Experimental Design and Analysis

Zach Branson and Howard Seltman

Preface

This book is intended as required reading material for Experimental Design for the Behavioral and Social Sciences (36-309 and 36-749) at Carnegie Mellon University (CMU). This text is largely based on text written by Professor Howard Seltman, who taught Experimental Design for the Behavioral and Social Sciences for about 10 years at CMU. I did my undergraduate at CMU, and even though I did not take 36-309 during my undergrad, I wrote my senior honors thesis with Howard, who taught me the fundamentals about experimental design and causal inference, which are now my main research interests. During my PhD, I would sometimes visit this textbook to remind myself of these fundamentals. So, it feels very appropriate to revisit this textbook and shape it into my own as I teach this course.

Despite its title, this course involves much more than experimental design or applications in the behavioral and social sciences. In truth, a better title for this course is Experimental Design and Analysis, and that is the title of this book. Experiments are the gold standard of scientific inquiry in many fields, not just the behavioral and social sciences. Why are experiments so ubiquitously considered the gold standard of scientific inquiry? To answer this, let's consider what is perhaps the most well-known mantra of statistics: "Correlation does not imply causation." This begs the question: What *does* imply causation? Many researchers would argue that experiments can imply causation when they are *well-designed* and *well-analyzed*. This course and text aim to teach you how to design and analyze an experiment such that you can make causal claims with scientific merit and impact.

Most experiments involve subjects (e.g., computer chips, mice, humans, classrooms) being randomized to treatments (e.g., drugs, placebos, advertisements, policies) with the aim of characterizing some truth about the world. Pretty much every experiment has two stages: A design stage and an analysis stage. The design stage happens before the experiment is actually conducted, and it addresses questions like: How many (and which) subjects and treatments should we study? How should we randomize subjects to treatments? What information should we record throughout the experiment? Meanwhile, the analysis stage happens after the experiment is conducted, and it utilizes the data resulting from the experiment to answer scientific questions: Does a new drug do a better job treating a disease? Is one teaching style more effective than another? Do the answers to these questions depend on the patients or students we're talking about? There are many choices to make when designing and analyzing an experiment, and we will

use statistical methods (e.g., t-tests, ANOVA, regression, power analyses, hierarchical models) to aid us in these choices. Throughout, we will use computational tools in `R` and datasets motivated by real experiments to get hands-on experience with these methods, such that you will be well-equipped to design, analyze, and interpret real experiments.

Before we dive into those topics, I should say a few things about the structure of this text. Because experiments are so ubiquitous in many fields of study, you probably already have some preconceptions about what is "good" experimental evidence, what constitutes cause-and-effect, and so on. Indeed, a great deal of the design and analysis of experiments consists of common sense and conceptual understanding, and there's only a small dose of statistics and mathematics. That said, some statistical and mathematical tools can be quite useful for understanding the concepts and common sense underpinning experimental design and analysis, which is why 36-200 (the "Stat 101" of CMU) is a prerequisite for this course. The first four chapters of this book act as a review of these prerequisite concepts that will pop up time and time again. Chapter 1 is an overview of what an experiment is and what you should expect to learn from this course. Chapter 2 discusses different types of variables we will encounter in datasets, Chapter 3 reviews the fundamentals of probability—which we will use to better characterize the variables from Chapter 2—and Chapter 4 reviews exploratory data analysis (EDA) techniques that we use for getting a preliminary understanding of our data.

Even though these first four chapters are considered a review, they consist of a lot of material—about one fourth of the book—and I do not necessarily expect you to understand the ins-and-outs of all of the material in those chapters. Like I said, these topics will pop up time and time again, and each time, you will have a chance to understand that topic from a new perspective. Indeed, even though this book follows a linear path from chapter to chapter, a more appropriate trajectory would be a spiral or cyclone: When we first encounter a topic, we may only graze it shallowly, but we will revisit many topics over and over with a deeper understanding each time, eventually "getting at the core" of what's going on in an experiment. So think of the first few chapters as just an initial sweep through this statistical spiral.

That said, there is much more to this text than just reading about statistics. One key idea in this course is that you cannot really learn statistics without doing statistics. Even if you will never analyze data again, getting some hands-on experience analyzing data will take your understanding of experiments (even if

you are only reading about them) to a whole new level. I don't think it makes much difference which statistical package you use for your analyses, but for practical reasons we must standardize on a particular package in this course, and that is `R`. Throughout the book, many chapters include "How to do it in `R`" sections to help you when you go to the computer to conduct analyses. We will engage with many concrete data scenarios throughout this text to better understand how theory applies to practice.

Meanwhile, Chapters 5 through 10 are the real "meat" of the course, where we will discuss how to design and analyze various types of experiments. Chapters 5 and 6 discuss experiments with two or more treatment groups, Chapter 7 discusses the conceptual complications that arise from these experiments, Chapter 8 discusses experiments where multiple treatments are applied, and Chapters 9 and 10 extend these experiments to scenarios where we have additional information about experimental subjects we can take advantage of. Then, Chapters 11 through 15 discuss more advanced topics. For example, what complications arise when we conduct many analyses on the same experimental dataset? What happens when we give the same subject multiple treatments? And what do we do when faced with categorical and non-Normal outcomes?

In this text, I will use some typographical conventions that I hope will aid you in our journey through this statistical spiral of experimental design and analysis. First, as you can see from the last sentence in the previous paragraph, I will use capitalization when referring to the Normal distribution, because I don't want you to think that the Normal distribution has anything to do with the ordinary conversational meaning of "normal." Furthermore, to better pinpoint what exactly the "core message" of the text is, I will summarize key points in a box:

Key points are in boxes. They may be useful at review time to help you decide which parts of the material you know well and which you should reread.

In a similar vein, I will occasionally sum up a larger topic to make sure you haven't "lost the forest for the trees". These are double boxed and start with "In a nutshell":

In a nutshell: You can make better use of the text by paying attention to the typographical conventions.

However, some trees in the forest are really interesting. So, another convention is that optional material has a gray background:

> This text uses a minimal amount of mathematical theory, but many students want a deeper understanding of what they are doing statistically. Optional material (which exams will not be based on, but nonetheless will hopefully give you a more nuanced appreciation for experimental design and analysis) is in a gray box like this.

You may be interested in my background. I was born and raised in Louisville, Kentucky, and from 2010-2014 I did my undergrad at Carnegie Mellon University, where I got a B.S. in Economics & Statistics and a B.A. in Professional Writing. My token "how I got into statistics" story is when I started doing statistical consulting for Pittsburgh Public Schools (PPS) as an undergrad; after I presented my statistical analyses to the PPS school board, I realized how impactful statistics can be when you know how to communicate it to a wide range of people. My undergraduate senior honors thesis about propensity score analyses was advised by Howard Seltman, who wrote the original version of this textbook. I went on to receive a PhD in Statistics from Harvard University in May 2019. At Harvard, I developed a passion for experimental design and causal inference: I helped environmental engineers design experiments and I developed new techniques for designing other kinds of complex experiments (such as those in education); I also worked on methods for estimating causal effects when experiments aren't possible, and I've applied my methods to problems in economics, education, epidemiology, medicine, and political science. I am very excited to teach this course, and I hope to improve this course and text over the years. So, as you read this book and (hopefully) come to my class, please let me know if you ever have any feedback, positive or negative (like any experiment, it would be great to know what is working and what isn't!)

Zach Branson
August 2020

Contents

Chapter 1

The Big Picture

Why experimental design matters.

Much of the progress in the sciences comes from performing experiments. These may be of either an exploratory or a confirmatory nature. Experimental evidence can be contrasted with evidence obtained from other sources such as observational studies, anecdotal evidence, or "from authority". This book focuses on design and analysis of experiments. While not denigrating the roles of anecdotal and observational evidence, the substantial benefits of experiments (discussed below) make them one of the cornerstones of science.

Contrary to popular thought, many of the most important parts of experimental design and analysis require little or no mathematics. In many instances this book will present concepts that have a firm underpinning in statistical mathematics, but the underlying details are not given here. The reader may refer to any of the many excellent textbooks of mathematical statistics listed in the appendix for those details.

This book presents the two main topics of experimental design and statistical analysis of experimental results in the context of the large concept of scientific learning. All concepts will be illustrated with realistic examples, although sometimes the general theory is explained first.

Scientific learning is always an iterative process, as represented in Figure 1.1. If we start at Current State of Knowledge, the next step is choosing a current theory to test or explore (or proposing a new theory). This step is often called "Constructing a Testable Hypothesis". Any hypothesis must allow for *different*

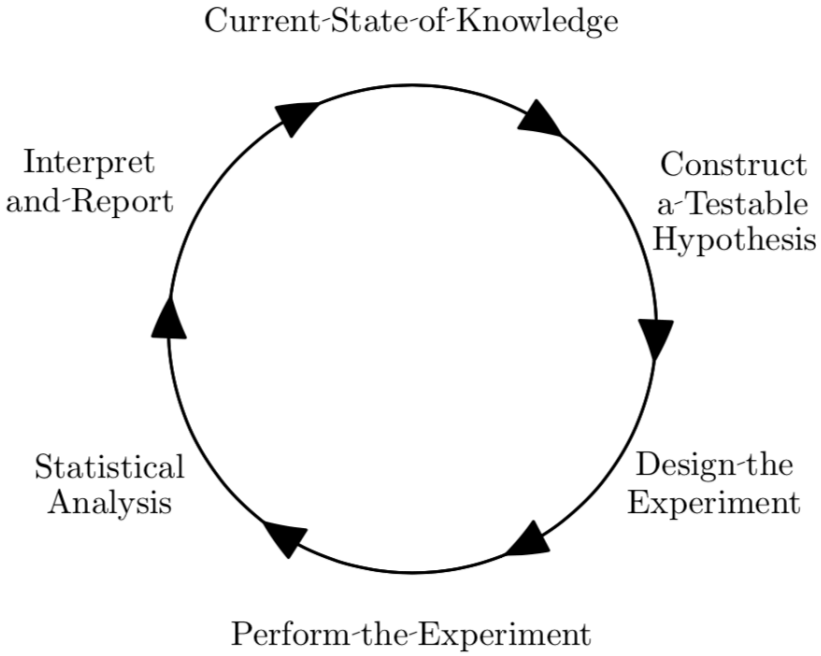


Figure 1.1: The circular flow of scientific learning

possible conclusions or it is pointless. For an exploratory goal, the different possible conclusions may be only vaguely specified. In contrast, much of statistical theory focuses on a specific, so-called "null hypothesis" (e.g., reaction time is not affected by background noise) which often represents "no treatment effect" or "no difference in treatment effects," usually in terms of some quantity being exactly equal to zero, as opposed to a more general, "alternative hypothesis" (e.g., reaction time changes as the level of background noise changes), which encompasses any amount of change other than zero. The next step in the cycle is to "Design an Experiment", followed by "Perform the Experiment", "Statistical Analysis" (formal and informal), and finally "Interpret and Report", which leads to possible modification of the "Current State of Knowledge".

Many parts of the "Design an Experiment" stage, as well as most parts of the "Statistical Analysis" and "Interpret and Report" stages, are common across many fields of science, while the other stages have many field-specific components. The focus of this book on the common stages is in no way meant to demean the importance of the other stages. You will learn the field-specific approaches in other courses, and the common topics here.

1.1 The importance of careful experimental design

Experimental design is a careful balancing of several features including "power", generalizability, various forms of "validity", practicality and cost. These concepts will be defined and discussed thoroughly in the next chapter. For now, you need to know that often an improvement in one of these features has a detrimental effect on other features. A thoughtful balancing of these features in advance will result in an experiment with the best chance of providing useful evidence to modify the current state of knowledge in a particular scientific field. On the other hand, it is unfortunate that many experiments are designed with avoidable flaws, and it is rarely the case that statistical analyses can make up for these flaws. This is an example of the old maxim "an ounce of prevention is worth a pound of cure". In fact, good experimental designs are chosen with certain statistical analyses in mind. Most statistical analyses are trustworthy only under certain assumptions, so we should ensure that those assumptions hold by design.

Our goal is always to actively design an experiment that has the best chance to produce meaningful, defensible evidence, rather than hoping that good statistical analysis may be able to correct for defects after the fact. Ideally, experiments are designed with particular statistical analyses in mind—i.e., experimental designs should complement statistical analyses, such that those analyses are more straightforward, interpretable, powerful, and convincing.

1.2 Overview of statistical analysis

Statistical analysis of experiments starts with graphical and non-graphical exploratory data analysis (EDA). EDA is useful for

- detection of mistakes
- checking of assumptions

- determining relationships among the explanatory variables
- assessing the direction and rough size of relationships between explanatory and outcome variables, and
- preliminary selection of appropriate models of the relationship between an outcome variable and one or more explanatory variables.

EDA always precedes formal (confirmatory) data analysis.

Most formal (confirmatory) statistical analyses are based on **models**. Statistical models are ideal, mathematical representations of observable characteristics. Models are best divided into two components. The structural component of the model (or **structural model**) specifies the relationships between explanatory variables and the mean (or other key feature) of the outcome variables. The "random" or "error" component of the model (or **error model**) characterizes the deviations of the individual observations from the mean. (Here, "error" does *not* indicate "mistake".) The two model components are also called "signal" and "noise" respectively. Statisticians realize that no mathematical models are perfect representations of the real world, but some are close enough to reality to be useful. A full description of a model should include all assumptions being made because statistical inference is impossible without assumptions, and sufficient deviation of reality from the assumptions will invalidate any statistical inferences.

A slightly different point of view says that models describe how the *distribution* of the outcome varies with changes in the explanatory variables.

Statistical models have both a structural component and a random component which describe means and the pattern of deviation from the mean, respectively.

A statistical test is always based on certain model assumptions about the population from which our sample comes. For example, a t-test includes the assumptions that the individual measurements are independent of each other, that the two groups being compared each have a Gaussian distribution, and that the standard

deviations of the groups are equal. The farther the truth is from these assumptions, the more likely it is that the t-test will give a misleading result. We will need to learn methods for assessing the truth of the assumptions, and we need to learn how "robust" each test is to violations of the assumptions, i.e., how far the assumptions can be "bent" before misleading conclusions are likely.

Understanding the assumptions behind every statistical analysis we learn is critical to judging whether or not the statistical conclusions are believable.

Statistical analyses can and should be framed and reported in different ways in different circumstances. But all statistical statements should at least include information about their level of uncertainty. The main reporting mechanisms you will learn about here are confidence intervals for unknown quantities and p-values and power estimates for specific hypotheses.

Here is an example of a situation where different ways of reporting give different amounts of useful information. Consider three different studies of the effects of a treatment on improvement on a memory test for which most people score between 60 and 80 points. First look at what we learn when the results are stated as 95% confidence intervals (full details of this concept are in later chapters) of $[-20, 40]$ points, $[-0.5, +0.5]$, and $[5, 7]$ points respectively. A statement that the first study showed a mean improvement of 10 points, the second of 0 points, and the third of 6 points (without accompanying information on uncertainty) is highly misleading! The third study lets us know that the treatment is almost certainly beneficial by a moderate amount, while from the first we conclude that the treatment may be quite strongly beneficial or strongly detrimental; we don't have enough information to draw a valid conclusion. And from the second study, we conclude that the effect is near zero. For these same three studies, the p-values might be, e.g., 0.35, 0.35 and 0.01 respectively. From just the p-values, we learn nothing about the magnitude or direction of any possible effects, and we cannot distinguish between the very different results of the first two studies. We only know that we have sufficient evidence to draw a conclusion that the effect is different from zero in the third study.

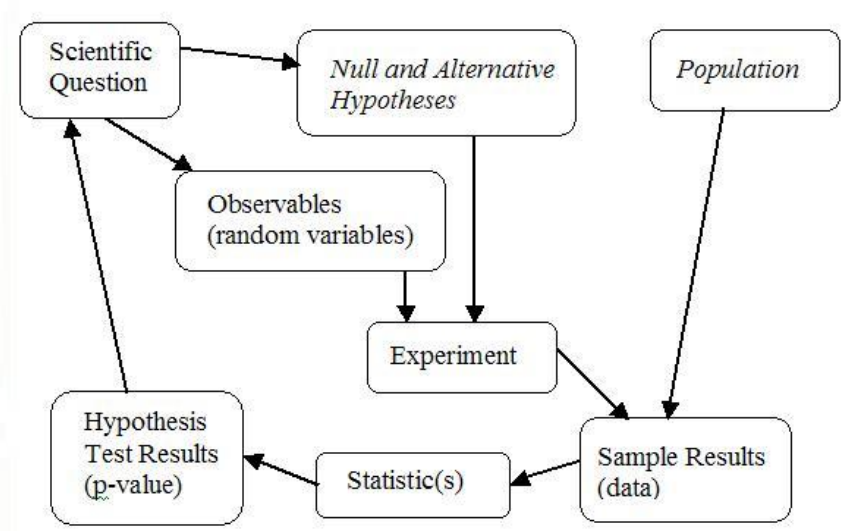


Figure 1.2: An oversimplified concept map.

p-values are not the only way to express inferential conclusions, and they are insufficient or even misleading in some cases.

1.3 What you should learn here

My expectation is that many of you, coming into the course, have a "concept-map" similar to Figure 1.2. This is typical of what students remember from a first course in statistics.

By the end of the book and course you should learn many things. You should be able to speak and write clearly using the appropriate technical language of statistics and experimental design. You should know the definitions of the key terms and understand the sometimes-subtle differences between the meanings of these terms in the context of experimental design and analysis as opposed to their meanings in ordinary speech. You should understand a host of concepts and their interrelationships. These concepts form a "concept-map" such as the one in Figure 1.3 that shows the relationships between many of the main concepts stressed in this course. The concepts and their relationships are the key to the practical use

of statistics in the social and other sciences. As a bonus to the creation of your own concept map, you will find that these maps will stick with you much longer than individual facts.

By actively working with data, you will gain intuition about how these concepts connect and what their implications are for real-world applications. This requires learning to use a specific statistical computer package. Many excellent packages exist and are suitable for this purpose. Examples here come from `R`, which has become ubiquitous in statistics, psychology, and other fields.

You should be able to design an experiment and discuss the choices that can be made and their competing positive and negative effects on the quality and feasibility of the experiment. You should know some of the pitfalls of carrying out experiments. It is critical to learn how to perform exploratory data analysis, assess data quality, and consider data transformations. You should also learn how to choose and perform the most common statistical analyses. And you should be able to assess whether the assumptions of the analysis are appropriate for the given data. You should know how to consider and compare alternative models. Finally, you should be able to interpret and report your results correctly so that you can assess how your experimental results may have changed the state of knowledge in your field.

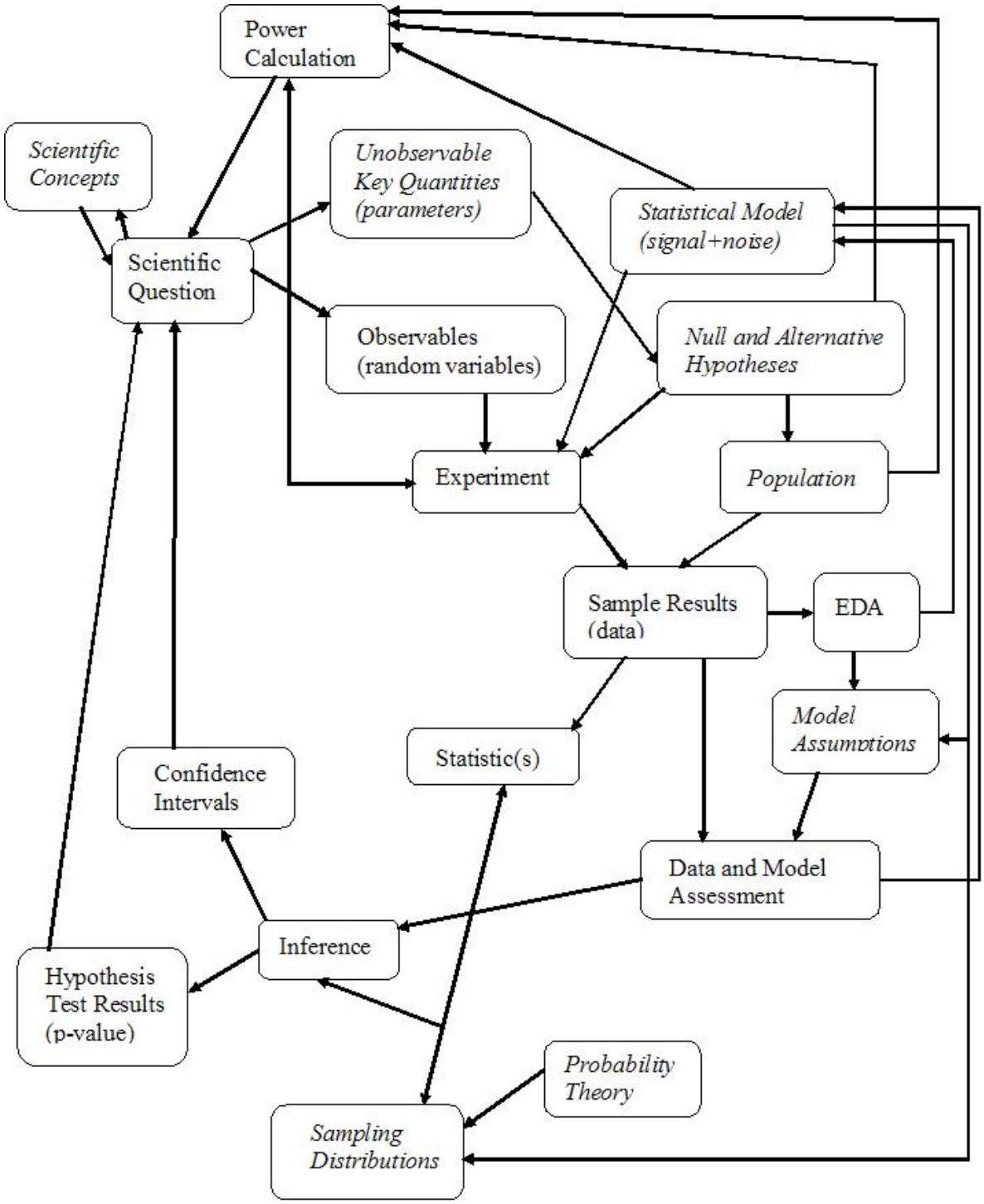


Figure 1.3: A reasonably complete concept map for this course.

Chapter 2

Defining and Classifying Data Variables

The link from scientific concepts to data quantities.

A key component of design of experiments is **operationalization**, which is the formal procedure that links scientific concepts to data collection. Operationalizations define **measures** or **variables** which are quantities of interest or which serve as the practical substitutes for the concepts of interest. For example, if you have a theory about what affects people's anger level, you need to operationalize the concept of anger. You might measure anger as the loudness of a person's voice in decibels, or some summary feature(s) of a spectral analysis of a recording of their voice, or where the person places a mark on a visual-analog "anger scale", or their total score on a brief questionnaire, etc. Each of these is an example of an operationalization of the concept of anger.

As another example, consider the concept of manual dexterity. You could devise a number of tests of dexterity, some of which might be "unidimensional" (producing one number) while others might be 'multidimensional"' (producing two or more numbers). Since your goal should be to convince both yourself and a wider audience that your final conclusions should be considered an important contribution to the body of knowledge in your field, you will need to make the choice carefully. Of course one of the first things you should do is investigate whether standard, acceptable measures already exist. Alternatively you may need to define your own measure(s) because no standard ones exist or because the

existing ones do not meet your needs (or perhaps because they are too expensive).

One more example is cholesterol measurement. Although this seems totally obvious and objective, there is a large literature on various factors that affect cholesterol, and enumerating some of these may help you understand the importance of very clear and detailed operationalization. Cholesterol may be measured as "total" cholesterol or various specific forms (e.g., HDL). It may be measured on whole blood, serum, or plasma, each of which gives somewhat different answers. It also varies with the time and quality of the last meal and the season of the year. Different analytic methods may also give different answers. All of these factors must be specified carefully to achieve the best measure.

In general, statistical analyses are only as useful as the types of variables and operationalizations that we feed into those analyses. If the variables at our disposal are not scientifically meaningful, then statistical analyses (regardless of how fancy they are) will be limited in the types of scientific questions they can answer. Thus, having a full understanding of the variables in our data is key to pinpointing the types of questions that can and cannot be answered from the data at hand.

2.1 What makes a "good" variable?

Regardless of what we are trying to measure, the qualities that make a good measure of a scientific concept are high reliability, absence of bias, low cost, practicality, objectivity, high acceptance, and high concept validity. **Reliability** is essentially the inverse of the statistical concept of variance, and a rough equivalent is "consistency". Statisticians also use the word "precision". In other words, the less variable (or more precise) the measure, the more reliable that measure is.

Bias refers to the difference between the measure and some "true" value. A difference between an *individual* measurement and the true value is called an "error" (which implies the practical impossibility of perfect precision, rather than the making of mistakes). The bias is the *average* difference over many measurements. Ideally the bias of a measurement process should be zero. For example, a measure of weight that is made with people wearing their street clothes and shoes has a positive bias equal to the average weight of the shoes and clothes across all subjects.

Precision or reliability refers to the reproducibility of repeated measurements, while bias refers to how far the average of many measurements is from the true value.

All other things being equal, when two measures are available, we will choose the less expensive and easier to obtain (more practical) measures. Measures that have a greater degree of subjectivity are generally less preferable. Although devising your own measures may improve upon existing measures, there may be a trade off with acceptability, resulting in reduced impact of your experiment on the field as a whole.

Construct validity is a key criterion for variable definition. Under ideal conditions, after completing your experiment you will be able to make a strong claim that changing your explanatory variable(s) in a certain way (e.g., doubling the amplitude of a background hum) causes a corresponding change in your outcome (e.g., score on an irritability scale). But if you want to convert that to meaningful statements about the effects of auditory environmental disturbances on the psychological trait or construct called "irritability", you must be able to argue that the scales have good construct validity for the traits, namely that the operationalization of background noise as an electronic hum has good construct validity for auditory environmental disturbances, and that your irritability scale really measures what people call irritability. Although construct validity is critical to the impact of your experimentation, its detailed understanding belongs separately to each field of study, and will not be discussed much in this book beyond the discussion in Chapter 3.

Construct validity is the link from practical measurements to meaningful concepts.

2.2 Classification by role

There are two different independent systems of classification of variables that you must learn in order to understand the rest of this book. The first system is based

on the role of the variable in the experiment and the analysis. The general terms used most frequently in this text are explanatory variables vs. outcome variables.

An experiment is designed to test the effects of some intervention on one or more measures, which are therefore designated as **outcome variables**. Much of this book deals with the most common type of experiment in which there is only a single outcome variable measured on each experimental unit (person, animal, factory, etc.) A synonym for outcome variable is dependent variable, often abbreviated DV.

