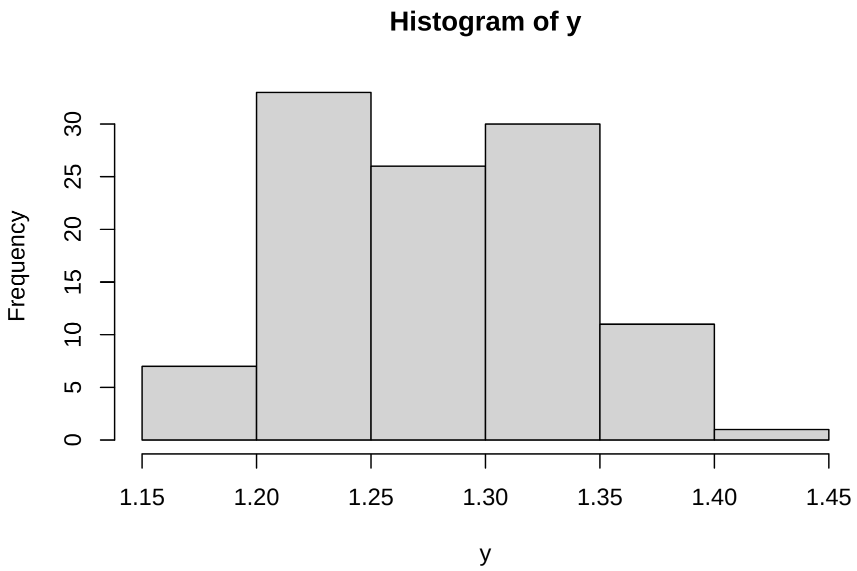
$$\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} = \frac{\sigma^2}{n+\sigma^2/\sigma_0^2} \approx \sigma^2/n$$

where the approximation holds for large n. The upshot is that the posterior is almost the same as the sampling distribution of $\overline{X}$ when n is large. Consequently, frequentist and Bayesian inferences (e.g. 95% confidence/credible intervals for μ) will agree when the sample size is large. The second remark is that the posterior distribution turned out to be the same kind of distribution as the prior distribution. The relationship between the prior, likelihood, and posterior in which the prior and posterior are the same type of distribution is called *conjugacy*. In general, the prior and posterior are only conjugate for certain choices of prior for certain likelihood—conjugacy is the exception, not the rule.

12.3.1 Example

A certain transformation of iron levels in water sampled from sites in the Mississippi river is well-modeled by a normal distribution. Previous studies suggest mean iron level around 1.5 (on the transformed scale) with standard deviation 0.05—so the prior distribution is set to be $N(\mu_0 = 1.5, \sigma_0^2 = 0.05^2)$. 108 iron readings are obtained in a new experiment with a mean of 1.277. Assuming a normal variance of $\sigma^2 = 0.05^2$ the posterior is $N(1.279, 0.05^2/109)$. The sampling distribution of $\overline{X}$, on the other hand, is $N(1.277, 0.05^2/108)$. The two are plotted below in red and blue, respectively. The 95% confidence interval for μ is (1.2675, 1.2864) while the corresponding credible interval is (1.2696, 1.2884).

```
library(SASmixed)
y <- (Multilocation$Fe^{-1/2}-1)/(-1/2)
hist(y)
```

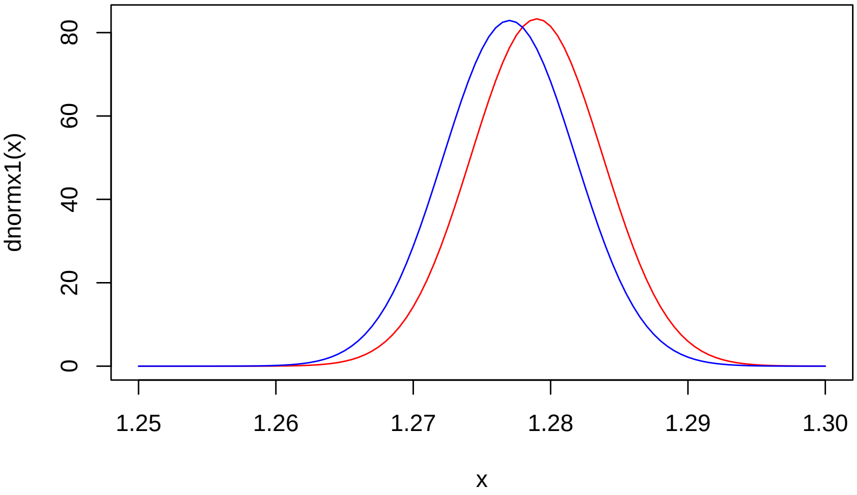


```
mean(y)
```

```
## [1] 1.276776
```

```
dnormx1 <- function(x) dnorm(x, 1.279,sqrt(0.05*0.05/109))
dnormx2 <- function(x) dnorm(x, 1.277,sqrt(0.05*0.05/108))

curve(dnormx1, 1.25,1.3, col = 'red')
curve(dnormx2, 1.25,1.3, col = 'blue', add = TRUE)
```


dnormx1(x)
0
20
40
60
80
1.25
1.26
1.27
1.28
1.29
1.30
x