The second main role a variable may play is that of an explanatory variable. **Explanatory variables** include variables purposely manipulated in an experiment (e.g., and mostly importantly, variables denoting different treatments) and variables that are not purposely manipulated, but are thought to possibly affect the outcome. Complete or partial synonyms include independent variable (IV), covariate, blocking factor, and predictor variable. Clearly, classification of the role of a variable is dependent on the specific experiment, and variables that are outcomes in one experiment may be explanatory variables in another experiment. For example, the score on a test of working memory may be the outcome variable in a study of the effects of an herbal tea on memory, but it is a possible explanatory factor in a study of the effects of different mnemonic techniques on learning calculus.

Most simple experiments have a single dependent or outcome variable plus one or more independent or explanatory variables.

In many studies, at least part of the interest is on how the effects of one explanatory variable on the outcome depends on the level of another explanatory variable. In statistics this phenomenon is called **interaction**. In some areas of science, the term **moderator variable** is used to describe the role of the secondary explanatory variable. For example, in the effects of the herbal tea on memory, the effect may be stronger in young people than older people, so age would be considered a moderator of the effect of tea on memory.

In more complex studies there may potentially be an intermediate variable in a causal chain of variables. If the chain is written A⇒B⇒C, then interest may focus on whether or not it is true that A can cause its effects on C only by changing B. If that is true, then we define the role of B as a mediator of the effect of A on C. An example is the effect of herbal tea on learning calculus. If this effect exists but

operates only through herbal tea improving working memory, which then allows better learning of calculus skills, then we would call working memory a **mediator** of the effect.

2.3 Classification by statistical type

A second classification of variables is by their statistical type. It is critical to understand the type of a variable for three reasons. First, it lets you know what type of information is being collected; second it defines (restricts) what types of statistical models are appropriate; and third, via those statistical model restrictions, it helps you choose what analysis is appropriate for your data.

Students often have difficulty knowing "which statistical test to use". The answer to that question always starts with variable classification:

Classifying variables by their roles and by their statistical types are the first two and the most important steps to choosing a correct analysis for an experiment.

There are two main types of variables, each of which has two subtypes according to this classification system:

Quantitative Variables
- **Discrete Variables**
- **Continuous Variables**

Categorical Variables
- **Nominal Variables**
- **Ordinal Variables**

Both categorical and quantitative variables are often recorded as numbers, so this is not a reliable guide to the major distinction between categorical and quantitative variables. **Quantitative variables** are those for which the recorded numbers encode magnitude information based on a true quantitative scale. The best way to check if a measure is quantitative is to use the **subtraction test**. If two experimental units (e.g., two people) have different values for a particular measure,

then you should subtract the two values, and ask yourself about the meaning of the difference. If the difference can be interpreted as a *quantitative* measure of difference between the subjects, and if the meaning of each quantitative difference is the same for any pair of values with the same difference (e.g., 1 vs. 3 and 10 vs. 12), then this is a quantitative variable. Otherwise, it is a categorical variable.

For example, if the measure is age of the subjects in years, then for all of the pairs 15 vs. 20, 27 vs. 33, 62 vs. 67, etc., the difference of 5 indicates that the subject in the pair with the large value has lived 5 more years than the subject with the smaller value, and this is a quantitative variable. Other examples that meet the subtraction test for quantitative variables are age in months or seconds, weight in pounds or ounces or grams, length of index finger, number of jelly beans eaten in 5 minutes, number of siblings, and number of correct answers on an exam.

Examples that fail the subtraction test, and are therefore categorical, not quantitative, are eye color coded 1=blue, 2=brown, 3=gray, 4=green, 5=other; race where 1=Asian, 2=Black, 3=Caucasian, 4=Other; grade on an exam coded 4=A, 3=B, 2=C, 1=D, 0=F; type of car where 1=SUV, 2=sedan, 3=compact and 4=subcompact; and severity of burn where 1=first degree, 2=second degree, and 3=third degree. While the examples of eye color and race would only fool the most careless observer into incorrectly calling them quantitative, the latter three examples are trickier. For the coded letter grades, the average difference between an A and a B may be 5 correct questions, while the average difference between a B and a C may be 10 correct questions, so this is not a quantitative variable. (On the other hand, if we call the variable quality points, as is used in determining grade point average, it can be used as a quantitative variable.) Similar arguments apply for the car type and burn severity examples, e.g., the size or weight difference between SUV and sedan is not the same as between compact and subcompact. (These three variables are discussed further below.)

Once you have determined that a variable is quantitative, it is often worthwhile to further classify it into discrete (also called counting) vs. continuous. Here the test is the **midway test**. If, for *every* pair of values of a quantitative variable the value midway between them is a meaningful value, then the variable is **continuous**, otherwise it is **discrete**. Typically discrete variables can only take on whole numbers (but all whole numbered variables are *not* necessarily discrete). For example, age in years is continuous because midway between 21 and 22 is 21.5 which is a meaningful age, even if we operationalized age to be age at the last birthday or age at the nearest birthday.

Other examples of continuous variables include weights, lengths, areas, times, and speeds of various kinds. Examples of discrete variables include number of jelly beans eaten, number of siblings, number of correct questions on an exam, and number of incorrect turns a rat makes in a maze. For none of these does an answer of, say, $3\frac{1}{2}$, make sense.

There are examples of quantitative variables that are not clearly categorized as either discrete or continuous. These generally have many possible values and strictly fail the midpoint test, but are practically considered to be continuous because they are well approximated by continuous probability distributions. One fairly silly example is mass; while we know that you can't have half of a molecule, for all practical purposes we can have a mass half-way between any two masses of practical size, and no one would even think of calling mass discrete. Another example is the ratio of teeth to forelimb digits across many species; while only certain possible values actually occur and many midpoints may not occur, it is practical to consider this to be a continuous variable. One more example is the total score on a questionnaire which is comprised of, say, 20 questions each with a score of 0 to 5 as whole numbers. The total score is a whole number between 0 and 100, and technically is discrete, but it may be more practical to treat it as a continuous variable.

It is worth noting here that as a practical matter most models and analyses do not distinguish between discrete and continuous *explanatory* variables, while many do distinguish between discrete and continuous quantitative *outcome* variables.

Measurements with meaningful magnitudes are called quantitative. They may be discrete (only whole number counts are valid) or continuous (fractions are at least theoretically meaningful).

Categorical variables simply place explanatory or outcome variable characteristics into (non-quantitative) categories. The different values taken on by a categorical variable are often called **levels**. If the levels simply have arbitrary names then the variable is **nominal**. But if there are at least three levels, and if every reasonable person would place those levels in the same (or the exact reverse) order, then the variable is **ordinal**. The above examples of eye color and race are nominal categorical variables. Other nominal variables include car make or model, political party, gender, and personality type. The above examples of exam grade,

car type, and burn severity are ordinal categorical variables. Other examples of ordinal variables include liberal vs. moderate vs. conservative for voters or political parties; severe vs. moderate vs. mild vs. no itching after application of a skin irritant; and disagree vs. neutral vs. agree on a policy question.

It may help to understand ordinal variables better if you realize that most ordinal variables, at least theoretically, have an underlying quantitative variable. Then the ordinal variable is created (explicitly or implicitly) by choosing "cut-points" of the quantitative variable between which the ordinal categories are defined. Also, in some sense, creation of ordinal variables is a kind of "super-rounding", often with different spans of the underlying quantitative variable for the different categories. See Figure 2.1 for an example based on the old IQ categorizations. Note that the categories have different widths and are quite wide (more than one would typically create by just rounding).

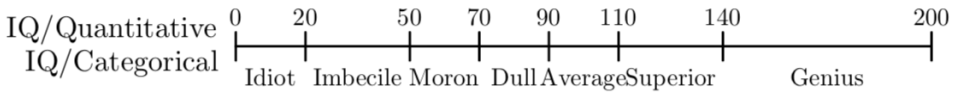


Figure 2.1: Old IQ categorization

It is worth noting here that the best-known statistical tests for categorical *outcomes* do not take the ordering of ordinal variables into account, although there certainly are good tests that do so. On the other hand, when used as *explanatory variables* in most statistical tests, ordinal variables are usually either "demoted" to nominal or "promoted" to quantitative.

2.4 Tricky cases

When categorizing variables, most cases are clear-cut, but some may not be. If the data are recorded directly as categories rather than numbers, then you only need to apply the "reasonable person's order" test to distinguish nominal from ordinal. If the results are recorded as numbers, apply the subtraction test to distinguish quantitative from categorical. When trying to distinguish discrete quantitative from continuous quantitative variables, apply the midway test and ignore the degree of rounding.

An additional characteristic that is worth paying attention to for quantitative variables is the range, i.e., the minimum and maximum possible values. Variables that are limited to between 0 and 1 or 0% and 100% often need special consideration, as do variables that have other arbitrary limits.

When a variable meets the definition of quantitative, but it is an explanatory variable for which only two or three levels are being used, it is usually better to treat this variable as categorical.

Finally we should note that there is an additional type of variable called an "order statistic" or "rank" which counts the placement of a variable in an ordered list of all observed values, and while strictly an ordinal categorical variable, is often treated differently in statistical procedures.

Chapter 3

Review of Probability

A review of the portions of probability useful for understanding experimental design and analysis.

The material in this section is intended as a review of the topic of probability as covered in the prerequisite course (36-201 at CMU). The material in gray boxes is beyond what you may have previously learned, but may help the more mathematically minded reader to get a deeper understanding of the topic. You need not memorize any formulas or even have a firm understanding of this material at the start of the class. But I do recommend that you at least skim through the material early in the semester. Later, you can use this chapter to review concepts that arise as the class progresses.

For the earliest course material, you should have a basic idea of what a random variable and a probability distribution are, and how a probability distribution defines event probabilities. You also need to have an understanding of the concepts of parameter, population, mean, variance, standard deviation, and correlation.

3.1 Definition(s) of probability

We could choose one of several technical definitions for **probability**, but for our purposes it refers to an assessment of the likelihood of the various possible outcomes in an experiment or some other situation with a "random" outcome.

Note that in probability theory the term "outcome" is used in a more general

sense than the outcome vs. explanatory variable terminology that is used in the rest of this book. In probability theory the term "outcome" applies not only to the "outcome variables" of experiments but also to "explanatory variables" if their values are not fixed. For example, the dose of a drug is normally fixed by the experimenter, so it is not an outcome in probability theory, but the age of a randomly chosen subject, even if it serves as an explanatory variable in an experiment, is not "fixed" by the experimenter, and thus can be an "outcome" under probability theory.

The collection of all possible outcomes of a particular random experiment (or other well defined random situation) is called the **sample space**, usually abbreviated as **S** or Ω (omega). The outcomes in this set (list) must be exhaustive (cover all possible outcomes) and mutually exclusive (non-overlapping), and should be as simple as possible.

For a simple example consider an experiment consisting of the tossing of a six sided die. One possible outcome is that the die lands with the side with one dot facing up. I will abbreviate this outcome as 1du (one dot up), and use similar abbreviations for the other five possible outcomes (assuming it can't land on an edge or corner). Now the sample space is the set {1du, 2du, 3du, 4du, 5du, 6du}. We use the term **event** to represent any subset of the sample space. For example {1du}, {1du, 5du}, and {1du, 3du, 5du}, are three possible events, and most people would call the third event "odd side up". One way to think about events is that they can be defined before the experiment is carried out, and they either occur or do not occur when the experiment is carried out. In probability theory we learn to compute the chance that events like "odd side up" will occur based on assumptions about things like the probabilities of the elementary outcomes in the sample space.

Note that the "true" outcome of most experiments is not a number, but a physical situation, e.g., "3 dots up" or "the subject chose the blue toy". For convenience sake, we often "map" the physical outcomes of an experiment to integers or real numbers, e.g., instead of referring to the outcomes 1du to 6du, we can refer to the numbers 1 to 6. Technically, this mapping is called a **random variable**, but more commonly and informally we refer to the unknown numeric outcome itself (before the experiment is run) as a "random variable". Random variables commonly are represented as upper case English letters towards the end of the alphabet, such as Z, Y or X. Sometimes the lower case equivalents are used to represent the actual outcomes after the experiment is run. For example, if Z denotes a binary random

variable in an experiment, there is some probability that $Z = 1$ and some probability that $Z = 0$; meanwhile, z denotes a particular value or "realization" of Z after the experiment is run (i.e., z is just a number—in this case, 0 or 1—it isn't random).

Random variables are maps from the sample space to the real numbers, but they need not be one-to-one maps. For example, in the die experiment we could map all of the outcomes in the set {1du, 3du, 5du} to the number 0 and all of the outcomes in the set {2du, 4du, 6du} to the number 1, and call this random variable Y. If we call the random variable that maps to 1 through 6 as X, then random variable Y could also be thought of as a map from X to Y where the odd numbers of X map to 0 in Y and the even numbers to 1. Often the term **transformation** is used when we create a new random variable out of an old one in this way. It should now be obvious that many, many different random variables can be defined/invented for a given experiment.

A few more basic definitions are worth learning at this point. A random variable that takes on only the numbers 0 and 1 is commonly referred to as an **indicator (random) variable**. It is usually named to match the set that corresponds to the number 1. So in the previous example, random variable Y is an indicator for even outcomes. For any random variable, the term **support** is used to refer to the set of possible real numbers defined by the mapping from the physical experimental outcomes to the numbers. Therefore, for random variables we use the term "event" to represent any subset of the support.

Ignoring certain technical issues, probability theory is used to take a basic set of assigned (or assumed) probabilities and use those probabilities (possibly with additional assumptions about something called independence) to compute the probabilities of various more complex events.

The core of probability theory is about measuring the chances of occurrence of events based on a set of assumptions about the underlying probability processes.

One way to think about probability is that it quantifies how much we can know when we cannot know something exactly. Probability theory is deductive, in the sense that it involves making assumptions about a random (not completely predictable) process, and then deriving valid statements about what is likely to

happen based on mathematical principles. For this course a fairly small number of probability definitions, concepts, and skills will suffice.

For those students who are unsatisfied with the loose definition of probability above, here is a brief descriptions of three different approaches to probability, although it is not necessary to understand this material to continue through the chapter. If you want even more detail, I recommend *Comparative Statistical Inference* by Vic Barnett.

Valid probability statements do not claim what events will happen, but rather which are likely to happen. The starting point is sometimes a judgment that certain events are a priori equally likely. Then using only the additional assumption that the occurrence of one event has no bearing on the occurrence of another separate event (called the assumption of independence), the likelihood of various complex combinations of events can be worked out through logic and mathematics. This approach has logical consistency, but cannot be applied to situations where it is unreasonable to assume equally likely outcomes and independence.

A second approach to probability is to define the probability of an outcome as the limit of the long-term fraction of times that outcome occurs in an ever-larger number of independent trials. This allows us to work with basic events that are not equally likely, but has a disadvantage that probabilities are assigned through observation. Nevertheless this approach is sufficient for our purposes, which are mostly to figure out what would happen if certain probabilities are assigned to some events.

A third approach is subjective probability, where the probabilities of various events are our subjective (but consistent) assignments of probability. This has the advantage that events that only occur once, such as the next presidential election, can be studied probabilistically. Despite the seemingly bizarre premise, this is a valid and useful approach which may give different answers for different people who have different beliefs, but still helps calculate your rational but personal probability of future uncertain events, given your prior beliefs.

Regardless of which definition of probability you use, the calculations we need

are basically the same. First we need to note that probability applies to some well-defined unknown or future situation in which some outcome will occur, the list of possible outcomes is well defined, and the exact outcome is unknown. If the outcome is categorical or discrete quantitative (see Section 2.3), then each possible outcome gets a probability in the form of a number between 0 and 1 such that the sum of all of the probabilities is 1. This indicates that impossible outcomes are assigned probability zero, but assigning a probability zero to an event does not necessarily mean that that outcome is impossible (see below). (Note that a probability is technically written as a number from 0 to 1, but is often converted to a percent from 0% to 100%. In case you have forgotten, to convert to a percent multiply by 100, e.g., 0.25 is 25% and 0.5 is 50% and 0.975 is 97.5%.)

Every valid probability must be a number between 0 and 1 (or a percent between 0% and 100%).

We will need to distinguish two types of random variables. Discrete random variables correspond to the categorical variables plus the discrete quantitative variables of Chapter 2. Their support is a (finite or infinite) list of numeric outcomes, each of which has a non-zero probability. (Here we will loosely use the term "support" not only for the numeric outcomes of the random variable mapping, but also for the sample space when we do not explicitly map an outcome to a number.) Examples of discrete random variables include the result of a coin toss (the support using curly brace set notation is {H,T}), the number of tosses out of 5 that are heads ($\{0, 1, 2, 3, 4, 5\}$), the color of a random person's eyes ({blue, brown, green, other}), and the number of coin tosses until a head is obtained ($\{1, 2, 3, 4, 5, \ldots\}$). Note that the last example has an infinite sized support.

Continuous random variables correspond to the continuous quantitative variables of Chapter 2. Their support is a continuous range of real numbers (or rarely several disconnected ranges) with no gaps. When working with continuous random variables in probability theory we think as if there is no rounding, and each value has an infinite number of decimal places. In practice we can only measure things to a certain number of decimal places; for example, actual measurement of the continuous variable "length" might be 3.14, 3.15, etc., which does have gaps. But we approximate this with a continuous random variable rather than a discrete random variable because more precise measurement is possible in theory.

A strange aspect of working with continuous random variables is that each particular outcome in the support has probability zero, while none is actually impossible. The reason each outcome value has probability zero is that otherwise the probabilities of all of the events would add up to more than 1. So for continuous random variables we usually work with intervals of outcomes to say, e.g, that the probability that an outcome is between 3.14 and 3.15 might be 0.02 while each real number in that range, e.g., π (exactly), has zero probability. Examples of continuous random variables include ages, times, weights, lengths, etc. All of these can theoretically be measured to an infinite number of decimal places.

It is also possible for a random variable to be a mixture of discrete and continuous random variables, e.g., if an experiment is to flip a coin and report 0 if it is heads and the time it was in the air if it is tails, then this variable is a mixture of the discrete and continuous types because the outcome "0" has a non-zero (positive) probability, while all positive numbers have a zero probability (though intervals between two positive numbers would have probability greater than zero.)

3.2 Probability mass functions and density functions

.

A **probability mass function** (pmf) is just a full description of the possible outcomes and their probabilities for some discrete random variable. In some situations it is written in simple list form, e.g.,

$$f(x) = \begin{cases} 0.25 & \text{if} \quad x = 1 \\ 0.35 & \text{if} \quad x = 2 \\ 0.40 & \text{if} \quad x = 3 \end{cases}$$

where f(x) is the probability that random variable X takes on value x, with f(x)=0 implied for all other x values. We can see that this is a valid probability distribution

because each probability is between 0 and 1 and the sum of all of the probabilities is 1.00. In other cases we can use a formula for f(x), e.g.

$$f(x) = \left(\frac{4!}{(4-x)!\ x!} \right) p^x (1-p)^{4-x} \text{ for } x = 0, 1, 2, 3, 4$$

which is the so-called binomial distribution with parameters 4 and p.

It is not necessary to understand the mathematics of this formula for this course, but if you want to try you will need to know that the exclamation mark symbol is pronounced "factorial" and $r!$ represents the product of all the integers from 1 to r. As an exception, $0! = 1$.

This particular pmf represents the probability distribution for getting x "successes" out of 4 "trials" when each trial has a success probability of p independently. This formula is a shortcut for the five different possible outcome values. If you prefer, you can calculate the five different probabilities and use the first form for the pmf (i.e., compute $f(x)$ for $x = 0$, $x = 1$, $x = 2$, $x = 3$, and $x = 4$, and write it out in list form).

Another example is the so-called geometric distribution, which represents the outcome for an experiment in which we count the number of independent trials until the first success is seen. The pmf is:

$$f(x) = p(1-p)^{x-1} \text{ for } x = 1, 2, 3, \ldots$$

and it can be shown that this is a valid distribution with the sum of this infinitely long series equal to 1.00 for any value of p between 0 and 1. This pmf cannot be written in list form (or rather, if you tried to write it in list form, it would take you literally an infinite amount of time!)

By definition a random variable takes on numeric values (i.e., it maps real experimental outcomes to numbers). Therefore it is easy and natural to think about the pmf of any discrete quantitative experimental variable, whether it is explanatory or outcome. For categorical experimental variables, we do not need to assign numbers to the categories, but we always can do that, and then it is easy to consider that variable as a random variable with a finite pmf. Of course, for nominal categorical variables the order of the assigned numbers is meaningless, and for ordinal categorical variables it is most convenient to use consecutive integers for the assigned numeric values.

Probability mass functions apply to discrete outcomes. A pmf is just a list (usually written as a formula) of all possible outcomes for a given random variable and the probabilities for each outcome.

For continuous random variables, we use a somewhat different method for summarizing all of the information in a probability distribution. This is the **probability density function** (pdf), usually represented as "f(x)", which does not represent probabilities directly but from which the probability that the outcome falls in a certain range can be calculated using integration from calculus. (If you don't remember integration from calculus, don't worry, it is OK to skip over the details.)

One of the simplest pdf's is that of the uniform distribution, where all real numbers between a and b are equally likely and numbers less than a or greater than b are impossible. The pdf is:

$$f(x) = 1/(b-a) \text{ for } a \leq x \leq b$$

The general probability formula for any continuous random variable is

$$\Pr(t \leq X \leq u) = \int_t^u f(x)dx.$$

In this formula $\int \cdot \, dx$ means that we must use calculus to carry out integration.

Note that we use capital X for the random variable in the probability statement because this refers to the potential outcome of an experiment that has not yet been conducted, while the formulas for pdf and pmf use lower case x because they represent calculations done for each of several possible outcomes of the experiment. Also note that, in the pdf *but not* the pmf, we could replace either or both $\leq$ signs with $<$ signs because the probability that the outcome is *exactly* equal to t or u (to an infinite number of decimal places) is zero.

So for the continuous uniform distribution, for any $a \leq t \leq u \leq b$,

$$\Pr(t \leq X \leq u) = \int_t^u \frac{1}{b-a} dx = \frac{u-t}{b-a}.$$

You can check that this always gives a number between 0 and 1, and the probability of any individual outcome (where u=t) is zero, while the probability that the outcome is some number between a and b is 1 (u=a, t=b). You can also see that, e.g., the probability that X is in the middle third of the interval from a to b is $\frac{1}{3}$, etc.

Of course, there are many interesting and useful continuous distributions other than the continuous uniform distribution. Some other examples are given below. Each is fully characterized by its probability density function.

3.2.1 Reading a pdf

In general, we often look at a plot of the probability density function, f(x), vs. the possible outcome values, x. This plot is high in the regions of likely outcomes and low in less likely regions. The well-known standard Gaussian distribution (see 3.2) has a bell-shaped graph centered at zero with about two thirds of its area between x = -1 and x = +1 and about 95% between x = -2 and x = +2. But a pdf can have many different shapes.

It is worth understanding that many pdf's come in "families" of similarly shaped curves. These various curves are named or "indexed" by one or more numbers called parameters (but there are other uses of the term parameter; see Section 3.5). For example that family of Gaussian (also called Normal) distributions is indexed by the mean and variance (or standard deviation) of the distribution. The t-distributions, which are all centered at 0, are indexed by a single parameter called the degrees of freedom. The chi-square family of distributions is also indexed by a single degree of freedom value. The F distributions are indexed by two degrees of freedom numbers designated numerator and denominator degrees of freedom.

In this course we will not do any integration. We will use tables or a computer

program to calculate probabilities for continuous random variables. We don't even need to know the formula of the pdf because the most commonly used formulas are known to the computer by name. Sometimes we will need to specify degrees of freedom or other parameters so that the computer will know which pdf of a family of pdf's to use.

Despite our heavy reliance on the computer, getting a feel for the idea of a probability density function is critical to the level of understanding of data analysis and interpretation required in this course. At a minimum you should realize that a pdf is a curve with outcome values on the horizontal axis and the vertical height of the curve tells which values are likely and which are not. The total area under the curve is 1.0, and the area under the curve between any two "x" values is the probability that the outcome will fall between those values.

For continuous random variables, we calculate the probability that the outcome falls in some interval, not that the outcome exactly equals some value. This calculation is normally done by a computer program which uses integral calculus on a "probability density function."

3.3 Probability calculations

This section reviews the most basic probability calculations. It is worthwhile, but not essential to become familiar with these calculations. For many readers, the boxed material may be sufficient. You won't need to memorize any of these formulas for this course.

Remember that in probability theory we don't worry about where probability assignments (a pmf or pdf) come from. Instead we are concerned with how to calculate other probabilities given the assigned probabilities. Let's start with calculation of the probability of a "complex" or "compound" event that is constructed from the simple events of a discrete random variable.

For example, if we have a discrete random variable that is the number of correct answers that a student gets on a test of 5 questions, i.e. integers in the set $\{0, 1, 2, 3, 4, 5\}$, then we could be interested in the probability that the student gets an even number of questions correct, or less than 2, or more than 3, or between

Event	Probability	Calculation
T=0	0.10	Assigned
T=1	0.26	Assigned
T=2	0.14	Assigned
T=3	0.21	Assigned
T=4	0.24	Assigned
T=5	0.05	Assigned
T$\in \{0, 2, 4\}$	0.48	0.10+0.14+0.24
T<2	0.36	0.10+0.26
T$\leq$2	0.50	0.10+0.26+0.14
T$\leq$4	0.29	0.24+0.05
T$\geq$0	1.00	0.10+0.26+0.14+0.21+0.24+0.05

Table 3.1: Disjoint Addition Rule

3 and 4, etc. All of these probabilities are for outcomes that are subsets of the sample space of all 6 possible "elementary" outcomes, and all of these subsets are the union (joining together) of some of the 6 possible "elementary" outcomes. In the case of any complex outcome that can be written as the union of some other disjoint (non-overlapping) outcomes, the probability of the complex outcome is the sum of the probabilities of the disjoint outcomes. To complete this example look at Table 3.1 which shows assigned probabilities for the elementary outcomes of the random variable we will call T (the test outcome) and for several complex events.

You should think of the probability of a complex event such as T<2, usually written as Pr(T<2) or P(T<2), as being the chance that, when we carry out a random experiment (e.g., test a student), the outcome will be any one of the outcomes in the defined set (0 or 1 in this case). Note that (implicitly) outcomes not mentioned are impossible, e.g., Pr(T=17) = 0. Also something must happen: Pr(T$\geq$0) = 1.00 or Pr(T $\in \{0, 1, 2, 3, 4, 5\}$) = 1.00. It is also true that the probability that nothing happens is zero: Pr($T \in \phi$) = 0, where ϕ means the "empty set".

Calculate the probability that any of several non-overlapping events occur in a single experiment by adding the probabilities of the individual events.

The addition rule shows us how to calculate the probability of non-overlapping events, but what if the events are overlapping? Using the above 5-question test example, we consider defining event E as the set $\{T : 1 \leq T \leq 3\}$ read as all values of outcome T such that 1 is less than or equal to T and T is less than or equal to 3. Of course E = $\{1, 2, 3\}$. Now define F = $\{T : 2 \leq T \leq 4\}$ or F = $\{2, 3, 4\}$. The union of these sets, written $E \cup F$ is equal to the set of outcomes $\{1, 2, 3, 4\}$. To find Pr($E \cup F$) we could try adding Pr(E) + Pr(F), but we would be double counting the elementary events in common to the two sets, namely $\{2\}$ and $\{3\}$, so the correct solution is to add first, and then subtract for the double counting. We define the intersection of two sets as the elements that they have in common, and use notation like $E \cap F = \{2, 3\}$ or, in situations where there is no chance of confusion, just $EF = \{2, 3\}$. Then the rule for the probability of the union of two sets is:

$$\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(E \cap F).$$

For our example, Pr(E F) = 0.61 + 0.59 - 0.35 = 0.85, which matches the direct calculation Pr($\{1, 2, 3, 4\}$) = 0.26 + 0.14 + 0.21 + 0.24. It is worth pointing out again that if we get a result for a probability that is not between 0 and 1, we are sure that we have made a mistake!

Note that it is fairly obvious that Pr $A \cap B$ = Pr $B \cap A$ because $A \cap B = B \cap A$, i.e., the two events are equivalent sets. Also note that there is a complicated general formula for the probability of the union of three or more events, but you can just apply the two event formula, above, multiple times to get the same answer.

If two events overlap, calculate the probability that either event occurs as the sum of the individual event probabilities minus the probability of the overlap.

Another useful rule is based on the idea that something in the sample space must happen and on the definition of the complement of a set. The complement of a set, say E, is written E^c and is a set made of all of the elements of the sample space that are not in set E. Using the set E above, $E^c = \{0, 4, 5\}$. The rule is:

$$\Pr(E^c) = 1 - \Pr(E).$$

In our example, $\Pr\{0,4,5\} = 1 - \Pr\{1,2,3\} = 1 - 0.61 = 0.39$.

Calculate the probability that an event will *not* occur as 1 minus the probability that it will occur.

Another important concept is **conditional probability**. At its core, conditional probability means reducing the pertinent sample space. For instance we might want to calculate the probability that a random student gets an odd number of questions correct while ignoring those students who score over 4 points. This is usually described as finding the probability of an odd number given $T \leq 4$. The notation is Pr(T is odd$|T \leq 4$) , where the vertical bar is pronounced "given". (The word "given" in a probability statement is usually a clue that conditional probability is being used.) For this example we are excluding the 5% of students who score a perfect 5 on the test. Our new sample space must be "renormalized" so that its probabilities add up to 100%. We can do this by replacing each probability by the old probability divided by the probability of the reduced sample space, which in this case is (1-0.05)=0.95. Because the old probabilities of the elementary outcomes in the new set of interest, $\{0,1,2,3,4\}$, add up to 0.95, if we divide each by 0.95 (making it bigger), we get a new set of 5 (instead of 6) probabilities that add up to 1.00. We can then use these new probabilities to find that the probability of interest is $0.26/0.95 + 0.21/0.95 = 0.495$.

Or we can use a new probability rule:

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

In our current example, we have

$$\begin{aligned} \Pr\left(T \in \{1,3,5\} | T \leq 4\right) &= \frac{\Pr(T \in \{1,3,5\} \cap T \leq 4)}{\Pr(T \leq 4)} \\ &= \frac{\Pr(T) \in \{1,3\}}{1 - \Pr(T = 5)} = \frac{0.26 + 0.21}{0.95} = 0.495 \end{aligned}$$

If we have partial knowledge of an outcome or are only interested in some selected outcomes, the appropriate calculations require use of the conditional probability formulas, which are based on using a new, smaller sample space.

The next set of probability concepts relates to **independence** of events. (Sometimes students confuse disjoint and independent; be sure to keep these concepts separate.) Two events, say E and F, are independent if the probability that event E happens, Pr(E), is the same whether or not we condition on event F happening. That is $\Pr(E) = \Pr(E|F)$. If this is true then it is also true that $\Pr(F) = \Pr(F|E)$. We use the term **marginal probability** to distinguish a probability like Pr(E) that is not conditional on some other probability. The marginal probability of E is the probability of E *ignoring* the outcome of F (or any other event). The main idea behind independence and its definition is that knowledge of whether or not F occurred does not change what we know about whether or not E will occur. It is in this sense that they are independent of each other.

Note that independence of E and F also means that Pr(E∩F) = Pr(E)Pr(F), i.e., the probability that two independent events both occur is the product of the individual (marginal) probabilities.

Continuing with our five-question test example, let event A be the event that the test score, T, is greater than or equal to 3, i.e., A=$\{3, 4, 5\}$, and let B be the event that T is even. Using the union rule (for disjoint elements or sets) Pr(A) = 0.21 + 0.24 + 0.05 = 0.50, and Pr(B) = 0.10 + 0.14 + 0.24 = 0.48. From the conditional probability formula

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(T = 4)}{\Pr(B)} = \frac{0.24}{0.48} = 0.50$$

and

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(T = 4)}{\Pr(A)} = \frac{0.24}{0.50} = 0.48.$$

Since $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$, events A and B are independent. We therefore can calculate that Pr(AB) = Pr(T=4) = Pr(A) Pr(B) = 0.50 (0.48) = 0.24 (which we happened to already know in this example).

If A and B are independent events, then we can calculate the probability of their intersection as the product of the marginal probabilities. If they are not independent, then we can calculate the probability of the intersection from an equation that is a rearrangement of the conditional probability formula:

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) \text{ or } \Pr(A \cap B) = \Pr(B|A)\Pr(A).$$

For our example, one calculation we can make is

$$\begin{aligned} \Pr(\text{T is even} \cap T < 2) &= \Pr(\text{T is even}|T < 2)\Pr(T < 2) \\ &= [0.10/(0.10 + 0.26)] \cdot (0.10 + 0.26) = 0.10. \end{aligned}$$

Although this is not the easiest way to calculate $\Pr(\text{T is even}|T < 2)$ for this problem, the small bag of tricks described in this chapter can come in handy for making certain calculations when only certain pieces of information are conveniently obtained.

A contrasting example is to define event G=$\{0, 2, 4\}$, and let H=$\{2, 3, 4\}$. Then G$\cap$H=$\{2, 4\}$. We can see that Pr(G)=0.48 and Pr(H)=0.59 and $\Pr(G \cap H)$=0.38. From the conditional probability formula

$$\Pr(G|H) = \frac{\Pr(G \cap H)}{\Pr(H)} = \frac{0.38}{0.59} = 0.644.$$

So, if we have no knowledge of the random outcome, we should say there is a 48% chance that T is even. But if we have the partial outcome that T is between 2 and 4 inclusive, then we revise our probability estimate to a 64.4% chance that T is even. Because these probabilities differ, we can say that event G is *not* independent of event H. We can "check" our conclusion by verifying that the probability of G$\cap$H (0.38) is *not* the product of the marginal probabilities, $0.48 \cdot 0.59 = 0.2832$.

Independence also applies to random variables. Two random variables are independent if knowledge of the outcome of one does not change the (conditional) probability of the other. In technical terms, if $\Pr(X|Y = y) = \Pr(X)$ for all values of y, then X and Y are independent random variables. If two random variables are independent, and if you consider any event that is a subset of the X outcomes and any other event that is a subset of the Y outcomes, these events will be independent.

At an intuitive level, events are independent if knowledge that one event has or has not occurred does not provide new information about the probability of the other event. Random variables are independent if knowledge of the outcome of one does not provide new information about the probabilities of the various outcomes of the other. In most experiments it is reasonable to assume that the outcome for any one subject is independent of the outcome of any other subject. If two events are independent, the probability that both occur is the product of the individual probabilities.

3.4 Populations and samples

In the context of experiments, observational studies, and surveys, we make our actual measurements on individual **observational units** . These are commonly people (subjects, participants, etc.) in the social sciences, but can also be schools, social groups, economic entities, archaeological sites, etc. (In some complicated situations we may make measurements at multiple levels, e.g., school size and students' test scores, which makes the definition of experimental units more complex.)

We use the term **population** to refer to the entire set of actual or potential observational units. So for a study of working memory, we might define the population as all U.S. adults, as all past present and future human adults, or we can use some other definition. In the case of, say, the U.S. census, the population is reasonably well defined (although there are problems, referred to in the census literature as "undercount") and is large, but finite. For experiments, the definition of population is often not clearly defined, although such a definition can be very important. See Section 7.2 for more details. Often we consider such a population to be theoretically infinite, with no practical upper limit on the number of potential subjects we could test.

For most studies (other than a census), only a subset of all of the possible experimental units of the population are actually selected for study, and this is called the **sample** (not to be confused with sample space). An important part of the understanding of the idea of a sample is to realize that each experiment is conducted on a particular sample, but hypothetically could have been conducted on many other different samples. For theoretically correct inference, the sample

should be randomly selected from the population. If this is not true, we call the sample a **convenience sample**, and we lose many of the theoretical properties required for correct inference.

Even though we often need to use samples in science, it is very important to remember that we are usually interested in learning about populations, not samples. Inference from samples to populations is the goal of most statistical analyses.

3.5 Parameters describing distributions

As mentioned above, the probability distribution of a random variable (pmf for a discrete random variable or pdf for a continuous random variable) completely describes its behavior in terms of the chances that various events will occur. It is also useful to work with certain fixed quantities that either completely characterize a distribution within a family of distributions or otherwise convey useful information about a distribution. These are called **parameters**. Parameters are fixed quantities that characterize theoretical probability distributions. (I am using the term "theoretical distribution" to focus on the fact that we are assuming a particular mathematical form for the pmf or pdf.)

The term parameter may be somewhat confusing because it is used in several slightly different ways. Parameters may refer to the fixed constants that appear in a pdf or pmf. Note that these are somewhat arbitrary because the pdf or pmf may often be rewritten (technically, re-parameterized) in several equivalent forms. For example, the binomial distribution is most commonly written in terms of a probability, but can just as well be written in terms of odds.

Another related use of the term parameter is for a summary measure of a particular (theoretical) probability distribution. These are most commonly in the form of **expected values**. Expected values can be thought of as long-run averages of a random variable or some computed quantity that includes the random variable. For discrete random variables, the expected value is just a probability weighted average, i.e., the **population mean**. For example, if a random variable takes on (only) the values 2 and 10 with probabilities 5/6 and 1/6 respectively, then the expected value of that random variable is $2(5/6)+10(1/6)=20/6$. To be a bit more concrete, if someone throws a die each day and gives you \$10 if 5 comes up and \$2 otherwise, then over n days, where n is a large number, you will end up with very

close to $\$\frac{20 \cdot n}{6}$, or about \$3.67(n).

The notation for expected value is $E[\cdot]$ or $E(\cdot)$ where, e.g., $E[X]$ is read as "expected value of X" and represents the population mean of X. Other parameters such as variance, skewness and kurtosis are also expected values, but of expressions involving X rather than of X itself.

The more general formula for expected value is

$$E[g(X)] = \sum_{i=1}^{k} g(x_i)p_i = \sum_{i=1}^{k} g(x_i)f(x_i)$$

where $E[\cdot]$ or $E(\cdot)$ represents "expected value", $g(X)$ is any function of the random variable X, k (which may be infinity) is the number of values of X with non-zero probability, the x_i values are the different values of X, and the p_i values (or equivalently, $f(x_i)$) are the corresponding probabilities. Note that it is possible to define $g(X) = X$, i.e., $g(x_i) = x_i$, to find $E(X)$ itself.

The corresponding formula for expected value of a continuous random variable is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Of course if the support is smaller than the entire real line, the pdf is zero outside of the support, and it is equivalent to write the integration limits as only over the support.

To help you think about this concept, consider a discrete random variable, say W, with values -2, -1, and 3 with probabilities 0.5, 0.3, 0.2 respectively. $E(W) = -2(0.5) - 1(0.3) + 3(0.2) = -0.7$. What is $E(W^2)$? This is equivalent to letting $g(W) = W^2$ and finding $E(g(W)) = E(W^2)$. Just calculate W^2 for each W and take the weighted average: $E(W^2) = 4(0.5) + 1(0.3) + 9(0.2) = 4.1$. It is also equivalent to define, say, $U = W^2$. Then we can express $f(U)$ as U has values 4, 1, and 9 with probabilities 0.5, 0.3, and 0.2 respectively. Then $E(U) = 4(0.5) + 1(0.3) + 9(0.2) = 4.1$, which is the same answer.

Different parameters are generated by using different forms of $g(x)$.

Name	**Definition**	**Symbol**
mean	$E[X]$	μ
variance	$E[(X-\mu)^2]$	σ^2
standard deviation	$\sqrt{\sigma^2}$	σ
skewness	$E[(X-\mu)^3]/\sigma^3$	γ_1
kurtosis	$E[(X-\mu)^4]/\sigma^4 - 3$	γ_2

Table 3.2: Common parameters and their definitions as expected values.

You will need to become familiar with several parameters that are used to characterize theoretical population distributions. Technically, many of these are defined using the expected value formula (optional material) with the expressions shown in Table 3.2. You only need to become familiar with the names and symbols and their general meanings, not the "Definition" column. Note that the symbols shown are the most commonly used ones, but you should not assume that these symbols always represent the corresponding parameters or vice versa.

3.5.1 Central tendency: mean and median

The **central tendency** refers to ways of specifying where the "middle" of a probability distribution lies. Examples of central tendency include the mean and median parameters. The mean is the *probability-weighted average* of all possible values of a random variable. In other words: If we imagine taking an infinite number of random draws from the probability distribution of a random variable, the mean is the average of those infinitely-many random draws (note that values that occur more frequently will carry more weight when computing the average). Another way to think about the mean is that, if the random variable is some monetary payout, the mean is the appropriate amount to bet to come out even in the long term. Finally, another interpretation of the mean is that it represents what would be a fair redistribution of whatever the random variable represents among all subjects (for example: the "mean income" of a community is the amount of money each

person in a community would have if the total amount of money were redistributed equally). On the other hand, the median is the value that splits the distribution in half so that there is a 50/50 chance of a random value from the distribution occurring above or below the median.

The median has a more technical definition that applies even in some less common situations such as when a distribution does not have a single unique median. The median is any m such that $\mathrm{P}(X \leq m) \geq \frac{1}{2}$ and $\mathrm{P}(X \geq m) \geq \frac{1}{2}$.

3.5.2 Spread: variance and standard deviation

The **spread** of a distribution most commonly refers to the variance or standard deviation parameter, although other quantities such as interquartile range are also measures of spread.

The **population variance** is the mean squared distance of any value from the mean of the distribution, but you only need to think of it as a measure of spread on a different scale from standard deviation. The **standard deviation** is defined as the square root of the variance. It is not as useful in statistical formulas and derivations as the variance, but it has several other useful properties, so both variance and standard deviation are commonly calculated in practice. The standard deviation is in the same units as the original measurement from which it is derived. For each theoretical distribution, the intervals $[\mu-\sigma,\ \mu+\sigma]$, $[\mu-2\sigma,\ \mu+2\sigma]$, and $[\mu-3\sigma,\ \mu+3\sigma]$ include fixed known amounts of the probability. It is worth memorizing that *for Gaussian distributions only* these fractions are 0.683, 0.954, and 0.997 respectively. (I usually think of this as approximately 2/3, 95% and 99.7%.) Also, exactly 95% of the Gaussian distribution is in $[\mu-1.96\sigma,\ \mu+1.96\sigma]$.

When the standard deviation of repeated measurements is proportional to the mean, then instead of using standard deviation, it often makes more sense to measure variability in terms of the **coefficient of variation**, which is the s.d. divided by the mean.

There is a special statistical theorem (called Chebyshev's inequality) that applies to *any* shaped distribution and states that at least $\left(1 - \frac{1}{k^2}\right) \times 100\%$ of the values are within k standard deviations from the mean. For example, the interval $[\mu - 1.41\sigma,\ \mu + 1.41\sigma]$ holds at least 50% of the values, $[\mu - 2\sigma,\ \mu + 2\sigma]$ holds at least 75% of the values, and $[\mu - 3\sigma,\ \mu + 3\sigma]$ holds at least 89% of the values.

3.5.3 Skewness and kurtosis

The **population skewness** of a distribution is a measure of asymmetry (zero is symmetric). If a distribution is "pulled out" towards higher values (to the right), then it has positive **skewness**. If it is pulled out toward lower values, then it has negative skewness. A symmetric distribution, e.g., the Gaussian distribution, has zero skewness.

The **population kurtosis** of a distribution measures how far away a distribution is from a Gaussian distribution in terms of peakedness vs. flatness. Compared to a Gaussian distribution (which has kurtosis $\gamma_2 = 0$), a distribution with negative kurtosis has "rounder shoulders" and "thin tails", while a distribution with a positive kurtosis has more a more sharply shaped peak and "fat tails".

3.5.4 Miscellaneous comments on distribution parameters

Mean, variance, skewness and kurtosis are examples of **moments** of random variables. Specifically, the mean is the first (non-central) moment, and the variance, skewness, and kurtosis are the 2nd, 3rd, and 4th (central) moments. In general, the r^{th} non-central moment of X is the expected value of X^r, and the r^{th} central moment of X is the expected value of $(X - E[X])^r$. There are formulas for calculating central moments from non-central moments, e.g., $\sigma^2 = E(X^2) - E(X)^2$, where σ^2 denotes the variance of X.

It is important to realize that for any particular distribution (but not family of distributions) each parameter is a fixed constant. Also, you will recognize that these parameter names are the same as the names of statistics that can be calculated for and used as descriptions of **samples** rather than probability distributions (see next chapter). The prefix "population" is sometimes used as a reminder that we are talking about the fixed numbers for a given probability distribution rather than the corresponding sample values. For example, $E[X]$ denotes the population mean (some unknown number), and $\bar{X}$ commonly denotes the sample mean (i.e., the average value of X for the units in the sample—this is a random variable, because it will vary from sample to sample).

It is worth knowing that any formula applied to one or more parameters creates a new parameter. For example, if μ_1 and μ_2 are parameters for some population, say, the mean dexterity with the subjects' dominant and non-dominant hands, then $\log(\mu_1)$, μ_2^2, $\mu_1 - \mu_2$ and $(\mu_1 + \mu_2)/2$ are also parameters.

In addition to the parameters in the above table, which are the most common descriptive parameters that can be calculated for any distribution, fixed constants in a pmf or pdf, such as degrees of freedom (see below) or the n in the binomial distribution are also (somewhat loosely) called parameters.

Technical note: For some distributions, parameters such as the mean or variance may be infinite.

Parameters such as (population) mean and (population) variance are fixed quantities that characterize a given probability distribution. The (population) skewness characterizes symmetry, and (population) kurtosis characterizes symmetric deviations from Normality. Corresponding sample statistics can be thought of as sample estimates of the population quantities.

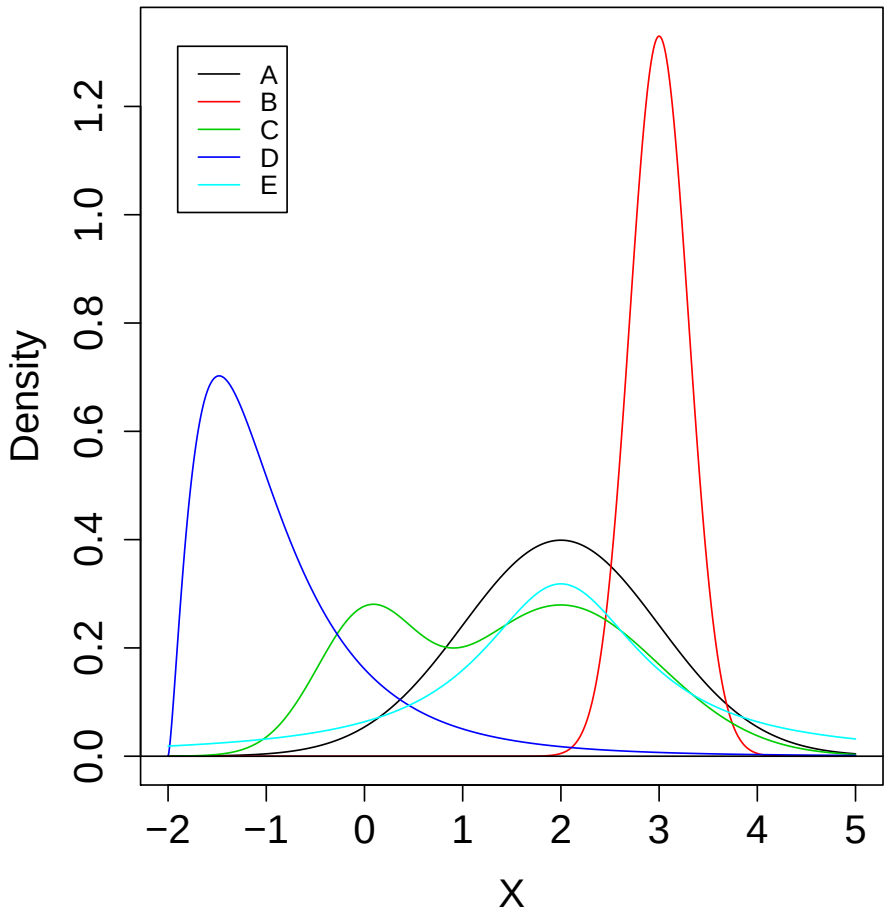


Figure 3.1: Various probability density functions.

3.5.5 Examples

As a review of the concepts of theoretical population distributions (in the continuous random variable case) let's consider a few examples.

Figure 3.1 shows five different pdf's representing the (population) probability distributions of five different continuous random variables. By the rules of pdf's, the area under each of the five curves equals exactly 1.0, because that represents the probability that a random outcome from a distribution is between -infinity and +infinity. (The area shown, between -2 and +5 is slightly less than 1.0 for each distribution because there is a small chance that these variables could have an outcome outside of the range shown.) You can see that distribution **A** is a unimodal (one peak) symmetric distribution, centered around 2.0. Although you cannot see it by eye, it has the perfect bell-shape of a Gaussian distribution. Distribution **B** is also Gaussian in shape, has a different central tendency (shifted higher or rightward), and has a smaller spread. Distribution **C** is bimodal (two peaks) so it cannot be a Gaussian distribution. Distribution **D** has the lowest center and is asymmetric (skewed to the right), so it cannot be Gaussian. Distribution **E** appears similar to a Gaussian distribution, but while symmetric and roughly bell-shaped, it has "tails" that are too fat to be a true bell-shaped, Gaussian distribution.

So far we have been talking about the parameters of a given, known, theoretical probability distribution. A slightly different context for the use of the term parameter is in respect to a real world population, either finite (but usually large) or infinite. As two examples, consider the height of all people living on the earth at 3:57 AM GMT on September 10, 2007, or the birth weights of all of the Sprague-Dawley breed of rats that could possibly be bred. The former is clearly finite, but large. The latter is perhaps technically finite due to limited resources, but may also be thought of as (practically) infinite. Each of these must follow some true distribution with fixed parameters, but these are practically unknowable. The best we can do with experimental data is to make an estimate of the fixed, true, unknowable parameter value. For this reason, I call parameters in this context "secrets of nature" to remind you that they are not random and they are not practically knowable.

3.6 Multivariate distributions: joint, conditional, and marginal

The concepts of this section are fundamentals of probability, but for the typical user of statistical methods, only a passing knowledge is required. More detail is given here for the interested reader.

So far we have looked at the distribution of a single random variable at a time. Now we proceed to look at the **joint distribution** of two (or more) random variables. First consider the case of two categorical random variables. As an example, consider the population of all cars produced in the world in 2006. (I'm just making up the numbers here.) This is a large finite population from which we might sample cars to do a fuel efficiency experiment. If we focus on the categorical variable "origin" with levels "US","Japanese", and "Other", and the categorical variable "size" with categorical variable "Small", "Medium" and "Large", then Table 3.3 would represent the joint distribution of origin and size in this population.

origin / size	Small	Medium	Large	Total
US	0.05	0.10	0.15	
Japanese	0.20	0.10	0.05	
Other	0.15	0.15	0.05	
Total				1.00

Table 3.3: Joint distribution of car origin and size.

These numbers come from categorizing all cars, then dividing the total in each combination of categories by the total cars produced in the world in 2006, so they are "relative frequencies". But because we are considering this the whole population of interest, it is better to consider these numbers to be the probabilities of a (joint) pmf. Note that the total of all of the probabilities is 1.00. Reading this table we can see, e.g., that 20% of all 2006 cars were small Japanese cars, or equivalently, the probability that a randomly chosen 2006 car is a small Japanese car is 0.20.

The joint distribution of X and Y is summarized in the joint pmf, which can be tabular or in formula form, but in either case is similar to the one variable pmf of Section 3.2 except that it defines a probability for each combination of levels of X and Y.

This idea of a joint distribution, in which probabilities are given for the combination of levels of two categorical random variables, is easily extended to three or more categorical variables.

The joint distribution of a pair of categorical random variables represents the probabilities of combinations of levels of the two individual random variables.

origin / size	Small	Medium	Large	Total
US	0.05	0.10	0.15	0.30
Japanese	0.20	0.10	0.05	0.35
Other	0.15	0.15	0.05	0.35
Total	0.40	0.35	0.25	(1.00)

Table 3.4: Marginal distributions of car origin and size.

Table 3.4 adds the margins to the previous table, which are defined as the row sums and column sums (labeled "Total"). Note that both the right vertical and bottom horizontal margins add to 1.00, and so they each represent a probability distribution, in this case of origin and size respectively. These distributions are called the **marginal distributions** and each represents the pmf of one of the variable *ignoring* the other variable. That is, a marginal distribution is the distribution of any particular variable when we don't pay any attention to the other variable(s). If we had only studied car origins, we would have found the population distribution to be 30% US, 35% Japanese and 35% other.

It is important to understand that every variable we measure is marginal with respect to all of the other variables that we could measure on the same units or subjects, and which we do not in any way control (or in other words, which we let vary freely).

The marginal distribution of any variable with respect to any other variable(s) is just the distribution of that variable ignoring the other variable(s).

origin / size	Small	Medium	Large	Total
US	0.167	0.333	0.400	1.000
Japanese	0.571	0.286	0.143	1.000
Other	0.429	0.429	0.142	1.000

Table 3.5: Conditional distributions of car size given its origin.

The third and final definition for describing distributions of multiple characteristics of a population of units or subjects is the **conditional distribution** which relates to conditional probability (see Page 30). As shown in Table 3.5, the conditional distribution refers to fixing the level of one variable, then "renormalizing" to find the probability level of the other variable when we only focus on or consider those units or subjects that meet the condition of interest.

So if we focus on Japanese cars only (technically, we condition on cars being Japanese) we see that 57.1% of those cars are small, which is very different from either the marginal probability of a car being small (0.40) or the joint probability of a car being small and Japanese (0.20). The formal notation here is Pr(size=small|origin=Japanese) = 0.571, which is read as "the probability of a car being small given that the car is Japanese equals 0.571". Note that the probabilities in Table 3.5 are obtained by focusing on a particular row of Table 3.4 and then dividing the "Small," "Medium," and "Large" probabilities by the "Total" probability in that row. Dividing by the row total is what we mean by "renormalizing"—it alters the probabilities in that row such that they add up to one.

It is important to realize that there is another set of conditional distributions for this example that we have not looked at. As an exercise, try to find the conditional distributions of "origin" given "size", which differ from the distributions of "size" given "origin" of Table 3.5.

It is interesting and useful to note that an equivalent alternative to specifying the complete joint distribution of two categorical (or quantitative)

random variables is to specify the marginal distribution of one variable, and the conditional distributions for the second variable at each level of the first variable. For example, you can reconstruct the joint distribution for the cars example from the marginal distribution of "origin" and the three conditional distributions of "size given origin". This leads to another way to think about marginal distributions as the distribution of one variable *averaged over* the distribution of the other.

The distribution of a random variable conditional on a particular level of another random variable is the distribution of the first variable when the second variable is fixed to the particular level.

The concepts of joint, marginal and conditional distributions transfer directly to two continuous distributions, or one continuous and one discrete distribution, but the details will not be given here. Suffice it to say that the joint pdf of two continuous random variables, say X and Y, is a formula with both xs and ys in it.

3.6.1 Covariance and Correlation

For two quantitative variables, the basic parameters describing the strength of their relationship are **covariance** and **correlation**. For both, larger absolute values indicate a stronger relationship, and positive numbers indicate a codirectional relationship while negative numbers indicate an contradirectional relationship. For both, a value of zero is called uncorrelated. Covariance depends on the scale of measurement, while correlation does not. For this reason, correlation is easier to understand, and we will focus on that here, although if you look at the gray box below, you will see that covariance is used as in intermediate in the calculation of correlation—in fact, correlation is just the standardized version of covariance. (Note that here we are concerned with the "population" or "theoretical" correlation. The sample version is covered in the EDA chapter.)

Correlation describes both the strength and direction of the (linear) relationship between two variables. Correlations run from -1.0 to +1.0. A negative correlation

indicates an "inverse" relationship such that population units that are low for one variable tend to be high for the other (and vice versa), while a positive correlation indicates a "direct" relationship such that population units that are low in one variable tend to be low in the other (also high with high). A zero correlation (also called **uncorrelated**) indicates that the "best fit straight line" (see the chapter on Regression) for a plot of X vs. Y is horizontal, suggesting no relationship between the two random variables. Technically, independence of two variables (see above) implies that they are uncorrelated, but the reverse is not necessarily true.

For a correlation of +1.0 or -1.0, Y can be perfectly predicted from X with no error (and vice versa) using a linear equation. For example if X is temperature of a rat in degrees C and Y is temperature in degrees F, then $Y = 9/5 * C + 32$, exactly, and the correlation is +1.0. And if X is height in feet of a person from the floor of a room with an 8 foot ceiling and Y is distance from the top of the head to the ceiling, then $Y = 8 - X$, exactly, and the correlation is -1.0. For other variables like height and weight, the correlation is positive, but less than 1.0. And for variables like IQ and length of the index finger, the correlation is presumably 0.0.

It should be obvious that the correlation of any variable with itself is 1.0. Let us represent the population correlation between random variable X_i and random variable X_j as $\rho_{i,j}$. Because the correlation of X with Y is the same as Y with X, it is true that $\rho_{i,j} = \rho_{j,i}$. We can compactly represent the relationships between multiple variables with a **correlation matrix** which shows all of the pairwise correlations in a square table of numbers (square matrix). An example is given in Table 3.6 for the case of 4 variables. As with all correlation matrices, the matrix is symmetric with a row of ones on the main diagonal. For some actual population and variables, we could put numbers instead of symbols in the matrix, and then make statements about which variables are directly vs. inversely vs. not correlated, and something about the strengths of the correlations.

Variable	X_1	X_2	X_3	X_4
X_1	1	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$
X_2	$\rho_{2,1}$	1	$\rho_{2,3}$	$\rho_{2,4}$
X_3	$\rho_{3,1}$	$\rho_{3,2}$	1	$\rho_{3,4}$
X_4	$\rho_{4,1}$	$\rho_{4,2}$	$\rho_{4,3}$	1

Table 3.6: Population correlation matrix for four variables.

There are several ways to measure "correlation" for categorical variables and choosing among them can be a source of controversy that we will not cover here. But for quantitative random variables covariance and correlation are mathematically straightforward.

The population covariance of two quantitative random variables, say X and Y, is calculated by computing the expected value (population mean) of the quantity $(X - \mu_X)(Y - \mu_Y)$ where μ_X is the population mean of X and μ_Y is the population mean of Y across all combinations of X and Y. For continuous random variables this is the double integral

$$\text{Cov}_{X,Y} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$

where $f(x, y)$ is the joint pdf of X and Y.

For discrete random variables we have the analogous form

$$\text{Cov}_{X,Y} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_X)(y - \mu_Y) f(x, y)$$

where $f(x, y)$ is the joint pmf, and $\mathcal{X}$ and $\mathcal{Y}$ are the respective supports of X and Y.

As an example consider a population consisting of all of the chickens of a particular breed (that only lives 4 years) belonging to a large multi-farm poultry company in January of 2007. For each chicken in this population we have X equal to the number of eggs laid in the first week of January and Y equal to the age of the chicken in years. The joint pmf of X and Y is given in Table 3.7. As usual, the joint pmf gives the probabilities that a random subject will fall into each combination of categories from the two variables.

We can calculate the (marginal) mean number of eggs from the marginal distribution of eggs as $\mu_X = 0(0.35) + 1(0.40) + 2(0.25) = 0.90$ and the mean age as $\mu_Y = 1(0.25) + 2(0.40) + 3(0.20) + 4(0.15) = 2.25$ years.

The calculation steps for the covariance are shown in Table 3.8. The population covariance of X and Y is 0.075 (exactly). The (weird) units are "egg years".

Population correlation can be calculated from population covariance and the two individual standard deviations using the formula

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}.$$

In this case $\sigma_X^2 = (0-0.9)^2(0.35)+(1-0.9)^2(0.40)+(2-0.9)^2(0.25) = 0.59$. Using a similar calculation for σ_Y^2 and taking square roots to get standard deviation from variance, we get

$$\rho_{X,Y} = \frac{0.075}{0.7681 \cdot 0.9937} = 0.0983$$

which indicates a weak positive correlation: older hens lay more eggs.

Y (year) / X (eggs)	0	1	2	Margin
1	0.10	0.10	0.05	0.25
2	0.15	0.15	0.10	0.40
3	0.05	0.10	0.05	0.20
4	0.05	0.05	0.05	0.15
Margin	0.35	0.40	0.25	1.00

Table 3.7: Chicken example: joint population pmf.

In a nutshell: When dealing with two (or more) random variables simultaneously it is helpful to think about joint vs. marginal vs. conditional distributions. This has to do with what is fixed vs. what is free to vary, and what adds up to 100%. The parameter that describes the strength of relationship between two random variables is the correlation, which ranges from -1 to +1.

X	Y	X-0.90	Y-2.25	Pr	Pr·(X-0.90)(Y-2.25)
0	1	-0.90	-1.25	0.10	0.11250
1	1	0.10	-1.25	0.10	-0.00125
2	1	1.10	-1.25	0.05	-0.06875
0	2	-0.90	-0.25	0.15	0.03375
1	2	0.10	-0.25	0.15	-0.00375
2	2	1.10	-0.25	0.10	-0.02750
0	3	-0.90	0.75	0.05	-0.03375
1	3	0.10	0.75	0.10	0.00750
2	3	1.10	0.75	0.05	0.04125
0	4	-0.90	1.75	0.05	-0.07875
1	4	0.10	1.75	0.05	0.00875
2	4	1.10	1.75	0.05	0.09625
			Total	1.00	0.07500

Table 3.8: Covariance calculation for chicken example.

3.7 Key application: sampling distributions

In this course we will generally be concerned with analyzing a **simple random sample** of size n which indicates that we randomly and independently choose n subjects from a large or infinite population for our experiment. (For practical issues, see Section 7.2.) Then we make one or more measurements, which are the realizations of some random variable. Often we combine these values into one or more **statistics**. A statistic is defined as any formula or "recipe" that can be explicitly calculated from observed data. Note that the formula for a statistic must not include unknown parameters. *When thinking about a statistic for a particular sample, always remember that this is only one of many possible values that we could have gotten for this statistic, based on the random nature of the sampling.*

If we think about random variable X for a sample of size n it is useful to consider this a multivariate situation, i.e., the outcome of the random trial is X_1 through X_n and there is a probability distribution for this multivariate outcome. If we have simple random sampling, this n-fold pmf or pdf is calculable from the distribution of the original random variable and the laws of probability with independence. Technically we say that X_1 through X_n are **iid** which stands for independent and identically distributed, which indicates that distribution of the outcome for, say,

X_1	X_2	X_3	Probability
F	F	F	0.216
M	F	F	0.144
F	M	F	0.144
F	F	M	0.144
F	M	M	0.096
M	F	M	0.096
M	M	F	0.096
M	M	M	0.064
Total			1.000

Table 3.9: Multivariate pmf for animal gender.

the third subject, is the same as for any other subject and is independent of (does not depend on the outcome of) the outcome for every other subject.

An example should make this clear. Consider a simple random sample of size $n = 3$ from a population of animals. The random variable we will observe is sex (treated as male or female here) of each animal in the sample. Furthermore, we will call X the random variable, and we will call X_1, X_2, and X_3 the particular *realizations* (or observations) of this random variable for the sample we observe. Let's say we know the true probability that an animal is male is equal to 0.4. Then the probability that an animal is female is 0.6. We can work out the multivariate pmf case by case, as is shown in Table ??. For example, the chance that the outcome is FMF in that order is (0.6)(0.4)(0.6)=0.144.

Using this multivariate pmf, we can easily calculate the pmf for derived random variables (statistics) such as Y=the number of females in the sample: Pr(Y=0)=0.064, Pr(Y=1)=0.288, Pr(Y=2)=0.432, and Pr(Y=3)=0.216.

Now think carefully about what we just did. We found the probability distribution of random variable Y, the number of females in a sample of size three. This is called the **sampling distribution** of Y, which refers to the fact that Y is a random quantity which varies from sample to sample over many possible samples (or experimental runs) that *could* be carried out if we had enough resources. We can find the sampling distribution of various sample quantities constructed from the data of a random sample. These quantities are **sample statistics**, and can take many different forms. Among these are the sample versions of mean, variance,

standard deviation, etc. Quantities such as the sample mean or sample standard deviation (see section 4.2) are often used as estimates of the corresponding population parameters. The sampling distribution of a sample statistic is then the key way to evaluate *how good of an estimate* a sample statistic is. In addition, we use various sample statistics and their sampling distributions to make probabilistic conclusions about statistical hypotheses, usually in the form of statements about population parameters.

Much of the statistical analysis of experiments is grounded in calculation of a sample statistic, computation of its sampling distribution (using a computer), and using the sampling distribution to draw inferences about statistical hypotheses.

3.8 Central limit theorem

The Gaussian (also called bell-shaped or Normal) distribution is a very common one. The central limit theorem (CLT) explains why many real-world variables follow a Gaussian distribution.

It is worth reviewing here what "follows a particular distribution" really means. A random variable follows a particular distribution if the observed probability of each outcome for a discrete random variable or the observed probabilities of a reasonable set of intervals for a continuous random variable are well approximated by the corresponding probabilities of some named distribution (see Common Distributions, below). Roughly, this means that a histogram of the actual random outcomes is quite similar to the theoretical histogram of potential outcomes defined by the pmf (if discrete) or pdf (if continuous). For example, for any Gaussian distribution with mean μ and standard deviation σ, we expect 2.3% of values to fall below $\mu-2\sigma$, 13.6% to fall between μ–2σ and μ–σ, 34.1% between μ–σ and μ, 34.1% between μ and μ+σ, 13.6% between μ+σ and μ+2σ, and 2.3% above μ+2σ. In practice we would check a finer set of divisions and/or compare the shapes of the actual and theoretical distributions either using histograms or a special tool called the quantile-quantile plot.

In non-mathematical language, the "CLT" says that *whatever* the pmf or pdf of a variable is, if we randomly sample a "large" number (say k) of independent

values from that random variable, the sum or mean of those k values, if collected repeatedly, will have a Normal distribution. It takes some extra thought to understand what is going on here. The process I am describing here takes a sample of (independent) outcomes, e.g., the weights of all of the rats chosen for an experiment, and calculates the mean weight (or sum of weights). Then we consider the less practical process of repeating the whole experiment many, many times (taking a new sample of rats each time). If we would do this, the CLT says that a histogram of all of these mean weights across all of these experiments would show a Gaussian shape, even if the histogram of the individual weights of any one experiment were not following a Gaussian distribution. By the way, the distribution of the means across many experiments is usually called the "sampling distribution of the mean".

For practical purposes, a number as small as 20 (observations per experiment) can be considered "large" when invoking the CLT if the original distribution is not very bizarre in shape and if we only want a reasonable approximation to a Gaussian curve. And for almost all original distributions, the larger k is, the closer the distribution of the means or sums are to a Gaussian shape.

It is usually fairly easy to find the mean and variance of the sampling distribution (see Section 3.7) of a statistic of interest (mean or otherwise), but finding the *shape* of this sampling distribution is more difficult. The Central Limit Theorem lets us predict the (approximate) shape of the sampling distribution for sums or means. And this additional shape information is usually all that is needed to construct valid confidence intervals and/or p-values.

But wait, there's more! The central limit theorem also applies to the sum or mean of many *different* independent random variables as long as none of them strongly dominates the others. So we can invoke the CLT as an explanation for why many real-world variables happen to have a Gaussian distribution. It is because they are the result of many small independent effects. For example, the weight of 12-week-old rats varies around the mean weight of 12-week-old rats due to a variety of genetic factors, differences in food availability, differences in exercise, differences in health, and a variety of other environmental factors, each of which adds or subtracts a little bit relative to the overall mean.

See one of the theoretical statistics texts listed in the bibliography for a proof of the CLT.

The Central Limit Theorem is the explanation why many real-world random variables tend to have a Gaussian distribution. It is also the justification for assuming that if we could repeat an experiment many times, any sample mean that we calculate once per experiment would follow a Gaussian distribution over the many experiments.

3.9 Common distributions

A brief description of several useful and commonly used probability distributions is given here. The casual reader will want to just skim this material, then use it as reference material as needed.

The two types of distributions are discrete and continuous (see Section 3.1), which are fully characterized by their pmf or pdf respectively. In the notation section of each distribution we use "$X \sim$" to mean "X is distributed as".

What does it mean for a random variable to follow a certain distribution? It means that the pdf or pmf of that distribution fully describes the probabilities of events for that random variable. Note that each of the named distributions described below are a family of related individual distributions from which a specific distribution must be specified using an index or pointer into the family, usually called a parameter (or sometimes using 2 parameters). For a theoretical discussion, where we assume a particular distribution and then investigate what properties follow, the pdf or pmf is all we need.

For data analysis, we usually need to choose a theoretical distribution that we think will well approximate our measurement for the population from which our sample was drawn. This can be done using information about what assumptions lead to each distribution, looking at the support and shape of the sample distribution, and using prior knowledge of similar measurements. Usually we choose a family of distributions, then use statistical techniques to estimate the parameter that chooses the particular distribution that best matches our data. Also, after carrying out a statistical test that assumes a particular family of distributions, we use model checking, such as residual analysis, to verify that our choice was a good one.

3.9.1 Binomial distribution

The **binomial distribution** is a discrete distribution that represents the number of successes in n independent trials, each of which has success probability p. Binomial distributions often come up in applications with binary data (e.g., polls assessing whether or not people approve of a particular policy, education data recording whether or not a student graduates from high school, or epidemiological records denoting the number of deaths and survivals of a disease). All of the (infinite) different values of n and p define a whole family of different binomial distributions. The outcome of a random variable that follows a binomial distribution is a whole number from 0 to n (i.e., n+1 different possible values). If $n = 1$, the special name **Bernoulli distribution** may be used. If random variable X follows a Bernoulli distribution with parameter p, then stating that $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$ fully defines the distribution of X.

If we let X represent the random outcome of a binomial random variable with parameters n and p, and let x represent any particular outcome (as a whole number from 0 to n), then the pmf of a binomial distribution tells us the probability that the outcome will be x:

$$\Pr(X = x) = f(x) = \left(\frac{n!}{(n - x)!\, x!} \right) p^x (1 - p)^{n-x}.$$

As a reminder, the exclamation mark symbol is pronounced "factorial" and $r!$ represents the product of all the integers from 1 to r. As an exception, $0! = 1$.

The true, theoretical mean of a binomial distribution is np and the variance is $np(1 - p)$. These refer to the ideal for an infinite population. For a sample, the sample mean and variance will be similar to the theoretical values, and the larger the sample, the more sure we are that the sample mean and variance will be very close to the theoretical values.

As an example, if you buy a lottery ticket for a daily lottery choosing your lucky number each of 5 different days in a lottery with a 1/500 chance of winning each time, then knowing that these chances are independent, we could call the number of times (out of 5) that you win Y, and state that Y is distributed according to a binomial distribution with $n = 5$ and $p = 0.002$. We now know that if many people each independently buy 5 lottery tickets they will each have an outcome between 0 and 5, and the mean of all of those outcomes will be (close to) $np = 5(0.002) = 0.01$ and the variance will be (close to) $np(1 - p) = 5(0.002)(0.998) = 0.00998$ (with

sd=$\sqrt{0.0098} = 0.0999$.)

In this example we can calculate $n! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, and for x=2, $(n-x)! = 3! = 3 \cdot 2 \cdot 1 = 6$ and $x! = 2! = 2 \cdot 1 = 2$. So

$$\Pr(X = 2) = \left(\frac{120}{6 \cdot 2}\right) 0.002^2 (0.998)^3 = 0.0000398.$$

Roughly 4 out of 100,000 people will win twice in 5 days.

It is sometimes useful to know that with large n a binomial random variable with parameter p approximates a Normal distribution with mean np and variance $np(1-p)$ (except that there are gaps in the binomial because it only takes on whole numbers).

Common notation is $X \sim \text{bin}(n, p)$.

3.9.2 Multinomial distribution

The **multinomial distribution** is a discrete distribution that can be used to model situations where a subject has n trials each of which independently can result in one of k different values which occur with probabilities $(p_1, p_2, \ldots, p_k)$, where $p_1 + p_2 + \ldots + p_k = 1$. The outcome of a multinomial is a list of k numbers adding up to n, each of which represents the number of times a particular value was achieved. In this sense, the multinomial distribution is just a generalization of the binomial distribution (i.e., the binomial distribution is a multinomial distribution with $k = 2$). Thus, multinomial distributions often come up in applications where there are $k \geq 3$ outcomes (e.g., surveys recording if people "strongly disagree", "disagree", ..., "strongly agree" with something, or data recording people's race or ethnicity). Note that the outcomes may be ordinal (as in the "strongly disagree"/"strongly agree" example) or not ordinal (as in the race/ethnicity example).

For random variable X following the multinomial distribution, the outcome is the list of values $(x_1, x_2, \ldots, x_k)$ and the pmf is:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \left(\frac{n!}{x_1! \cdot x_2! \cdots x_k!}\right) p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

For example, consider a kind of candy that comes in an opaque bag and has three colors (red, blue, and green) in different amounts in each bag. If 30% of the bags have red as the most common color, 20% have green, and 50% have blue,

then we could imagine an experiment consisting of opening n randomly chosen bags and recording for each bag which color was most common. Here $k = 3$ and $p_1 = 0.30$, $p_2 = 0.20$, and $p_3 = 0.50$. The outcome is three numbers, e.g., x_1=number of times that red was most common, x_2=number of times blue is most common, and x_3=number of times green is most common. If we choose n=2, one calculation we can make is

$$\Pr(x_1 = 1, x_2 = 1, x_3 = 0) = \left(\frac{2!}{1! \cdot 1! \cdot 0!}\right) 0.30^1\ 0.20^1\ 0.50^0 = 0.12$$

and the whole pmf can be represented in this tabular form (where "# of Reds" means number of bags where red was most common, etc.):

x_1 (# of Reds)	x_2 (# of Blues)	x_3 (# of Greens)	Probability
2	0	0	0.09
0	2	0	0.04
0	0	2	0.25
1	1	0	0.12
1	0	1	0.30
0	1	1	0.20

Common notation is $X \sim \text{MN}(n, p_1, \ldots, p_k)$.

3.9.3 Poisson distribution

The **Poisson distribution** is a discrete distribution whose support is the non-negative integers $(0, 1, 2, \ldots)$. Many measurements that represent counts which have no theoretical upper limit—such as the number of times subjects click on an online ad, or the number of times that certain mental health services are used—follow a Poisson distribution. A Poisson distribution is applicable when the chance of a countable event is proportional to the time (or distance, etc.) available, when the chances of events in non-overlapping intervals are independent, and when the chance of two events in a very short interval is essentially zero.

A Poisson distribution has one parameter, usually represented as λ (lambda). The pmf is:

$$\Pr(X = x) = f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

The mean is λ and the variance is also λ. From the pmf, you can see that the probability of no events, $\Pr(X = 0)$, equals $e^{-\lambda}$.

If the data show a substantially larger variance than the mean, then a Poisson distribution is not appropriate. A common alternative is the **negative binomial distribution** which has the same support, but has two parameters often denoted p and r. The negative binomial distribution can be thought of as the number of trials until the r^{th} success when the probability of success is p for each trial.

It is sometimes useful to know that with large λ a Poisson random variable approximates a Normal distribution with mean λ and standard deviation $\sqrt{\lambda}$ (except that there are gaps in the Poisson because it only takes on whole numbers).

Common notation is $X \sim \text{Pois}(\lambda)$.

3.9.4 Gaussian distribution

The **Gaussian or Normal distribution** is a continuous distribution with a symmetric, bell-shaped pdf curve as shown in Figure 3.2. The members of this family are characterized by two parameters, the mean and the variance (or standard deviation) usually written as μ and σ^2 (or σ). The support is all of the real numbers, but the "tails" are very thin, so the probability that X is more than 4 or 5 standard deviations from the mean is extremely small. The pdf of the Normal distribution is:

$$f(x) = \frac{1}{\sqrt{2}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Among the family of Normal distributions, the standard normal distribution, the one with $\mu = 0$ and $\sigma^2 = 1$ is special. It is the one for which you will find information about the probabilities of various intervals in textbooks. This is useful because the probability that the outcome will fall in, say, the interval from minus infinity to any arbitrary number x for a non-standard normal distribution, say, X, with mean $\mu \neq 0$ and standard deviation $\sigma \neq 1$ is the same as the probability that the outcome of a standard normal random variable, usually called Z, will be less than $z = \frac{x-\mu}{\sigma}$, where the formula for z is the "z-score" formula.

Of course, there is not really anything "normal" about the Normal distribution, so I always capitalize "Normal" or use Gaussian to remind you that we are just talking about a particular probability distribution, and not making any judgments about normal vs. abnormal. The Normal distribution is a very commonly used distribution (see CLT, in Section 3.8), and it is quite flexible in that the center and spread can be set to any values independently. (In fact, the Normal distribution is basically the only distribution where the mean and variance are independent.)

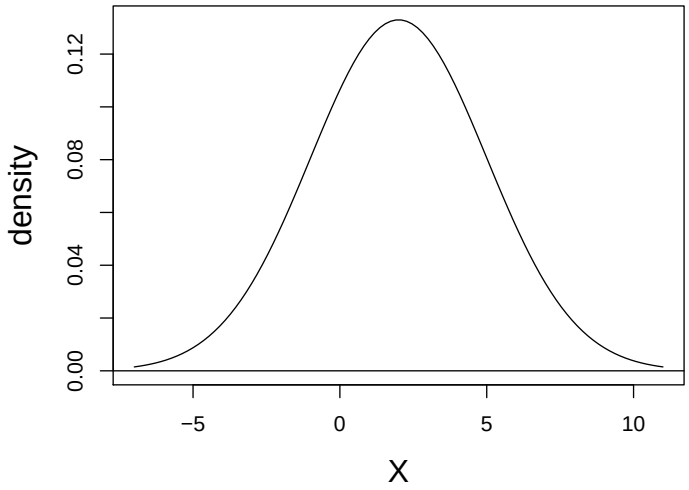


Figure 3.2: Gaussian bell-shaped probability density function

On the other hand, not every distribution that subjectively looks "bell-shaped" is a Normal distribution. Some distributions are flatter than a Normal distribution, with "thin tails" (negative kurtosis), and some distributions are more "peaked" than a Normal distribution and thus have "fatter tails" (called positive kurtosis). An example of this is the t-distribution (see below).

Common notation is $X \sim \mathrm{N}(\mu, \sigma^2)$.

3.9.5 t-distribution

The **t-distribution** is a continuous distribution with a symmetric, unimodal pdf centered at zero that has a single parameter called the "degrees of freedom" (df). In this context you can think of df as just an index or pointer which selects a single distribution out of a family of related distributions. For other ways to think about df, see Section 3.10. The support is all of the real numbers. The t-distribution has fatter tails than the Normal distribution, but it approaches the shape of the Normal distribution as the df increases. The t-distribution arises most commonly when analyzing one or more sample means and the standard deviation of the sampling distribution is estimated from the data rather than known (which is almost always the case). Intuitively, it is this additional uncertainty in the standard deviation (because we need to estimate it) that causes the widening of the distribution from Normal to t.

Common notation is $X \sim \mathrm{t}_{df}$.

3.9.6 Chi-square distribution

A **chi-square distribution** is a continuous distribution with support on the positive real numbers whose family is indexed by a single "degrees of freedom" parameter. A chi-square distribution with df equal to a has the same distribution as the sum of squares of a-many independent N(0,1) random variables. Thus, the chi-square distribution comes up quite often when we compute the sum of squares of independent z-scores. The mean is equal to the df and the variance is equal to twice the df.

Common notation is $X \sim \chi^2_{df}$.

3.9.7 F-distribution

The **F-distribution** is a continuous distribution with support on the positive real numbers. The family encompasses a large range of unimodal, asymmetric shapes determined by two parameters which are usually called numerator and denominator degrees of freedom. The F-distribution is very commonly used in analysis of experiments. If X and Y are two independent chi-square random variables with r and s df respectively, then $\frac{X/r}{Y/s}$ defines a new random variable that follows the F-distribution with r and s df. The mean is $\frac{s}{s-2}$ and the variance is a complicated function of r and s.

Common notation is $X \sim \mathrm{F}(r, s)$.

3.10 A note on degrees of freedom

Degrees of freedom are numbers that characterize specific distributions in a family of distributions. Often we find that a certain family of distributions is needed for a particular application, and then we need to calculate the degrees of freedom to know which specific distribution within the family is appropriate.

The most common situation is when we have a particular statistic and want to know its sampling distribution. If the sampling distribution falls in the "t" family as when performing a t-test, or in the "F" family when performing an ANOVA, or in several other families, we need to find the number of degrees of freedom to figure out which particular member of the family actually represents the desired sampling distribution. In this way, the degrees of freedom is similar to the sample mean: When we consider using a Normal distribution for our data, we dont just use any Normal distribution—we focus on the distribution whose mean is equal to the sample mean, and whose variance is equal to the sample variance. In fact, the mean and variance of the t-distribution, chi-square distribution, and F distribution are all characterized by their degrees of freedom. In short, the degrees of freedom is just the parameter of certain distributions.

Yet another way to think about degrees of freedom for a particular statistic is that they represent the number of independent pieces of information that go into the calculation of the statistic. This is why, for example, the sample variance, defined as $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$, has $n-1$ degrees of freedom. We can see that variance is computed conditional on $\bar{x}$ (i.e., we need to compute $\bar{x}$ before we can compute

variance); thus, once we know $\bar{x}$, we only need $n-1$ independent datapoints to compute the variance, because the last datapoint will be determined by $\bar{x}$ (which we already know). This is what people mean when they say that one degree of freedom is "used up" to determine the sample mean when computing the variance.

Consider 5 numbers with a mean of 10. To calculate the variance of these numbers we need to sum the squared deviations (from the mean). It really doesn't matter whether the mean is 10 or any other number: as long as all five deviations are the same, the variance will be the same. This make sense because variance is a pure measure of spread, not affected by central tendency. But by mathematically rearranging the definition of mean, it is not too hard to show that the sum of the deviations (not squared) is always zero. Therefore, the first four deviations can (freely) be any numbers, but then the last one is forced to be the number that makes the deviations add to zero, and we are not free to choose it. It is in this sense that five numbers used for calculating a variance or standard deviation have only four degrees of freedom (or independent useful pieces of information). In general, a variance or standard deviation calculated from n data values and one mean has $n-1$ df.

Another example is the "pooled" variance from k independent groups. If the sizes of the groups are n_1 through n_k, then each of the k individual variance estimates is based on deviations from a different mean, and each has one less degree of freedom than its sample size, e.g., $n_i - 1$ for group i. We also say that each numerator of a variance estimate, e.g., SS_i, has $n_i - 1$ df. The pooled estimate of variance is

$$s^2_{\text{pooled}} = \frac{\mathrm{SS}_1 + \cdots + \mathrm{SS}_k}{\mathrm{df}_1 + \cdots + \mathrm{df}_k}$$

and we say that both the numerator SS and the entire pooled variance has $\mathrm{df}_1 + \cdots + \mathrm{df}_k$ degrees of freedom, which suggests how many independent pieces of information are available for the calculation.

Chapter 4

Exploratory Data Analysis

A first look at the data.

As mentioned in Chapter 1, exploratory data analysis or "EDA" is a critical first step in analyzing the data from an experiment. Here are the main reasons we use EDA:

- detection of mistakes
- checking of assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables, and
- assessing the direction and rough size of relationships between explanatory and outcome variables.

Loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis.

Most importantly, EDA is usually an iterative process of scientific inquiry: When handed a dataset, you will have certain questions about the dataset, so you'll perform EDA to answer these questions, and maybe the EDA let's new questions arise, and so on. For example, as we will discuss, EDA can be used to assess the relationship between two random variables. If we see a positive correlation between two random variables (X and Y, say), and there is another categorical

random variable (C, say), a natural question to ask is: Does this positive relationship between X and Y hold across all values of C? This will lead to more EDA techniques, which may lead to more questions to ask, and so on. As we will see, there are two big benefits to this kind of iterative EDA: It allows us to detect nuanced relationships in our data, and it helps us assess what kinds of statistical analyses are most appropriate for our data.

4.1 Typical data format and the types of EDA

The data from an experiment are generally collected into a rectangular array (e.g., spreadsheet or database), most commonly with one row per experimental subject and one column for each subject identifier, outcome variable, and explanatory variable. Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable. (Some more complicated experiments require a more complex data layout.)

People are not very good at looking at a column of numbers or a whole spreadsheet and then determining important characteristics of the data. They find looking at numbers to be tedious, boring, and/or overwhelming. Exploratory data analysis techniques have been devised as an aid in this situation. Most of these techniques work in part by hiding certain aspects of the data while making other aspects more clear.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

Non-graphical methods generally involve calculating summary statistics, while graphical methods summarize the data in a diagrammatic or pictorial way. Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables. *It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.*

Throughout EDA, it is also important to keep in mind the role (outcome or explanatory) and type (categorical or quantitative) of the variable(s) being examined, because that will determine the type of EDA that is most appropriate for

the data at hand.

Although there are guidelines about which EDA techniques are useful in what circumstances, there is an important degree of looseness and art to EDA. Competence and confidence come with practice, experience, and close observation of others. Also, EDA need not be restricted to techniques you have seen before; sometimes you need to invent a new way of looking at your data.

The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical.

This chapter first discusses the non-graphical and graphical methods for looking at single variables, then moves on to looking at multiple variables at once. Univariate EDA focuses on summarizing key characteristics about the distribution of variables, while multivariate EDA focuses on investigating the relationship among variables.

4.2 Univariate non-graphical EDA

The data that come from making a particular measurement on all of the subjects in a sample represent our observations for a single characteristic, such as age, gender, speed at a task, or response to a stimulus. We should think of these measurements as representing a "sample distribution" of the variable, which in turn more or less represents the "population distribution" of the variable. The usual goal of univariate non-graphical EDA is to better appreciate the "sample distribution" and also to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution. Outlier detection is also a part of this analysis.

4.2.1 Categorical data

The characteristics of interest for a *categorical* variable are simply the range of values and the frequency (or relative frequency) of occurrence for each value. (For ordinal variables it is sometimes appropriate to treat them as quantitative variables using the techniques in the second part of this section.) Therefore, usually the most

useful univariate non-graphical techniques for categorical variables is some form of **tabulation** of the frequencies, usually along with calculation of the fraction (or percent) of data that falls in each category. For example, if we categorize subjects by College at Carnegie Mellon University as Dietrich, MCS, SCS and "other", then there is a true population of all students enrolled in the 2019 Fall semester. If we take a random sample of 20 students for the purposes of performing a memory experiment, we could list the sample "measurements" as Dietrich, Dietrich, MCS, other, other, SCS, MCS, other, Dietrich, MCS, SCS, SCS, other, MCS, MCS, Dietrich, MCS, other, Dietrich, SCS. Our EDA would look like this:

Statistic/College	**Dietrich**	**MCS**	**SCS**	**other**	**Total**
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

Note that it is useful to have the total count (frequency) to verify that we have an observation for each subject that we recruited. (Losing track of data is a common mistake, and EDA is very helpful for finding mistakes.). Also, we should expect that the proportions add up to 1.00 (or 100%) if we are calculating them correctly (count/total). Once you get used to it, you won't need both proportion (relative frequency) and percent, because they will be interchangeable in your mind.

A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.

4.2.2 Characteristics of quantitative data

Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.

The characteristics of the population distribution of a *quantitative* variable are its center, spread, modality (number of peaks in the pdf), shape (including "heaviness of the tails"), and outliers. (See Section 3.5.) Our observed data represent

just one sample out of an infinite number of possible samples. *The characteristics of our randomly observed sample are not inherently interesting, except to the degree that they represent the population that it came from.*

The **sample** of measurements that we observe for a particular variable and a particular experiment is the "sample distribution". We need to recognize that the sample distribution would very likely be different if we were to repeat the same experiment, due to selection of a different random sample, a different treatment randomization, and different random (incompletely controlled) experimental conditions. In addition, we can calculate "sample statistics" from the data, such as sample mean, sample variance, sample standard deviation, sample skewness and sample kurtosis. These again would vary for each repetition of the experiment, so they don't represent any deep truth, but rather represent some uncertain information about the underlying population distribution and its parameters, which are what we really care about.

Many of the sample's distributional characteristics are seen qualitatively in the univariate graphical EDA technique of a histogram (see Section 4.3.1). In most situations it is worthwhile to think of univariate non-graphical EDA as telling you about aspects of the histogram of the variable of interest. Again, these aspects are quantitative, but because they refer to just one of many possible samples from a population, they are best thought of as random (non-fixed) estimates of the fixed, unknown parameters (see Section 3.5) of the population distribution of interest.

If the quantitative variable does not have too many distinct values, a tabulation, as we used for categorical data, will be a worthwhile univariate, non-graphical technique. But mostly, for quantitative variables we are concerned with the numeric (non-graphical) measures which are the various **sample statistics**. In fact, sample statistics are generally thought of as estimates of the corresponding population parameters.

Figure 4.1 shows a histogram of a sample of size 200 from the infinite population characterized by distribution **C** of Figure 3.1 from Section 3.5. Remember that in that section we examined the parameters that characterize theoretical (population) distributions. Now we are interested in learning what we can (but not everything, because parameters are "secrets of nature") about these parameters from measurements on a (random) sample of subjects out of that population.

The bi-modality is visible, as is an **outlier** at X=-2. There is no generally recognized formal definition for outlier, but roughly it means values that are unlikely to occur according to a particular distribution (but nonetheless have occurred).

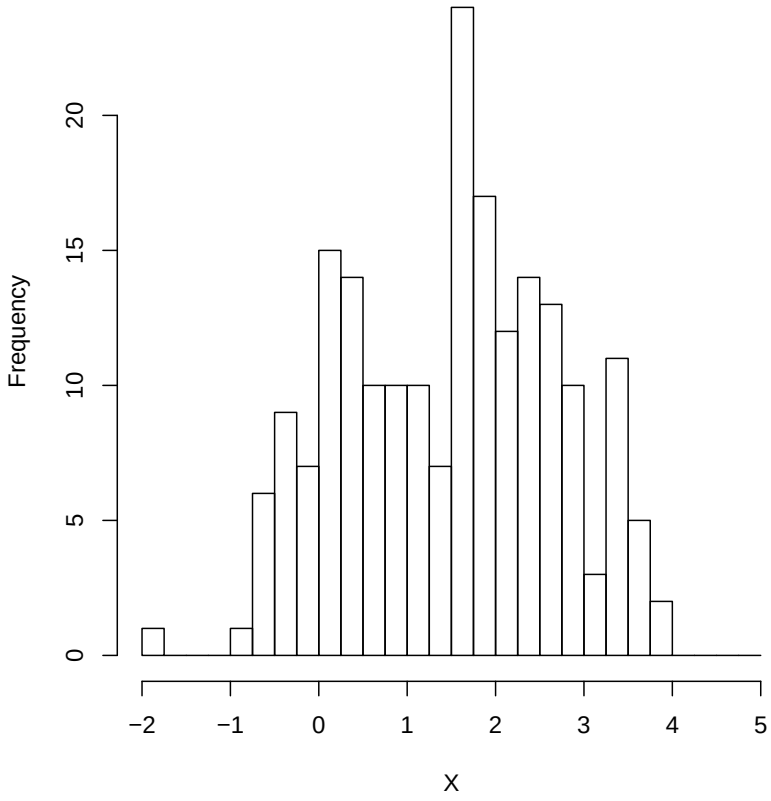


Figure 4.1: Histogram from distribution C.

This can also be thought of as sample data values which correspond to areas of the population pdf (or pmf) with low density (or probability). The definition of "outlier" for standard boxplots is described below (see 4.3.2). Another common definition of "outlier" considers any point more than a fixed number of standard deviations from the mean to be an "outlier", but these and other definitions are arbitrary and vary from situation to situation.

For quantitative variables (and possibly for ordinal variables) it is worthwhile to look at the central tendency, spread, skewness, and kurtosis the sample distribution. *But for categorical variables, none of these make any sense.* In what follows, we describe each of these four measures for quantitative variables in detail.

4.2.3 Central tendency

The **central tendency** or "location" of a distribution has to do with typical or middle values. The common, useful measures of central tendency are the statistics called (arithmetic) mean, median, and sometimes mode. Occasionally other means such as geometric, harmonic, truncated, or Winsorized means are used as measures of centrality. While most authors use the term "average" as a synonym for arithmetic mean, some use average in a broader sense to also include geometric, harmonic, and other means.

Assuming that we have n data values labeled x_1 through x_n, the formula for calculating the sample (arithmetic) **mean** is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The arithmetic mean is simply the sum of all of the data values divided by the number of values. It can be thought of as how much each subject gets in a "fair" re-division of whatever the data are measuring. For instance, the mean amount of money that a group of people have is the amount each would get if all of the money were put in one "pot", and then the money was redistributed to all people evenly. I hope you can see that this is the same as "summing then dividing by n".

For any symmetrically shaped distribution (i.e., one with a symmetric histogram or pdf or pmf) the mean is the point around which the symmetry holds. For non-symmetric distributions, the mean is the "center of mass" of probability distributions: If the histogram is cut out of some homogeneous stiff material such as cardboard, it will balance on a fulcrum placed at the mean.

For many descriptive quantities, there are both a sample and a population version. For a fixed finite population or for a theoretic infinite population described by a pmf or pdf, there is a single population mean which is a fixed, often unknown, value called the mean **parameter** (see Section 3.5). On the other hand, the "sample mean" will vary from sample to sample as different samples are taken, and so it is a random variable. The probability distribution of the sample mean is referred to as its **sampling distribution**. This term expresses the idea that any experiment could (at least theoretically, given enough resources) be repeated many times and various statistics such as the sample mean can be calculated each time, thus producing a distribution of the statistic across samples (i.e., a sampling distribution). However, realistically, we're only going to run an experiment once—we're not going to repeat it many times. This is where probability theory kicks in: We can often use probability theory to work out the exact distribution of the sample statistic, at least under certain assumptions, without having to repeat the experiment many times.

The **median** is another measure of central tendency. The sample median is the middle value after all of the values are put in an ordered list. If there are an even number of values, take the average of the two middle values. (If there are ties at the middle, some special adjustments are made by the statistical software we will use. In unusual situations for discrete random variables, there may not be a unique median.)

For symmetric distributions, the mean and the median coincide. For unimodal skewed (asymmetric) distributions, the mean is farther in the direction of the skewness (or "pulled out tail") of the distribution than the median is. Therefore, for many cases of skewed distributions, the median is preferred as a measure of central tendency. For example, according to the US Census Bureau 2004 Economic Survey, the median income of US families, which represents the income above and below which half of families fall, was $43,318. This seems a better measure of central tendency than the mean of $60,828, which indicates how much each family would have if we all shared equally. And the difference between these two numbers is quite substantial. Whether the mean or the median is the more appropriate statistic of interest will depend on the application at hand.

The median has a very special property called **robustness**. A sample statistic is "robust" if moving some data tends not to change the value of the statistic. The median is highly robust, because you can move nearly all of the upper half and/or lower half of the data values any distance away from the median without

changing the median. In contrast, a few very high or low values can greatly change the mean, making it less robust than the median. To put it another way, robust statistics are less sensitive to outliers or other extremes.

A rarely used measure of central tendency is the **mode**, which is the most likely or frequently occurring value. More commonly we simply use the term "mode" when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal). In symmetric, unimodal distributions, the mode equals both the mean and the median. In unimodal, skewed distributions the mode is on the other side of the median from the mean. In multi-modal distributions there is either no unique highest mode, or the highest mode may well be unrepresentative of the central tendency.

The most common measure of central tendency is the mean. For skewed distribution or when there is concern about outliers, the median may be preferred.

4.2.4 Spread

Several statistics are commonly used as a measure of the **spread** of a distribution, including variance, standard deviation, and interquartile range. Spread is an indicator of how far away from the center we are still likely to find data values.

The **variance** is a standard measure of spread. It is calculated for a list of numbers, e.g., the n observations of a particular measurement labeled x_1 through x_n, based on the n **sample deviations** (or just "deviations"). Then for any data value, x_i, the corresponding deviation is $(x_i - \bar{x})$, which is the signed (- for lower and + for higher) distance of the data value from the mean of all of the n data values. It is not hard to prove that the sum of all of the deviations of a sample is zero.

The variance of a population is defined as the mean squared deviation (see Section 3.5.2). The sample formula for the variance of observed data conventionally has $n-1$ in the denominator instead of n to achieve the property of "unbiasedness", which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here, σ^2). The most commonly used symbol for sample variance is s^2,

and the formula is

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

which is essentially the average of the squared deviations, except for dividing by $n-1$ instead of n. This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets. (In most cases, squaring is better than taking the absolute value because it puts special emphasis on highly deviant values.) As usual, a sample statistic like s^2 is best thought of as a characteristic of a particular sample (thus varying from sample to sample) which is used as an estimate of the single, fixed, true corresponding parameter value from the population, namely σ^2.

Another (equivalent) way to write the variance formula, which is particularly useful for thinking about ANOVA is

$$s^2 = \frac{\mathrm{SS}}{\mathrm{df}}$$

where SS is "sum of squared deviations", often loosely called "sum of squares", and df is "degrees of freedom" (see Section 3.10).

Because of the square, variances are always non-negative, and they have the somewhat unusual property of having squared units compared to the original data. So if the random variable of interest is a temperature in degrees, the variance has units "degrees squared", and if the variable is area in square kilometers, the variance is in units of "kilometers to the fourth power".

Variances have the very important property that they are additive for any number of different independent sources of variation. For example, the variance of a measurement which has subject-to-subject variability, environmental variability, and quality-of-measurement variability is equal to the sum of the three variances. This property is not shared by the "standard deviation".

The **standard deviation** is simply the square root of the variance. Therefore it has the same units as the original data, which helps make it more interpretable. The sample standard deviation is usually represented by the symbol s. For a theoretical Gaussian distribution, we learned in the previous chapter that mean plus or minus 1, 2 or 3 standard deviations holds 68.3, 95.4 and 99.7% of the probability respectively, and this should be approximately true for real data from a Normal distribution.

The variance and standard deviation are two useful measures of spread. The variance is the mean of the squares of the individual deviations. The standard deviation is the square root of the variance. For Normally distributed data, approximately 95% of the values lie within 2 sd of the mean.

A third measure of spread is the **interquartile range**. To define IQR, we first need to define the concepts of **quartiles**. The quartiles of a population or a sample are the three values which divide the distribution or observed data into even fourths. So one quarter of the data fall below the first quartile, usually written Q1; one half fall below the second quartile (Q2); and three fourths fall below the third quartile (Q3). The astute reader will realize that half of the values fall above Q2, one quarter fall above Q3, and also that Q2 is a synonym for the median. Once the quartiles are defined, it is easy to define the IQR as $IQR = Q3 - Q1$. By definition of quartiles, half of the data fall between Q1 and Q3. Thus, half of the data fall within an interval whose width equals the IQR. If the data are more spread out, then the IQR tends to increase, and vice versa.

The IQR is a more robust measure of spread than the variance or standard deviation. Any number of values in the top or bottom quarters of the data can be moved any distance from the median without affecting the IQR at all. More practically, a few extreme outliers have little or no effect on the IQR.

In contrast to the IQR, the **range** of the data is not very robust at all. The range of a sample is the distance from the minimum value to the maximum value: range = maximum - minimum. If you collect repeated samples from a population, the minimum, maximum and range tend to change drastically from sample to sample, while the variance and standard deviation change less, and the IQR least of all. The minimum and maximum of a sample may be useful for detecting outliers, especially if you know something about the possible reasonable values for your variable. They often (but certainly not always) can detect data entry errors such as typing a digit twice or transposing digits (e.g., entering 211 instead of 21 and entering 19 instead of 91 for data that represents ages of senior citizens.)

The IQR has one more property worth knowing: for normally distributed data *only*, the IQR approximately equals 4/3 times the standard deviation. This means that for Gaussian distributions, you can approximate the sd from the IQR by calculating 3/4 of the IQR.

The interquartile range (IQR) is a robust measure of spread.

4.2.5 Skewness and kurtosis

Two additional useful univariate descriptors are the skewness and kurtosis of a distribution. Skewness is a measure of asymmetry. Kurtosis is a measure of "peakedness" relative to a Gaussian shape. Sample estimates of skewness and kurtosis are taken as estimates of the corresponding population parameters (see section 3.5.3). If the sample skewness and kurtosis are calculated along with their standard errors, we can roughly make conclusions according to the following table where e is an estimate of skewness and u is an estimate of kurtosis, and $SE(e)$ and $SE(u)$ are the corresponding standard errors.

Skewness (e) or kurtosis (u)	**Conclusion**
$-2\text{SE(e)} < e < 2\text{SE(e)}$	not skewed
$e \leq -2\text{SE(e)}$	negative skew
$e \geq 2\text{SE(e)}$	positive skew
$-2\text{SE(u)} < u < 2\text{SE(u)}$	not kurtotic
$u \leq -2\text{SE(u)}$	negative kurtosis
$u \geq 2\text{SE(u)}$	positive kurtosis

For a positive skew, values far above the mode are more common than values far below, and the reverse is true for a negative skew. When a sample (or distribution) has positive kurtosis, then compared to a Gaussian distribution with the same variance or standard deviation, values far from the mean (or median or mode) are more likely, and the shape of the histogram is peaked in the middle, but with fatter tails. For a negative kurtosis, the peak is sometimes described has having "broader shoulders" than a Gaussian shape, and the tails are thinner, so that extreme values are less likely.

Skewness is a measure of asymmetry. Kurtosis is a more subtle measure of peakedness, relative to a Gaussian distribution.

4.3 Univariate graphical EDA

If we are focusing on data for a single variable on n subjects, i.e., a sample of size n, then in addition to looking at the various sample statistics discussed in the previous section, we also need to look graphically at the distribution of the sample. Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data; therefore, graphical methods, which are more qualitative and involve a degree of subjective analysis, are also useful.

4.3.1 Histograms

The most basic graph is the **histogram**, which includes bars that represent the frequency (count) or proportion (count/total count) of cases for a range of values. Typically the bars run vertically with the count (or proportion) axis running vertically. To manually construct a histogram, define the range of data for each bar (called a **bin**), count how many cases fall in each bin, and draw the bars high enough to indicate the count. For the simple data set found in EDA1.dat the histogram is shown in Figure 4.2. Besides getting the general impression of the shape of the distribution, you can read off facts like "there are two cases with data values between 1 and 2" and "there are 9 cases with data values between 2 and 3". Generally, values that fall exactly on the boundary between two bins are put in the lower bin, but this rule is not always followed.

Generally you will choose between about 5 and 30 bins, depending on the amount of data and the shape of the distribution. Of course, you need to see the histogram to know the shape of the distribution, so this may be an iterative process. It is often worthwhile to try a few different bin sizes/numbers because, especially with small samples, there may sometimes be a different shape to the histogram when the bin size changes. But usually the difference is small. Figure 4.3 shows three histograms of the same sample from a bimodal population using three different bin widths (5, 2 and 1). If you want to try on your own, the data are in EDA2.dat. The top panel appears to show a unimodal distribution. The middle panel correctly shows the bimodality. The bottom panel incorrectly suggests many modes. There is some art to choosing bin widths, and although often the automatic choices of a program like `R` are pretty good, they are certainly not always adequate.

It is very instructive to look at multiple samples from the same population to

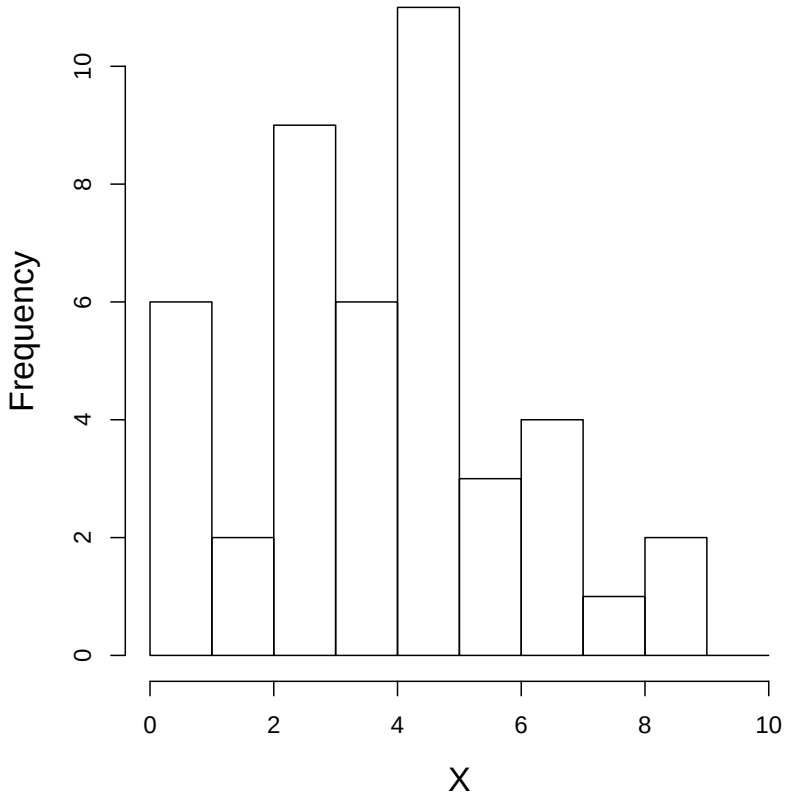


Figure 4.2: Histogram of EDA1.dat.

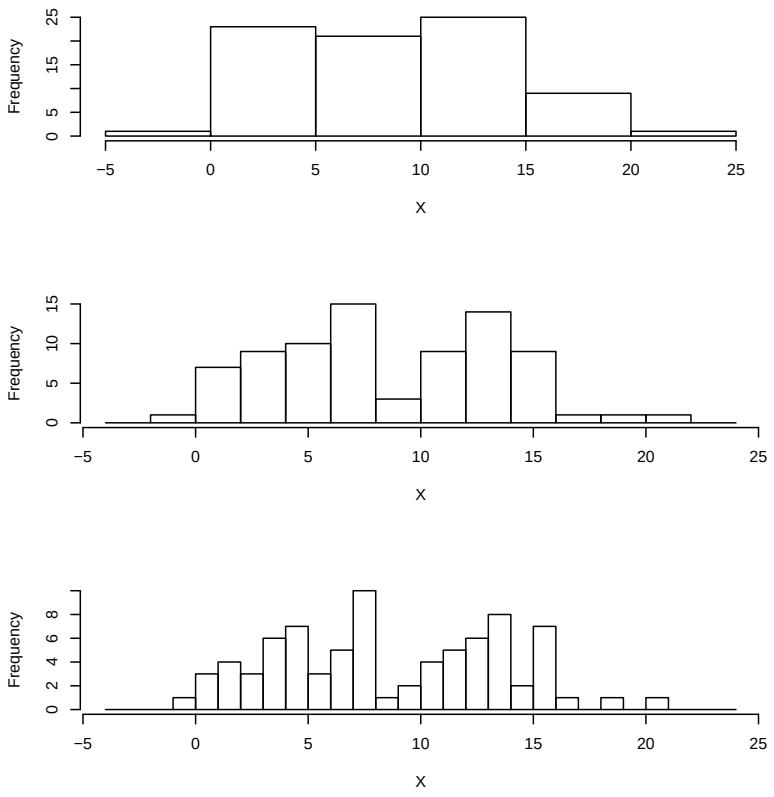


Figure 4.3: Histograms of EDA2.dat with different bin widths.

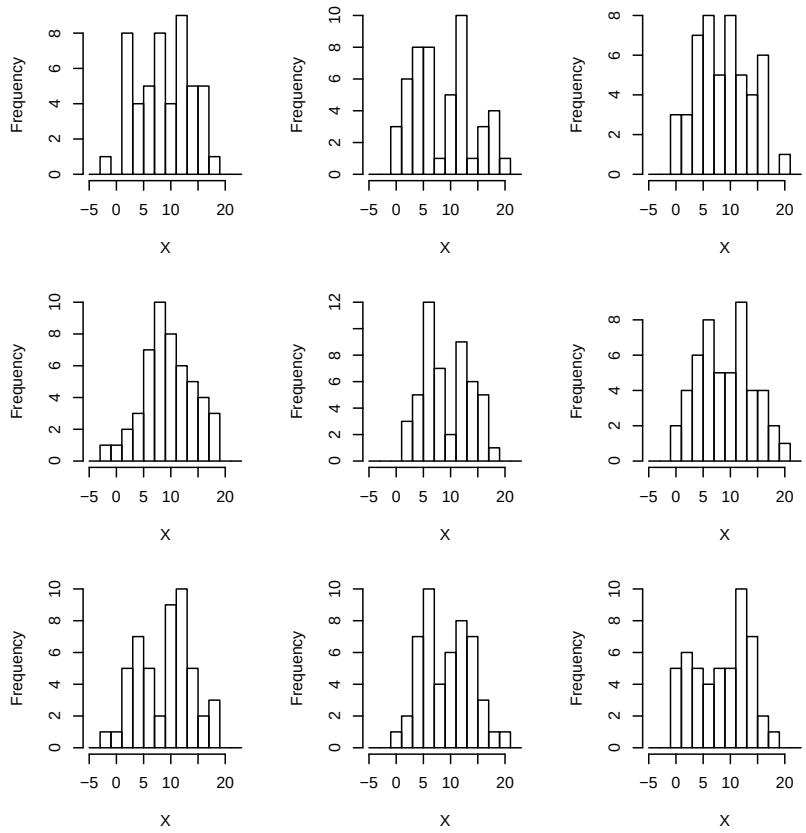


Figure 4.4: Histograms of multiple samples of size 50.

get a feel for the variation that will be found in histograms. Figure 4.4 shows histograms from multiple samples of size 50 from the same population as Figure 4.3, while 4.5 shows samples of size 100. Notice that the variability is quite high, especially for the smaller sample size, and that an incorrect impression (particularly of unimodality) is quite possible, just by the bad luck of taking a particular sample. In other words, a particular sample may not necessarily be representative of your population of interest, even when the sample is taken completely at random.

With practice, histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

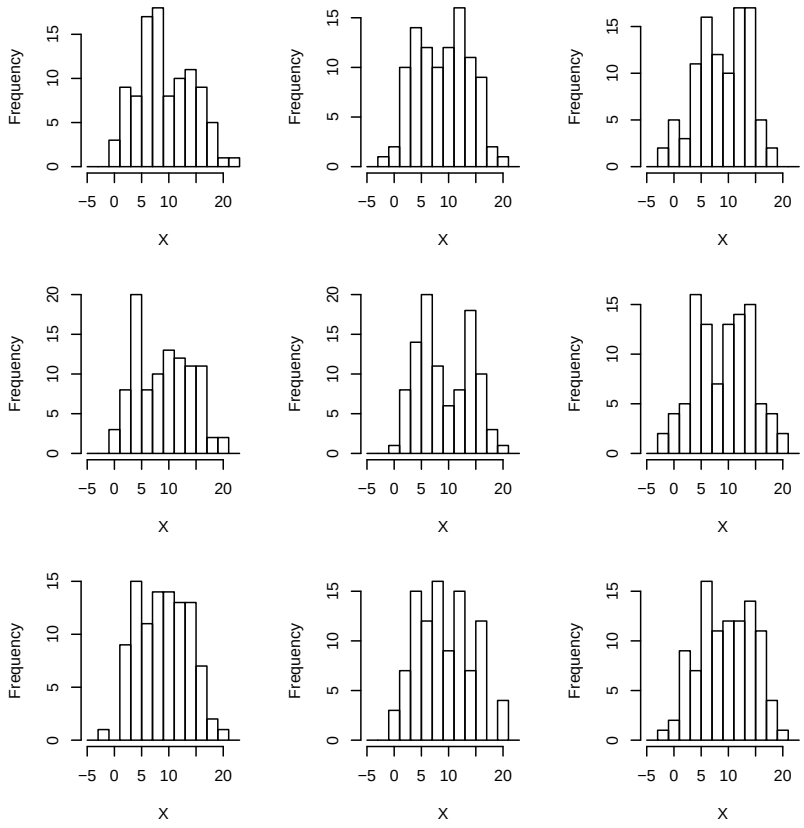


Figure 4.5: Histograms of multiple samples of size 100.

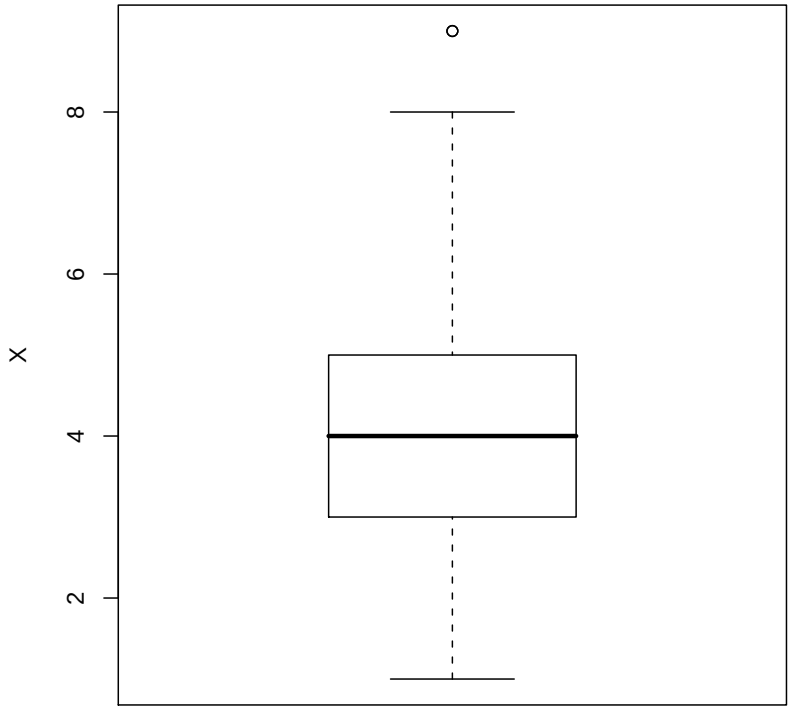


Figure 4.6: A boxplot of the data from EDA1.dat.

4.3.2 Boxplots

Another very useful univariate graphical technique is the **boxplot**. The boxplot will be described here in its vertical format, which is the most common, but a horizontal format also is possible. An example of a boxplot is shown in Figure 4.6, which again represents the data in EDA1.dat.

Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers, although they can be misleading about aspects such as multimodality. One of the best uses of boxplots is in the form of side-by-side boxplots (which we'll see in the multivariate graphical EDA section, Section 4.5).

Figure 4.7 is an annotated version of Figure 4.6. Here you can see that the boxplot consists of a rectangular box bounded above and below by "hinges" that

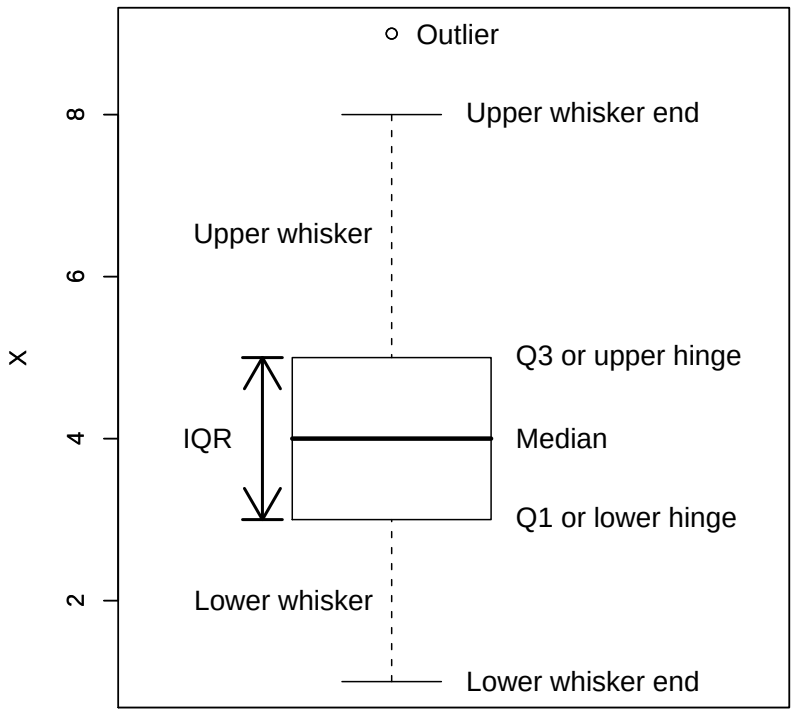


Figure 4.7: Annotated boxplot.

represent the quartiles Q3 and Q1 respectively, and with a horizontal "median" line through it. You can also see the upper and lower "whiskers", and a point marking an "outlier". The vertical axis is in the units of the quantitative variable.

Let's assume that the subjects for this experiment are hens and the data represent the number of eggs that each hen laid during the experiment. We can read certain information directly off of the graph. The median (**not mean!**) is 4 eggs, so no more than half of the hens laid more than 4 eggs and no more than half of the hens laid less than 4 eggs. (This is based on the technical definition of median; we would usually claim that half of the hens lay more than 4 eggs and half lay less than 4 eggs, knowing that this may be only approximately correct.) We can also state that one quarter of the hens lay less than 3 eggs and one quarter lay more than 5 eggs (again, this may not be exactly correct, particularly for small samples or a small number of different possible values). This leaves half of the hens, called the "central half", to lay between 3 and 5 eggs, so the interquartile range (IQR) is Q3-Q1=5-3=2.

The interpretation of the whiskers and outliers is slightly more complicated. Any data value more than 1.5 IQRs beyond its corresponding hinge in either direction is considered an "outlier" and is individually plotted. Sometimes values beyond 3.0 IQRs are considered "extreme outliers" and are plotted with a different symbol. In this boxplot, a single outlier is plotted corresponding to 9 eggs laid, although we know from Figure 4.2 that there are actually two hens that laid 9 eggs. This demonstrates a general problem with plotting whole number data, namely that multiple points may be superimposed, giving a wrong impression. (Jittering, circle plots, and star plots are examples of ways to correct this problem.) This is one reason why, e.g., combining a tabulation and/or a histogram with a boxplot is better than either alone.

Each whisker is drawn out to the most extreme data point that is less than 1.5 IQRs beyond the corresponding hinge. Therefore, the whisker ends correspond to the minimum and maximum values of the data *excluding* the "outliers".

Important: The term "outlier" is not well defined in statistics, and the definition varies depending on the purpose and situation. The "outliers" identified by a boxplot, which could be called "boxplot outliers" are defined as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1. This *does not* by itself indicate a problem with those data points. Boxplots are an exploratory technique, and you should consider designation as a boxplot outlier as just a suggestion that the points might be unusual. It is also important to realize that the number of

boxplot outliers depends strongly on the size of the sample. In fact, for data that are perfectly Normally distributed, we expect 0.70 percent (or about 1 in 150 cases) to be "boxplot outliers", with approximately half in either direction.

The boxplot information described above could be appreciated almost as easily if given in non-graphical format. The boxplot is useful because, with practice, all of the above and more can be appreciated at a quick glance. The additional things you should notice on the plot are the symmetry of the distribution and possible evidence of "fat tails". Symmetry is appreciated by noticing if the median is in the center of the box and if the whiskers are the same length as each other. For this purpose, as usual, the smaller the dataset the more variability you will see from sample to sample, particularly for the whiskers. In a skewed distribution we expect to see the median pushed in the direction of the shorter whisker. If the longer whisker is the top one, then the distribution is positively skewed (or skewed to the right, because higher values are on the right in a histogram). If the lower whisker is longer, the distribution is negatively skewed (or left skewed.) In cases where the median is closer to the longer whisker it is hard to draw a conclusion.

The term **fat tails** is used to describe the situation where a histogram has a lot of values far from the mean relative to a Gaussian distribution. This corresponds to positive kurtosis. In a boxplot, many outliers (more than the 1/150 expected for a Normal distribution) suggests fat tails (positive kurtosis), or possibly many data entry errors. Also, short whiskers suggest negative kurtosis, at least if the sample size is large.

Boxplots are excellent EDA plots because they rely on robust statistics like median and IQR rather than more sensitive ones such as mean and standard deviation. With boxplots it is easy to compare distributions (usually for one variable at different levels of another; see multivariate graphical EDA in Section 4.5) with a high degree of reliability because of the use of these robust statistics.

Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

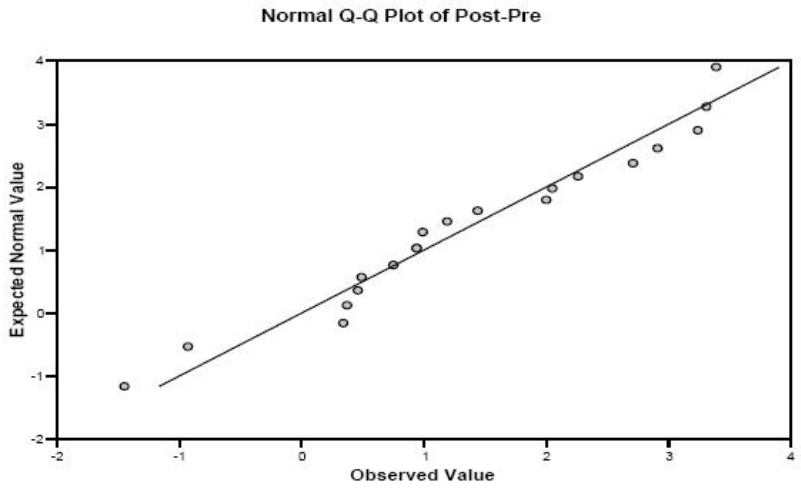


Figure 4.8: A quantile-normal plot.

4.3.3 Quantile-normal plots

The final univariate graphical EDA technique is the most complicated. It is called the **quantile-normal or QN plot** or more generality the **quantile-quantile or QQ plot**. It is used to see how well a particular sample follows a particular theoretical distribution. Although it can be used for any theoretical distribution, we will limit our attention to seeing how well a sample of data of size n matches a Gaussian distribution with mean and variance equal to the sample mean and variance. By examining the quantile-normal plot we can detect left or right skew, positive or negative kurtosis, and bimodality.

The example shown in Figure 4.8 shows 20 data points that are approximately normally distributed. **Do not confuse a quantile-normal plot with a simple scatter plot of two variables.** The title and axis labels are strong indicators that this is a quantile-normal plot. For many computer programs, the word "quantile" is also in the axis labels.

The main purpose of QN plots is to assess if a particular variable is approximately Normally distributed. Many statistical tests have the assumption that the outcome for any fixed set of values of the explanatory variables is approximately Normally distributed, and that is why QN plots are useful: If the assumption is grossly violated, the p-value and confidence intervals of those tests are likely inappropriate for the data at hand, suggesting that other statistical analysis tools

should be used. Relatedly, as we will see in the ANOVA and regression chapters, the most important situation where we use a QN plot to assess Normality assumptions is when we examine something called "residuals" (see Section 9.4). For basic interpretation of the QN plot, you just need to be able to distinguish between the two situations of "OK" (points fall randomly around the line) versus "non-normality" (points follow a strong curved pattern rather than following the line).

If you are still curious, here is a description of how the QN plot is created. Understanding this will help to understand the interpretation, but is not required in this course. Note that some programs swap the x and y axes from the way described here, but the interpretation is similar for all versions of QN plots. Consider the 20 values observed in this study. They happen to have an observed mean of 1.37 and a standard deviation of 1.36. Ideally, 20 random values drawn from a distribution that has a true mean of 1.37 and sd of 1.36 have a perfect bell-shaped distribution and will be spaced so that there is equal area (probability) in the region around each value in the bell curve.

In Figure 4.9 the dotted lines divide the bell curve up into 20 equally probable zones, and the 20 points are at the probability mid-points of each zone. These 20 points, which are more tightly packed near the middle than in the ends, are used as the "Expected Normal Values" in the QN plot of our actual data.

In summary, the sorted observed data values are plotted against "Expected Normal Values", and a diagonal line is added to help direct the eye towards a perfect straight line on the quantile-normal plot that represents a perfect bell shape for the observed data. If the data fall exactly on this line, then the observed quantiles are exactly the quantiles we would expect to see from a Normal distribution.

The interpretation of the QN plot is given here. If the axes are reversed in the computer package you are using, you will need to correspondingly change your interpretation. If all of the points fall on or nearly on the diagonal line (with a random pattern), this tells us that a histogram of the variable will show a bell shaped (Normal or Gaussian) distribution.

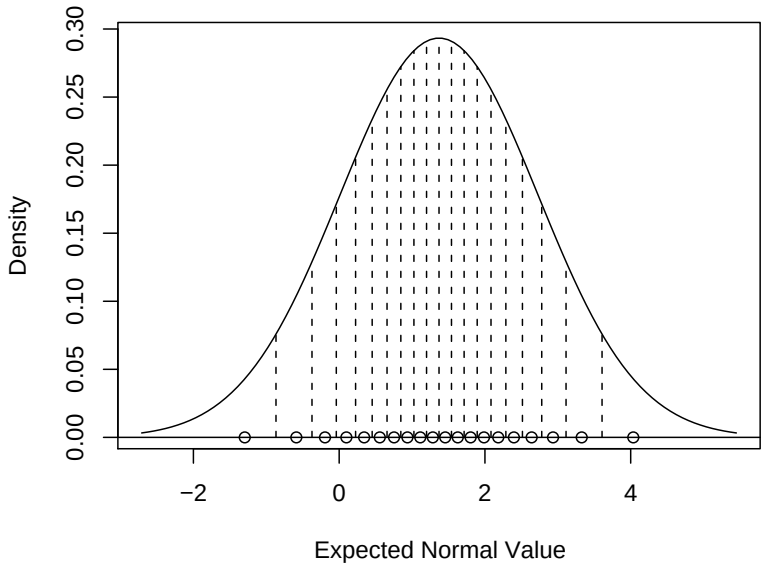


Figure 4.9: A way to think about QN plots.

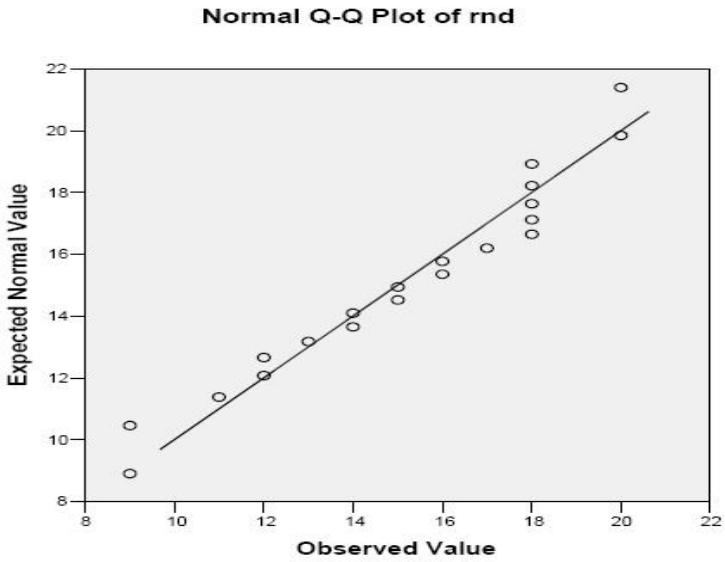


Figure 4.10: Quantile-normal plot with ties.

As an example, Figure 4.10 shows a QN plot of a toy dataset, where all of the points are basically on the reference line. However, we can see that there are several vertical bands of points; because the x-axis is "observed values", these bands indicate ties, i.e., multiple points with the same values (which all happen to be whole numbers). So, the data are discrete, meaning that we already know from the outset that the data are not *literally* Normally distributed. However, this QN plot suggests that these data appear to be at least approximately normally distributed.

In Figure 4.11 note that we have many points in a row that are on the same side of the line (rather than just bouncing around to either side), and that suggests that there is a real (non-random) deviation from Normality. The best way to think about these QN plots is to look at the low and high ranges of the Expected Normal Values. In each area, see how the observed values deviate from what is expected, i.e., in which direction the points deviate from the "perfect normal" line. Here we observe values that are too high in both the low and high ranges. So compared to a perfect bell shape, this distribution is pulled asymmetrically towards higher values, which indicates positive skew.

Also note that if you just *shift* a distribution to the right (without disturbing its symmetry) rather than skewing it, it will maintain its perfect bell shape, and the points remain on the diagonal reference line of the quantile-normal curve. This

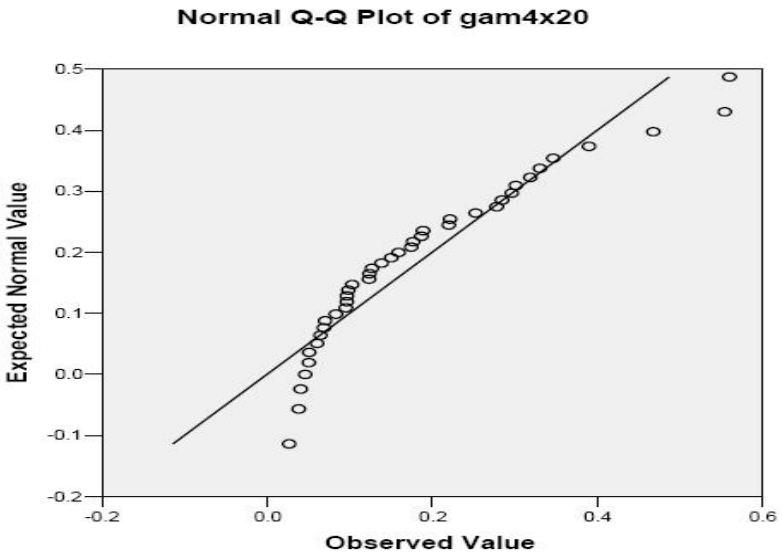


Figure 4.11: Quantile-normal plot showing right skew.

is because shifting a distribution doesn't change the distribution of its quantiles, which is what the QN plot displays.

Of course, we can also have a distribution that is skewed to the left, in which case the high and low range points are shifted (in the Observed Value direction) towards lower than expected values.

In Figure 4.12 the high end points are shifted too high and the low end points are shifted too low. These data show a positive kurtosis (fat tails). The opposite pattern is a negative kurtosis in which the tails are too "thin" to be bell shaped.

In Figure 4.13 there is a single point that is off the reference line, i.e. shifted to the right of where it should be. (Remember that the pattern of locations on the Expected Normal Value axis is fixed for any sample size, and only the position on the Observed axis varies depending on the observed data.) This pattern shows nearly Gaussian data with one "high outlier".

Finally, Figure 4.14 looks a bit similar to the "skew left" pattern, but the most extreme points tend to return to the reference line. This pattern is seen in bi-modal data, e.g., this is what we would see if we would mix strength measurements from controls and muscular dystrophy patients.

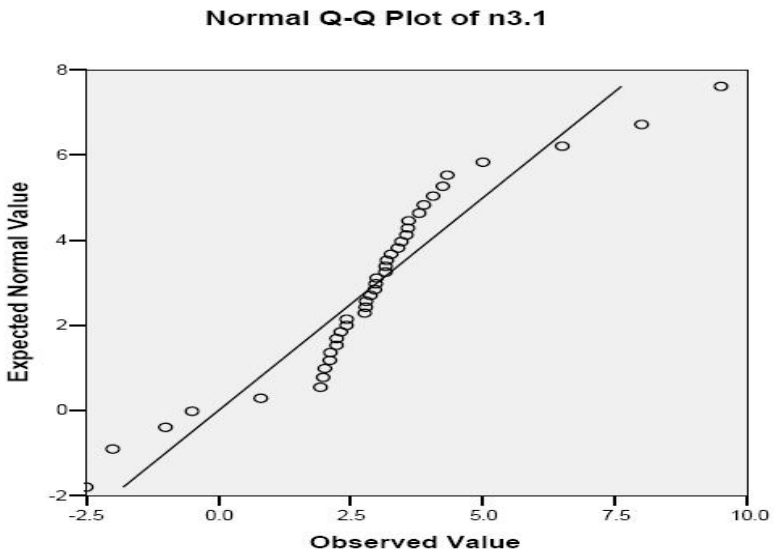


Figure 4.12: Quantile-normal plot showing fat tails.

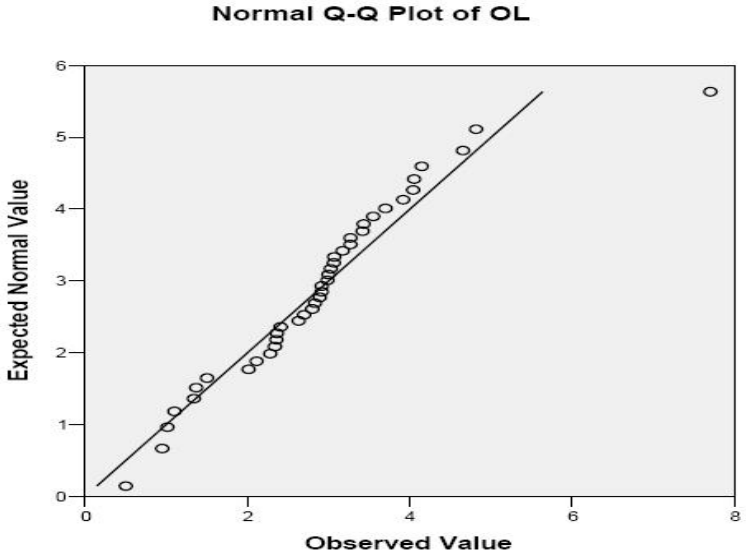


Figure 4.13: Quantile-normal plot showing a high outlier.

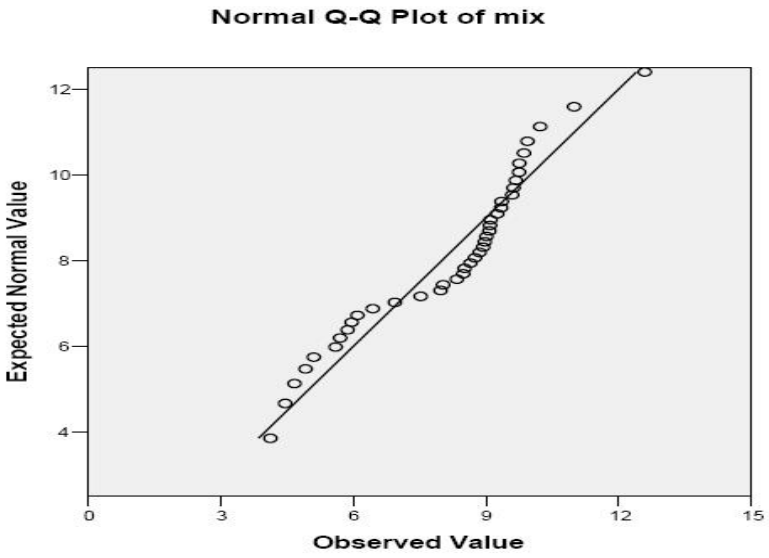


Figure 4.14: Quantile-normal plot showing bimodality.

Quantile-Normal plots allow detection of non-normality and diagnosis of skewness and kurtosis.

4.4 Multivariate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

4.4.1 Cross-tabulation

For categorical data (and quantitative data with only a few different values) an extension of tabulation called **cross-tabulation** is very useful. For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels. The two variables might be both explanatory, both outcome, or one of

each. Depending on the goals, row percentages (which add to 100% for each row), column percentages (which add to 100% for each column) and/or cell percentages (which add to 100% over all cells) are also useful.

Here is an example of a cross-tabulation. Consider the data in table 4.1. For each subject we observe sex (viewed through a binary lens) and age as categorical variables.

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Table 4.1: Sample Data for Cross-tabulation

Table 4.2 shows the cross-tabulation.

We can easily see that the total number of young females is 2, and we can calculate, e.g., the corresponding cell percentage is $2/11 \times 100 = 18.2\%$, the row percentage is $2/5 \times 100 = 40.0\%$, and the column percentage is $2/7 \times 100 = 28.6\%$.

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Table 4.2: Cross-tabulation of Sample Data

Cross-tabulation can be extended to three (and sometimes more) variables by making separate two-way tables for two variables at each level of a third variable. For example, we could make separate age by gender tables for different education levels.

Cross-tabulation is the most common bivariate non-graphical EDA technique.

4.4.2 Correlation for categorical data

Another statistic that can be calculated for two categorical variables is their correlation. But there are many forms of correlation for categorical variables, and that material is currently beyond the scope of this book.

4.4.3 Univariate statistics by category

For one categorical variable (usually explanatory) and one quantitative variable (usually outcome), it is common to produce some of the standard univariate non-graphical statistics for the quantitative variables separately for each level of the categorical variable, and then compare the statistics across levels of the categorical variable. Comparing the means is an informal version of ANOVA (or of a linear regression that uses only categorical explanatory variables), and comparing medians is a robust informal version of ANOVA. Comparing measures of spread is also a good informal test of the assumption of equal variances used in ANOVA and linear regression.

Especially for a categorical explanatory variable and a quantitative outcome variable, it is useful to produce a variety of univariate statistics for the quantitative variable at each level of the categorical variable.

4.4.4 Correlation and covariance

For two quantitative variables, the basic statistics of interest are the sample covariance and/or sample correlation, which correspond to and are estimates of the corresponding population parameters from Section 3.6.1. The sample covariance is a measure of how much two variables "co-vary", i.e., how much (and in what direction) should we expect one variable to change when the other changes. Note that, just like the sample mean, the sample covariance is just an average—specifically, it is the average of the product of two variables (instead of the average of a single variable). See the technical text below for more details.

Sample covariance is calculated by computing (signed) deviations of each measurement from the average of all measurements for that variable. Then the deviations for the two measurements are multiplied together separately for each subject. Finally these values are averaged (actually summed and divided by n-1, to keep the statistic unbiased). The general formula for sample covariance is

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Note that the units on sample covariance are the products of the units of the two variables. Furthermore, it is worth noting that $\text{Cov}(X, X) = \text{Var}(X)$.

Positive covariance values suggest that when one measurement is above the mean the other will probably also be above the mean, and vice versa. Negative covariances suggest that when one variable is above its mean, the other is below its mean. And covariances near zero suggest that the two variables vary independently of each other.

Technically, independence implies zero correlation, but the reverse is not necessarily true.

Covariances tend to be hard to interpret, so we often use correlation instead. Correlation is the standardized version of covariance, which is why it always falls between -1 and +1, with -1 being a "perfect" negative linear correlation, +1 being a perfect positive linear correlation, and 0 indicating that X and Y are uncorrelated. The symbol r or $r_{x,y}$ is often used for sample correlations.

The formula for the sample correlation is

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{s_x s_y}$$

where s_x is the standard deviation of X and s_y is the standard deviation of Y. From this we can see that correlation is a standardized version of covariance, because we are dividing the covariance (which is the average of a product) by the product of the standard deviations.

If you want to see a "manual example" of calculating the sample covariance and correlation, consider an example using the data in table 4.3. For each subject we observe age and a strength measure.

Table 4.4 shows the calculation of covariance. The mean age is 50 and the mean strength is 19, so we calculate the deviation for age as age-50 and deviation for strength and strength-19. Then we find the product of the deviations and add them up. This total is -1106, and since n=11, the covariance of x and y is -1106/10=-110.6. The fact that the covariance is negative indicates that as age goes up strength tends to go down (and vice versa).

In this example, $s_x = 18.96$, $s_y = 6.39$, so $r = \frac{-110.6}{18.96 \cdot 6.39} = -0.913$. This is a strong negative correlation.

4.4.5 Covariance and correlation matrices

When we have many quantitative variables the most common non-graphical EDA technique is to calculate all of the pairwise covariances and/or correlations and assemble them into a matrix. Note that the covariance of X with X is the variance of X and the correlation of X with X is 1. In equation form, this means

Subject ID	Age	Strength
GW	38	20
JA	62	15
TJ	22	30
JMA	38	21
JMO	45	18
JQA	69	12
AJ	75	14
MVB	38	28
WHH	80	9
JT	32	22
JKP	51	20

Table 4.3: Covariance Sample Data

$\text{Cov}(X, X) = \text{Var}(X)$ and $\text{Cor}(X, X) = 1$. For example, the covariance matrix of Table 4.5 tells us that the variances of X, Y, and Z are 5, 7, and 4 respectively, the covariance of X and Y is 1.77, the covariance of X and Z is -2.24, and the covariance of Y and Z is 3.17.

Similarly, the correlation matrix in Table 4.6 tells us that the correlation of X and Y is 0.3, the correlation of X and Z is -0.5. and the correlation of Y and Z is 0.6. Note that, in a covariance matrix, the variances can always be found on the diagonals, and the covariances are on the off-diagonals. Furthermore, covariance matrices and correlation matrices are always symmetric, because $\text{Cov}(X, Y) = \text{Cov}(Y, X)$. In other words, the (i, j) and (j, i) elements of these matrices will always be the same.

The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

Subject ID	Age	Strength	Age-50	Str-19	product
GW	38	20	-12	+1	-12
JA	62	15	+12	-4	-48
TJ	22	30	-28	+11	-308
JMA	38	21	-12	+2	-24
JMO	45	18	-5	-1	+5
JQA	69	12	+19	-7	-133
AJ	75	14	+25	-5	-125
MVB	38	28	-12	+9	-108
WHH	80	9	+30	-10	-300
JT	32	22	-18	+3	-54
JKP	51	20	+1	+1	+1
Total			0	0	-1106

Table 4.4: Covariance Calculation

	X	Y	Z
X	5.00	1.77	-2.24
Y	1.77	7.0	3.17
Z	-2.24	3.17	4.0

Table 4.5: A Covariance Matrix

4.5 Multivariate graphical EDA

In this section we will discuss how to graph the relationship between multiple random variables (usually two). We'll focus on the cases where (1) there is one categorical variable and one quantitative variable, and (2) there are two quantitative variables.

4.5.1 Univariate graphs by category

When we have one categorical (usually explanatory) and one quantitative (usually outcome) variable, graphical EDA usually takes the form of "conditioning" on the categorical random variable. This simply indicates that we focus on all of the

	X	Y	Z
X	1.0	0.3	-0.5
Y	0.3	1.0	0.6
Z	-0.5	0.6	1.0

Table 4.6: A Correlation Matrix

subjects with a particular level of the categorical random variable, then make plots of the quantitative variable for those subjects. We repeat this for each level of the categorical variable, then compare the plots. The most commonly used plot is a **side-by-side boxplots**, as in Figure 4.15. Here we see the data from EDA3.dat, which consists of strength data for each of three age groups. You can see the downward trend in the median as the ages increase. By looking at the lengths of the three boxes in the plot, we can see that the spreads (IQRs) are similar for the three groups. Furthermore, because the median lines are approximately in the middle of each box, all three groups are roughly symmetrical, with one high strength outlier in the youngest age group.

Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable.

4.5.2 Scatterplots

For two quantitative variables, the basic graphical EDA technique is the scatterplot, which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset. *If one variable is explanatory and the other is outcome, it is a very, very strong convention to put the outcome on the y (vertical) axis.*

One or two additional categorical variables can be accommodated on the scatterplot by encoding the additional information in the symbol type and/or color. This way, a scatterplot can actually show several dimensions (more than just two!) An example is shown in Figure 4.16. Age vs. strength is shown, and different col-

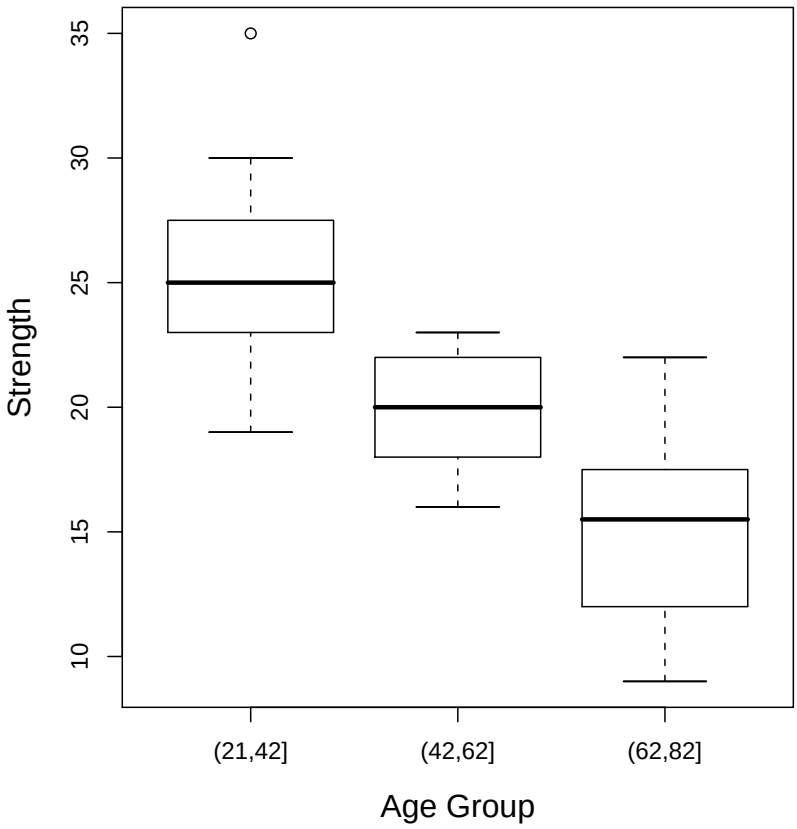


Figure 4.15: Side-by-side Boxplot of EDA3.dat.

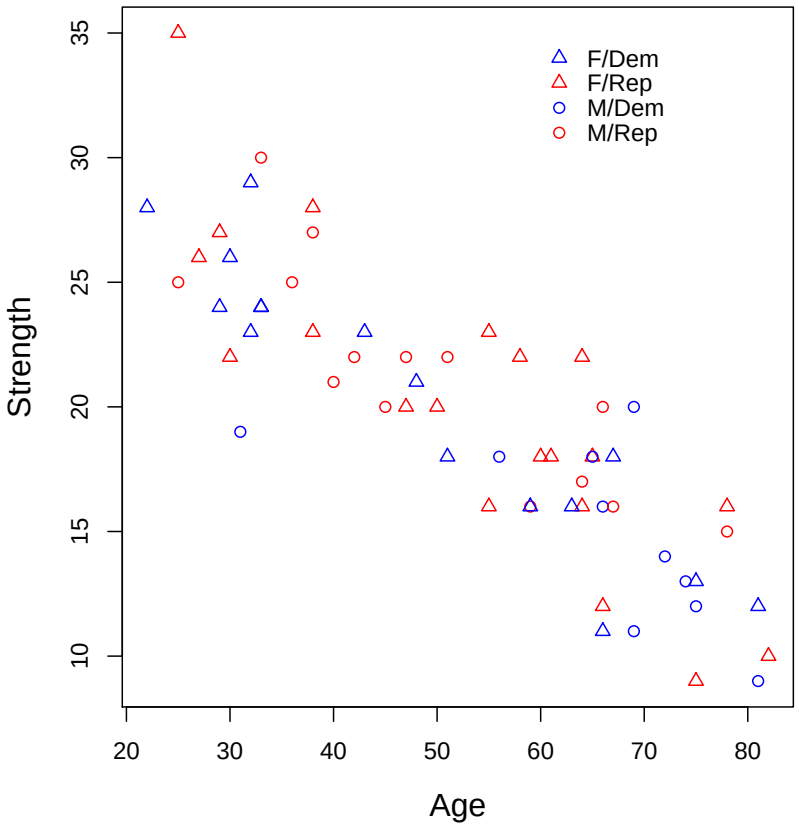


Figure 4.16: scatterplot with two additional variables.

ors and symbols are used to code political party and gender (both viewed through a binary lens). Thus, Figure 4.16 plots four dimensions.

In a nutshell: You should always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables. EDA is not an exact science – it is a very important art!

Chapter 5

The t-test and Basic Inference Principles

The t-test is used as an example of the basic principles of statistical inference.

One of the simplest situations for which we might design an experiment is the case of a nominal two-level explanatory variable and a quantitative outcome variable. This situation is quite ubiquitous: These types of experiments are commonly characterized as treatment-versus-control or Treatment A versus Treatment B experiments; the term A/B testing is also commonly used in the tech industry. Table 5.1 shows several examples. For all of these experiments, the treatments have two levels, and the treatment variable is nominal. Note in the table the various experimental units to which the two levels of treatment are being applied for these examples. If we randomly *assign* the treatments to these units, this will be a randomized experiment rather than an observational study, which justifies using the word "causes" rather than just "is associated with" to any statistically significant result. (This point will be discussed in depth in Chapter 7.) This chapter only discusses so-called "between subjects" explanatory variables, which means that we assume each experimental unit is exposed to only one of the two levels of treatment (even though that is not necessarily the most obvious way to run an experiment, such as the fMRI experiment in Table 5.1).

This chapter shows one way to perform statistical inference for the two-group, quantitative outcome experiment, namely the independent samples t-test. More importantly, the t-test is used as an example for demonstrating the basic principles

Experimental units	**Explanatory variable**	**Outcome variable**
people	placebo vs. vitamin C	time until the first cold symptoms
hospitals	control vs. enhanced hand washing	number of infections in the next six months
people	math tutor A vs. math tutor B	score on the final exam
people	neutral stimulus vs. fear stimulus	ratio of fMRI activity in the amygdala to activity in the hippocampus

Table 5.1: Some examples of experiments with a quantitative outcome and a nominal 2-level explanatory variable

of statistical inference that will be used throughout the book. Understanding these principles, along with some degree of theoretical underpinning, is key to using statistical results effectively. Among other things, you need to really understand what a p-value and a confidence interval tell us, and when they can and cannot be trusted. Because p-values and confidence intervals come up in most statistical analyses (not just t-tests), the topics we discuss in this chapter will give you a general picture of the statistical analysis of an experiment and a good foundation in the underlying theory. As usual, more advanced material—which will enhance your understanding but is not required for a fairly good understanding of the concepts—is shaded in gray.

Relatedly, in many ways, t-tests are connected to more complicated statistical methods that we'll examine throughout this book. For example, an alternative inferential procedure is one-way ANOVA, which always gives the same results as the t-test, and is the topic of the next chapter. More specifically, one-way ANOVA is an approach for comparing the means of $k \geq 2$ groups, and t-tests can be viewed as ANOVA with $k = 2$ groups. Furthermore, in this chapter and the ANOVA chapter, we assume that the only explanatory variable is the categorical treatment variable, but researchers often have other explanatory variables that they want to "control for" or "adjust for;" we discuss those methods (linear regression and ANCOVA) in Chapters 9 and 10. t-tests and ANOVA can be viewed as a special case of linear regression, where the only variable we are "controlling for" in the experiment is a categorical treatment variable.

As mentioned in the preface, it is hard to find a linear path for learning experimental design and analysis because so many of the important concepts are interdependent. For this chapter, we will assume that the subjects chosen to participate in the experiment are representative, and that each subject is randomly assigned to exactly one treatment. The reasons we should do these things and the consequences of not doing them are postponed until Chapter 7. For now, we will focus on the EDA and confirmatory analyses for a two-group between-subjects experiment with a quantitative outcome.

5.1 Case study from the field of Human-Computer Interaction (HCI)

This (fake) experiment is designed to determine which of two background colors for computer text is easier to read, as determined by the speed with which a task described by the text is performed. The study randomly assigns 35 university students to one of two versions of a computer program that presents text describing which of several icons the user should click on. The program measures how long it takes until the correct icon is clicked. This measurement is called "reaction time" and is measured in milliseconds (ms). The program reports the average time for 20 trials per subject. The two versions of the program differ in the background color for the text (yellow or cyan).

The data can be found in the file background.sav on this book's web data site. It is tab delimited with no header line and with columns for subject identification, background color, and response time in milliseconds. The coding for the color column is 0=yellow, 1=cyan. The data look like this:

Subject ID	Color	Time (ms)
NYP	0	859
⋮	⋮	⋮
MTS	1	1005

This dataset only has 35 rows (one for each subject) and three columns (one for each variable), but even for a dataset this small, it is hard to get a good idea of the differences in response time across the two colors just by looking at the numbers.

This motivates using exploratory data analyses (EDA) to better understand the data.

Here are some basic univariate EDA. There is no point in doing EDA for the subject IDs. For the categorical variable Color, the only useful non-graphical EDA is a tabulation of the two values. This can be done in R using the `table()` function:

```
> table(background$color)
 0  1
17 18
```

This shows us that 18 subjects were assigned to cyan and 17 were assigned to yellow. It's common for categorical variables to be coded with numbers, so it's very important to remember how your variables are coded (e.g., in this example, we had to remember that `color = 1` corresponds to cyan and `color = 0` corresponds to yellow).

Sometimes it's useful to look at group percentages instead of just group counts. This can be done using the `prop.table()` function in conjunction with `table()`:

```
> prop.table( table(background$color) )
        0         1
0.4857143 0.5142857
```

Other non-graphical exploratory analyses of Color - such as calculation of mean, variance, etc. - don't make sense because Color is a categorical variable. (It is possible to interpret the mean in this case because yellow is coded as 0 and cyan is coded as 1. The mean, 0.514, represents the fraction of cyan backgrounds.) For graphical EDA of the color variable you could make a pie or bar chart, but this really adds nothing to the simple 48.57% vs 51.43% breakdown.

For the quantitative variable Reaction Time, non-graphical EDA would include statistics like these:

```
#number of subjects
> nrow(background)
35
#mean of reaction time
> mean(background$time)
670.0286
#standard deviation of reaction time
> sd(background$time)
180.1518
#minimum and maximum values
> min(background$time); max(background$time)
```

```
291
1005
```

Here we can see that there are 35 reactions times that range from 291 to 1005 milliseconds, with a mean of 670.03 and a standard deviation of 180.15. (We can calculate that the variance is $180.15^2 = 32454.02$, and we could use the `summary()` function to calculate the median and IQR.) If we were to assume that the data follow a Normal distribution, then we could conclude that about 95% of the data fall within the mean plus or minus 2 sd, which is about 310 to 1030. But such an assumption is most likely incorrect, because if there is a difference in reaction times between the two colors, we would expect that the distribution of reaction times *ignoring color* would be some bimodal distribution that is a mixture of the two individual reaction time distributions for the two colors.

A histogram and/or boxplot of reaction time will further help you get a feel for the data. Ultimately, the purpose of this experiment is to assess if there is a difference in mean reaction time between (1) people who read text with a yellow background, and (2) people who read text with a cyan background. Thus, a natural thing to do is to compare the reaction times between the yellow group and the cyan group, which leads us to bivariate EDA (bivariate because we will be looking at the relationship between two variables: color and reaction time). For bivariate EDA, we want graphs and descriptive statistics for the quantitative outcome (dependent) variable Reaction Time broken down by the levels of the categorical explanatory variable (factor) Background Color. For example, the `aggregate()` function is a great way to look at summary statistics across different subgroups of the data:

```
> aggregate(time~color, data = background, FUN = summary)
color time.Min. time.1st Qu. time.Median time.Mean time.3rd Qu.
   time.Max.
  0  392.0000     559.0000    695.0000  679.6471     833.0000
   906.0000
  1  291.0000     475.7500    649.5000  660.9444     827.2500
   1005.0000
```

In words, the above code says, "Use the `summary` function for the `time` variable within different groups of the `color` variable." Already we can see that the cyan group has an average reaction time lower than the yellow group. We can also see this information graphically with a side-by-side boxplot, as shown in Figure . Here is the code used to make this boxplot:

```
> boxplot(time~color, data = background,
+   xlab = "Background Color", ylab = "Reaction Time (ms)",
```

```
+    names = c("Yellow", "Cyan"))
```

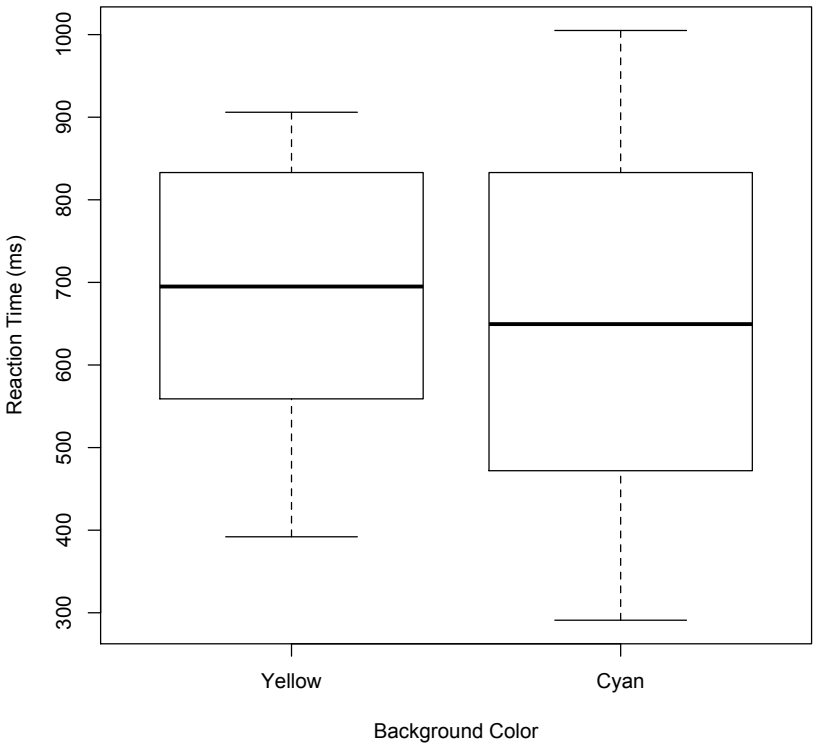


Figure 5.1: Boxplots of reaction time by color.

From these boxplots, we see that the cyan group has a lower median but a wider IQR, and the median line is approximately in the center of both boxes (suggesting symmetry of the distributions). In short, *within the sample*, the cyan group has shorter (but also more variable) reaction times, on average, as compared to the yellow group.

EDA for the two-group quantitative outcome experiment should include examination of sample statistics for mean, standard deviation, skewness, and kurtosis separately for each group, as well as boxplots and histograms.

Even though the cyan group measurements are on average lower than the yellow group measurements within the sample, that does not immediately suggest that it is definitely (or even likely) easier to read text with a cyan background than a yellow background. Indeed, for any random sample of subjects, there will *always* be some difference between the two groups. The question is if the difference observed in this sample is *significant*, and that's where statistical tools for inference will come in handy, as we discuss in the next section.

5.2 How classical statistical inference works

In this section you will see ways to think about the state of the real world at a level appropriate for scientific study, see how that plays out in experimentation, and learn how we match up the real world to the theoretical constructs of probability and statistics. In the next section you will see the details of how formal inference is carried out and interpreted.

How should we think about the real world with respect to a simple two group experiment with a continuous outcome? Obviously, if we were to repeat the entire experiment on a new set of subjects, we would (almost surely) get different results. The reasons that we would get different results are many, but they can be broken down into several main groups (see Section 7.5) such as measurement variability, environmental variability, treatment application variability, and subject-to-subject variability. The understanding that our experimental results are just one (random) set out of many possible sets of results is the foundation of statistical inference.

The key to standard (classical) statistical analysis is to consider what types of results we would expect if we repeated an experiment many times under specific conditions, and then compare the observed result to these hypothetical results and characterize how "typical" the observed result is.

5.2.1 The steps of statistical analysis

Most formal statistical analyses work like this:

1. Use your judgement to choose a model (mean and error components) that is a reasonable match for the data from the experiment. The model is expressed in terms of the population from which the subjects (and outcome variable) were drawn. Also, define parameters of interest.

2. Using the parameters, define a (point) null hypothesis and a (usually complex) alternative hypothesis which correspond to the scientific question of interest.

3. Choose (or invent) a statistic which has different distributions under the null and alternative hypotheses.

4. Calculate the null sampling distribution of the statistic.

5. Compare the observed (experimental) statistic to the null sampling distribution of that statistic to calculate a p-value for a specific null hypothesis (and/or use similar techniques to compute a confidence interval for a quantity of interest).

6. Perform some kind of assumption checks to validate the degree of appropriateness of the model assumptions (e.g., through EDA or formal statistical tests).

7. Use your judgement to interpret the statistical inference in terms of the underlying science.

Ideally there is one more step, which is the power calculation. This involves calculating the distribution of the statistic under one or more specific (point) alternative hypotheses *before* conducting the experiment so that we can assess the likelihood of getting a "statistically significant" result for various "scientifically significant" alternative hypotheses.

All of these points will now be discussed in more detail, both theoretically and using the HCI example. Focus is on the two group, quantitative outcome case, but the general principles apply to many other situations.

Classical statistical inference involves multiple steps, including definition of a model, definition of statistical hypotheses, selection of a statistic, computation of the sampling distribution of that statistic, computation of a p-value and/or confidence intervals, and interpretation of the statistical results as it relates to the science being studied.

5.2.2 Model and parameter definition

To start conducting statistical inference for the HCI experiment example, we will posit a model that characterizes the "data generating process" i.e., the process that generated the data we are analyzing. We will assume that the subjects are representative of some population of interest. In our two-treatment-group example, we most commonly consider the parameters of interest to be the population means of the outcome variable (true value without measurement error) for the two treatments, usually designated with the Greek letter mu (μ) and two subscripts. For now let's use μ_1 and μ_2, where in the HCI example μ_1 is the population mean of reaction time for subjects shown the yellow background and μ_2 is the population mean for those shown the cyan background. (A good alternative is to use μ_Y and μ_C, which are better mnemonically.)

It is helpful to think about the relationship between the treatment randomization and the population parameters in terms of **counterfactuals**. Although we have the measurement for each subject for the treatment (background color) to which they were assigned, there is also a "counterfactual" ("against the facts") result for the treatment they did not get. For example, in the HCI experiment, consider a subject who was assigned to the cyan color in the experiment. In this case, we *know* what the "cyan reaction time" was for that subject, but we *do not know* the "yellow reaction time" for that subject. Ideally, we could go back in time, assign that subject to read text with a yellow background, and see what happens. That way, we could know with certainty which treatment is better *for that subject* (in other words, we could obtain a random draw from $\mu_C - \mu_Y$, which is our main quantity of interest). However, for good or bad, we cannot go back in time: Ultimately, we only observe a random draw from μ_C but not μ_Y for that subject. The next best thing we can do is hope that the reaction times for subjects who were assigned to yellow background text are a good proxy for this counterfactual that

we can never know. In short, we hope that we can use random draws from μ_C and random draws from μ_Y to approximate random draws from $\mu_C - \mu_Y$ (which we can never obtain, because we can't assign individual subjects to cyan *and* yellow). When faced with these noisy hopes and approximates, it is important to keep in mind that we are most interested in unknowable population parameters, rather than sample statistics. Put another way, in essentially every experiment that we run, the sample means of the outcomes for the treatment groups will differ, *even if there is really no true difference between the outcome mean parameters for the two treatments in the population*, so focusing on those differences is not very meaningful.

It must be strongly emphasized that statistical inference is all about learning what we can about the (unknowable) population parameters and not about the sample statistics per se.

As mentioned in Section 1.2, a statistical model has two parts, the **structural model** and the **error model**. The structural model refers to defining the pattern of means for groups of subjects defined by explanatory variables, but it does not state what values these means take. In the case of the two group experiment, simply defining the population means (without making any claims about their equality or non-equality) defines the structural model. As we progress through the course, we will have more complicated structural models.

The error model (noise model) defines the variability of subjects "in the same group" around the mean for that group. The error model does not assume that we can predict the deviation of individual measurements from the group mean; rather, it assumes that the deviations follow some probability distribution.

For continuous outcome variables, the most commonly used error model is that for *each* treatment group the distribution of outcomes in the population is normally distributed, and furthermore that the population variances of the groups are equal. In addition, we assume that each error (deviation of an individual value from the group population mean) is statistically independent of every other error. The normality assumption is often approximately correct because (as stated in the CLT) the sum of many small non-Normal random variables will be normally distributed, and most outcomes of interest can be thought of as being affected in some additive way by many individual factors. On the other hand, it is not true

that *all* outcomes are normally distributed, so we need to check our assumptions before interpreting any formal statistical inferences.

The structural component of a statistical model defines the means of groups, while the error component describes the random pattern of deviation from those means.

5.2.3 Null and alternative hypotheses

The null and alternative hypotheses are statements about *the population parameters* that express different characterizations of the population (i.e., different ways of modeling the data generating process). Almost always the null hypothesis is a so-called point hypothesis in the sense that it defines a specific case (with an equal sign), and the alternative is a complex hypothesis in that it covers many different conditions with less than ($<$), greater than ($>$), or unequal ($\neq$) signs.

In the two-treatment-group case, the usual **null hypothesis** is that the two population means are equal, usually written as $H_0 : \mu_1 = \mu_2$, where the symbol H_0, read "H zero" or "H naught" indicates the null hypothesis. The null hypothesis usually denotes no treatment effect or no difference in treatment effects (i.e., it posits that null effects are occurring in the experiment).

In the two-treatment-group case, the usual **alternative hypothesis** is that the two population means are unequal, written as $H_1 : \mu_1 \neq \mu_2$ or $H_A : \mu_1 \neq \mu_2$ where H_1 or H_A are interchangeable symbols for the alternative hypothesis. (Occasionally, researchers may use an alternative hypothesis that states that one population mean is less than the other, but such a "one-sided hypothesis" should only be used when the opposite direction is truly impossible.) Note that there are really an infinite number of specific alternative hypotheses, e.g., $|\mu_0 - \mu_1| = 1$, $|\mu_0 - \mu_1| = 2$, etc. It is in this sense that the alternative hypothesis is complex, and this is an important consideration in power analysis, as we will discuss in detail in Chapter 11.

The null hypothesis specifies patterns of mean parameters corresponding to no difference in treatment effects, while the alternative hypothesis usually covers everything else.

5.2.4 Choosing a statistic

The next step is to find (or invent) a statistic that has a different distribution for the null and alternative hypotheses and for which we can calculate the null sampling distribution (see below). It is important to realize that the sampling distribution of the chosen statistic differs for each specific alternative, that there is almost always overlap between the null and alternative distributions of the statistic, and that the overlap is large for alternatives that reflect small effects and smaller for alternatives that reflect large effects.

For the two-treatment-group experiment with a quantitative outcome a commonly used statistic is the so-called "t" statistic which is the difference between the sample means (in either direction) divided by the (estimated) standard error (see below) of that difference. In general, when analyzing experiments, it is good to choose a statistic that measures some effect or difference of interest. For example, for the two-treatment-group experiment, we want to know if the two treatments are different (on average) in terms of the outcome. Thus, the difference in means is a very natural choice of statistic. In fact, under certain assumptions, it can be shown that this statistic is "optimal" (in terms of power), but a valid test does not require optimality, and other statistics are possible (e.g., a difference in medians instead of means). In fact, we will encounter situations where no one statistic is optimal, and different researchers might choose different statistics for their formal statistical analyses, depending on the application at hand.

Inference is usually based on a single statistic that, ideally, measures some effect or difference of interest.

The standard error of the difference between two sample means is the standard deviation of its sampling distribution. Statistical theory shows that under the assumptions of the t-test, the standard error of the difference is

$$\text{SE(diff)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where n_1 and n_2 are the group sample sizes, and σ_1^2 and σ_2^2 are the true variances within each group. In the t-test, we assume that these variances are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$, which is how we get the second equality. Note that this simplifies to $\sqrt{2}\sigma/\sqrt{n}$ when the sample sizes are equal.

In practice, the SE is estimated as a weighted average of the observed sample variances in each group:

$$\text{estimated SE(diff)} = \sqrt{\frac{s_1^2(df_1) + s_2^2(df_2)}{df_1 + df_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$. This estimated standard error has $n_1 + n_2 - 2 = df_1 + df_2$ degrees of freedom. Importantly, we are using the *pooled variance* estimator $\frac{s_1^2(df_1)+s_2^2(df_2)}{df_1+df_2}$ to estimate the (assumed to be common) variance σ^2—instead of using the sample variance across all subjects regardless of treatment group—because it allows the group means to differ (even if their variances are assumed to be the same).

5.2.5 Computing the null sampling distribution

The next step in the general scheme of formal (classical) statistical inference is to compute the **null sampling distribution** of the chosen statistic. The null sampling distribution of a statistic is the probability distribution of the statistic across repeated experiments under the conditions defined by the model assumptions and the null hypothesis. For our HCI example, we consider what would happen if the truth is that there is no difference in reaction times between the two background colors, and we repeatedly sample 35 subjects and randomly assign yellow to 17 of

them and cyan to 18 of them, and then calculate the t-statistic each time. The distribution of the t-statistics under these conditions is the null sampling distribution of the t-statistic appropriate for this experiment. Note that we are assuming that the treatment group sizes 17 and 18 are fixed as part of the design of the experiment. If the experiments of interest have these as random numbers, then the null sampling distribution should technically have these quantities vary as well.

For the HCI example, the null sampling distribution of the t-statistic can be shown to match a well known, named continuous probability distribution called the "t-distribution" (see Section 3.9). Actually there are an infinite number of t-distributions (a family of distributions) and these are named (indexed) by their "degrees of freedom" (df). For the two-group quantitative outcome experiment, the df of the t-statistic and its corresponding null sampling distribution is $(n_1 - 1) + (n_2 - 1)$, so we will use the t-distribution with $n_1 + n_2 - 2$ df to make our inferences. For the HCI experiment, this is 17+18-2=33 df.

The calculation of the mathematical form (pdf) of the null sampling distribution of any chosen statistic using the assumptions of a given model is beyond the scope of this book, but the general idea can be seen in Section 3.7.

Statistical inference utilizes the null sampling distribution of a chosen statistic based on modeling assumptions. Probability theory allows us to compute what the null sampling distribution is for a particular statistic of interest.

You may notice that the null hypothesis of equal population means is in some sense "complex" rather than "point" because the two means could be both equal to 600, 601, etc. It turns out that the t-statistic has the same null sampling distribution regardless of the exact value of the population mean (and of the population variance), although it does depend on the sample sizes, n_1 and n_2.

5.2.6 Finding the p-value

Once we have the null sampling distribution of a statistic, we can see whether or not the observed statistic is "typical" of the kinds of values that we would expect to see when the null hypothesis is true (which is the basic interpretation of the null sampling distribution of the statistic). If we find that the observed statistic is typical, then we conclude that our experiment has not provided evidence against the null hypothesis, and if we find it to be atypical, we conclude that we do have evidence against the null hypothesis.

The formal language we use is to either "reject" the null hypothesis (in favor of the alternative) or to "retain" the null hypothesis. The word "accept" is not a good substitute for retain (see below). The inferential conclusion to "reject" or "retain" the null hypothesis is simply a conjecture based on the evidence in the observed data. But whichever inference we make, there *is* an underlying truth (null or alternative) that we can never know for sure, and there is always a chance that we will be wrong in our conclusion even if we use all of our statistical tools correctly.

Classical statistical inference focuses on controlling the chance that we incorrectly reject the null hypothesis. By "incorrectly reject the null hypothesis, " we mean that we reject the null hypothesis when the underlying truth is that it is correct. We call the erroneous conclusion that the null hypothesis is incorrect when it is actually correct a **Type 1 error**. (But because the true state of the null hypothesis is unknowable, we never can be sure whether or not we have made a Type 1 error in any specific actual situation.) To use a court justice metaphor, a Type I error is synonymous to falsely convicting an innocent person of a crime. Understandably, we would like to limit how often this type of error occurs.

The usual way that we make a formal, objective reject vs. retain decision of the null hypothesis is to calculate a p-value. Formally, a **p-value** is the probability that any given experiment will produce a value of the chosen statistic equal to or more extreme than the value of the statistic observed in the actual experiment, under the assumption that the null hypothesis is true and the model assumptions are correct. Be careful: The latter half of this definition is as important as the first half.

A p-value is the probability that any given experiment will produce a value of the chosen statistic equal to or more extreme than the value of the statistic observed in the actual experiment, when the null hypothesis is true and the model assumptions are correct.

For the HCI example, the numerator of the t-statistic is the difference between the observed sample means. Therefore, values near zero support the null hypothesis of equal population means, while values far from zero in either direction support the alternative hypothesis of unequal population means. In our specific experiment, the t-statistic equals 0.30. A value of -0.30 would give exactly the same degree of evidence for or against the null hypothesis (and the direction of subtraction is arbitrary). Values smaller in absolute value than 0.30 are more suggestive that the underlying truth is equal population means, while larger values support the alternative hypothesis. So the p-value for this experiment is the probability of getting a t-statistic greater than 0.30 or less than -0.30 based on the null sampling distribution of the t-distribution with 33 df. As explained in chapter 3, this probability is equal to the corresponding area under the curve of the pdf of the null sampling distribution of the statistic. As shown in Figure 5.2 the chance that a random t-statistic is less than -0.30 if the null hypothesis is true is 0.382, as is the chance that it is above +0.30. So the p-value equals 0.382+0.382=0.764, i.e. 76.4% of null experiments would give a t-value this large or larger (in absolute value). We conclude that the observed outcome (t=0.30) is not uncommonly far from zero when the null hypothesis is true, so we have no reason to believe that the null hypothesis is false (assuming that our model assumptions are correct—namely, that the difference in means indeed follows a t-distribution).

The usual convention (and it is only a convention, not anything stronger) is to reject the null hypothesis if the p-value is less than or equal to 0.05 and retain it otherwise. Under some circumstances it is more appropriate to use numbers bigger or smaller than 0.05 for this **decision rule**. We call the cutoff value the **significance level** of a test, and use the symbol alpha (α), with the conventional alpha being 0.05. We use the phrase statistically significant at the 0.05 (or some other) level, when the p-value is less than or equal to 0.05 (or some other value). This indicates that if we have used a correct model, i.e., the model assumptions mirror reality and if the null hypothesis happens to be correct, then a result like ours or one even more "un-null-like" would happen at most 5% of the time. It is reasonable to say that because our result is atypical for the null hypothesis,

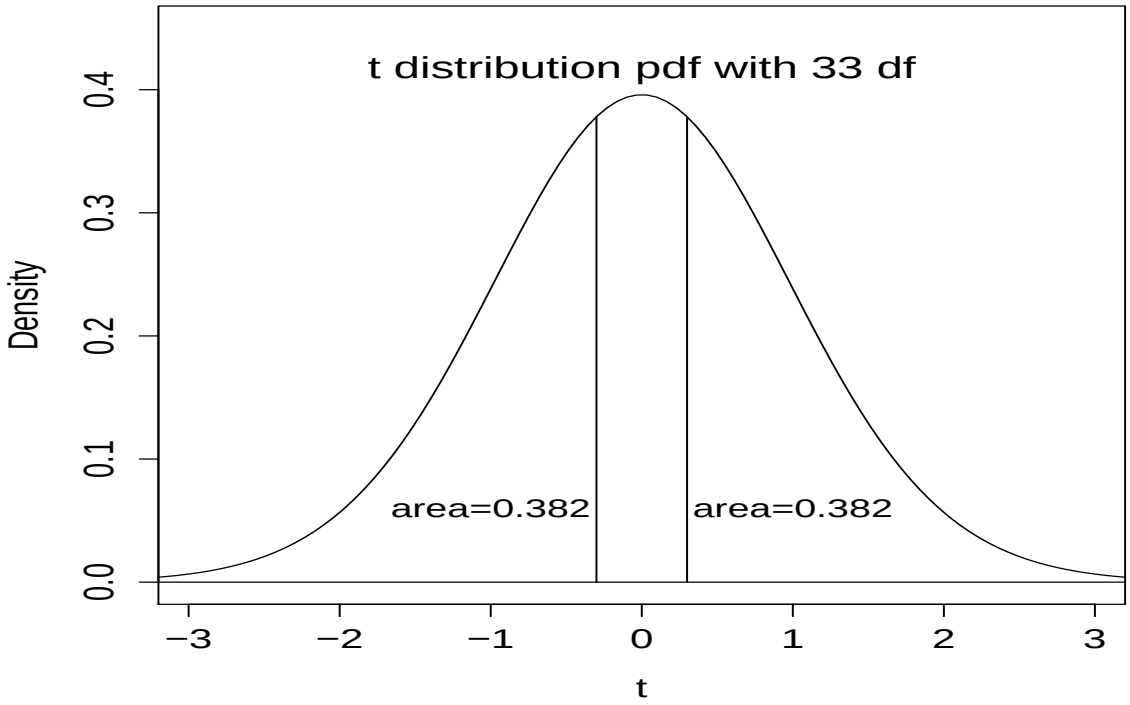


Figure 5.2: Calculation of the p-value for the HCI example

then claiming that the alternative hypothesis is true is appropriate. *But when we get a p-value of less than or equal to 0.05 and we reject the null hypothesis, it is completely incorrect to claim that there is only a 5% chance that we have made an error.* For more details, see Chapter 11.

Note that a "statistically non-significant" p-value is not necessarily an insignificant finding. For example, it may be very scientifically interesting that there is no statistically significant difference between two groups. So, even if you fail to reject the null hypothesis, you shouldn't necessarily conclude that there isn't anything interesting to conclude from your experiment!

The most common decision rule is to reject the null hypothesis if the p-value is less than or equal to 0.05 and to retain it otherwise.

It is important to realize that the p-value is a random quantity. If we could repeat our experiment (with no change in the underlying state of nature), then we would get a different p-value. What does it mean for the p-value to be "correct"? For one thing it means that we have made the calculation correctly, but since the computer is doing the calculation we have no reason to doubt that. What is more important is to ask whether the p-value that we have calculated is giving us appropriate information. For one thing, when the null hypothesis and modeling assumptions are really true (which we can never know for certain) an appropriate p-value will be less than 0.05 exactly 5% of the time over repeated experiments. So if the null hypothesis is true, and if you and 99 of your friends independently conduct experiments (which is what friends tend to do, right?), about five of you will get p-values less than or equal to 0.05, causing you to incorrectly reject the null hypothesis. Which five people this happens to has nothing to do with the quality of their research; it just happens because of bad luck!

And if an alternative hypothesis is true, then all we know is that the p-value will be less than or equal to 0.05 at least 5% of the time, but it might be as little 6% of the time. So a "correct" p-value does not protect you from making a lot of **Type 2 errors** which happen when you incorrectly retain the null hypothesis. With Type 2 errors, something interesting is going on in nature, but you miss it. See Section 5.2.10 for more on this "power" problem.

It is also important to keep in mind that all p-values are predicated on certain modeling assumptions being (at least approximately) correct. We would say that

a p-value is "incorrect" if the modeling assumptions it depends on are incorrect for the data at hand. Remember that a p-value is computed from a null sampling distribution, and we only know what that null sampling distribution is by assumption. If the modeling assumptions we used to compute this p-value are incorrect, then our statistic follows a *different null sampling distribution* that we do not know about. In this scenario, there is a mismatch between the actual null sampling distribution we *should* be using and the null sampling distribution we are *actually* using. As a result, we may well falsely reject the null hypothesis more than 5% of the time (i.e., we may have a higher Type 1 error rate), leading to inappropriate statistical inferences. Examples of different "assumption violations" for the t-test are non-normality of the distributions (though this is not a problem if the sample size is large, because of the CLT), unequal variance of the outcome measure for the two treatment groups, confounding of treatment group with important unmeasured explanatory variables, or lack of independence of the measures (for example, if some subjects are accidentally measured in both groups). If any of these "assumption violations" are sufficiently large, the p-value loses its validity—i.e., it becomes less useful (or even misleading) for conducting inferences about the science under study.

A p-value has meaning only if the correct null sampling distribution of the statistic has been used, i.e., if the assumptions of the test are (reasonably well) met. Computer programs generally give no warnings when they calculate p-values that are inappropriate for the data at hand.

5.2.7 Confidence intervals

Besides p-values, another way to express the evidence of an experiment is to compute one or more **confidence intervals**, often abbreviated CI. We would like to make a statement like, "We are sure that the difference between μ_1 and μ_2 is no more than 20 ms." However, that is not possible: Statistics is not about certainty, but rather about uncertainty. In short, confidence intervals tell us how certain we are about our uncertainty. We can only make statements such as, "We are 95% confident that the difference between μ_1 and μ_2 is no more than 20 ms." The choice of the percent confidence number is as arbitrary as the choice of alpha for

computing a p-value, and indeed there is a correspondence between the two. Typically, researchers report a 95% confidence interval (which corresponds to an alpha = 0.05 p-value), but we can choose another number, like 99% or 75%. However, when we do so, the width of the interval changes: High confidence requires wider intervals.

The actual computations are usually done by computer, but in many instances the idea of the calculation is simple.

If the underlying data are normally distributed, or if we are looking at a sum or mean with a large sample size (and can therefore invoke the CLT), then a confidence interval for a quantity (statistic) is computed as the statistic plus or minus the appropriate "multiplier" times the estimated standard error of the quantity. The multiplier used depends on both the desired confidence level (e.g., 95% vs. 90%) and the degrees of freedom for the standard error (which may or may not have a simple formula). The multiplier is based on the t-distribution which takes into account the uncertainty in the standard deviation used to estimate the standard error. We can use a computer or table of the t-distribution to find the multiplier as the value of the t-distribution for which plus or minus that number covers the desired percentage of the t-distribution with the correct degrees of freedom. If we call the quantity 1-(confidence percentage)/100 as alpha (α), then the multiplier is the 1-$\alpha/2$ quantile of the appropriate t-distribution.

If the data are not normally distributed and we cannot invoke the CLT, then the appropriate confidence interval may not be symmetric (and as a result, we cannot simply perform a "plus-or-minus" calculation), because the distribution of the statistic of interest may not be symmetric. This issue is largely outside the scope of this class, but we will discuss it briefly in Chapter 15.

For our HCI example the 95% confidence interval for the fixed, unknown, "secret-of-nature" that equals $\mu_1 - \mu_2$ is [-106.9, 144.4]. We are 95% confident that the mean reaction time is between 106.9 ms shorter and 144.4 ms longer for the yellow background compared to cyan. The meaning of this statement is that if all of the assumptions are met, and if we repeat the experiment many times, the *random* interval that we compute each time will contain the single, fixed, true

parameter value 95% of the time. Similar to the interpretation a p-value, if 100 competent researchers independently conduct the same experiment, by bad luck about five of them will unknowingly be incorrect if they claim that the 95% confidence interval that they correctly computed actually contains the true parameter value.

Note that, in the above example, we computed the confidence interval for $\mu_1 - \mu_2$ instead of comparing the confidence interval for μ_1 to the confidence interval for μ_2. It is sometimes common for researchers to see if the confidence interval of μ_1 and the confidence interval of μ_2 overlap, and then use this overlap—or lack thereof—to determine whether there is a significant difference between the two parameters, but this is incorrect. If the confidence interval of μ_1 and the confidence interval of μ_2 do not overlap, then it is true that we will reject the null hypothesis that $\mu_1 = \mu_2$; however, if the confidence intervals do overlap, it may still be the case that we will reject this null hypothesis. In general, when conducting inference on the difference between two parameters, it is important to look at the confidence interval of their difference and *not* the difference of their confidence intervals.

Confidence intervals are in many ways more informative than p-values. Their greatest strength is that they help a researcher focus on **substantive significance** in addition to statistical significance. Consider a bakery that does an experiment to see if an additional complicated step will reduce waste due to production of unsaleable, misshapen cupcakes. If the amount saved has a 95% CI of [0.1, 0.3] dozen per month with a p-value of 0.02, then even though this would be statistically significant (because zero is outside the confidence interval and yields a p-value < 0.05), it would not be substantively significant (because saving 0.1 to 0.3 dozen cupcakes per month is likely not a substantial difference in terms of business).

In contrast, if we had a 95% CI of [-30, 200] dozen per month with p=0.15, then even though this is not statistically significant, the inclusion of substantively important values like 175 dozen per month tells us that the experiment has not provided enough information to make a good, real world conclusion.

Finally, if we had a 95% CI of [-0.1, 0.2] dozen per month with p=0.15, we would conclude that even if a real non-zero difference exists, its magnitude is not enough to warrant adding the complex step to our cupcake making.

Confidence intervals can add a lot of important real world information to p-values and help us complement statistical significance with substantive significance.

5.2.8 Assumption checking

We have seen above that the p-value can be misleading or "wrong" if the model assumptions used to construct the statistic's sampling distribution are not close enough to the reality of the situation. To protect against being mislead, we usually assess the plausibility of model assumptions after conducting an analysis but before considering its conclusions.

Depending on the model, assumption checking can take several different forms. A major role is played by examining the model **residuals**. Remember that our standard model says that for each treatment group the best guess (the expected or predicted value) for each observation is defined by the means of the structural model. Then, the actual observed value for each observation will deviate above or below the true mean. The error component of our model describes the assumed distribution of these deviations, which are called **errors**. The residuals, which are defined as the observed minus expected value for each outcome measurement, are our best estimates of the unknowable, true errors for each subject. We will assess if the distribution of observed residuals is consistent with the error distribution posited by the model.

After fitting an initial model (through a hypothesis test, confidence interval, or some other inferential tool), assumption checking is needed to assess the plausibility of the assumptions underpinning the model.

For the two-treatment-group experiment discussed in this chapter, the only information we have about each subject is the treatment group they were assigned in the experiment. Thus, all subjects within each group are treated identically, and the predicted values within each group are the same. For the t-test, the observed group means *are* the two predicted values from which the residuals can be

computed. Then we can check if the residuals for each group follow a Normal distribution with equal variances for the two groups (or more commonly, we check the equality of the variances and check the normality of the combined set of residuals). At first, it may appear odd that we treat every subject within a group as identical, especially if we are dealing with human subjects, which tend to differ from each other. However, all we need for valid inference is that the two groups *on average* are the same, which is the case for randomized experiments. These nuances will be discussed in detail in Chapter 7.

Another important assumption is the independence of the errors. There should be nothing about the subjects that allows us to predict the sign or the magnitude of one subject's error just by knowing the value of another specific subject's error. As a trivial example, if we have identical twins in a study, it may well be true that their errors are not independent. This might also apply to close friends in some studies. The worst case is to apply both treatments to each subject, and then pretend that we used two independent samples of subjects. In this worst case, the two measurements are obviously correlated (and thus dependent), because they came from the same subject. Usually there is no way to check the independence assumption from the data; we just need to think about how we conducted the experiment to consider whether the assumption might have been violated. In some cases, because the residuals can be looked upon as a substitute for the true unknown errors, certain residual analyses may shed light on the independent errors assumption.

You can be sure that the underlying reality of nature is never perfectly captured by our models. This is why statisticians often say "all models are wrong, but some are useful" (coined by George E.P. Box). It takes some experience to judge how badly the assumptions can be bent before the inferences are broken. For now, a rough statement can be made about the independent samples t-test: we need to worry about the reasonableness of the inference if the normality assumption is strongly violated, if the equal variance assumption is moderately violated, or if the independent errors assumption is mildly violated. We say that a statistical test is **robust** to a particular model violation if the p-value remains approximately "correct" even when the assumption is moderately or severely violated.

All models are wrong, but some are useful. It takes experience and judgement to evaluate model adequacy.

5.2.9 Subject matter conclusions

Applying subject matter knowledge to statistical inference tools like p-values and confidence intervals is one key form of relating statistical conclusions back to the subject matter of the experiment. An analysis is incomplete if you stop at reporting the p-value and/or CI without returning to the original scientific question(s). In short, statistics is more than just number-crunching: Throughout EDA and statistical analyses, you should have scientific questions at the forefront of your mind. The main goal of any statistical analysis is to answer scientific questions, not compute p-values or some other quantity.

5.2.10 Power

Earlier in this section, we discussed how we can perform a hypothesis test to retain or reject a null hypothesis. If we *correctly* reject a null hypothesis, then we have detected that some alternative hypothesis is true. Intuitively, larger deviances from the null hypothesis should be easier to detect—for example, if there is (in truth) a huge difference in response between two treatment groups, then statistical tools should easily detect such a deviance from the null of no difference. However, tiny deviances will be much harder to detect. In short, our power to detect deviances from the null depends (in part) on the magnitude of that deviance.

The **power** of an experiment is defined for specific alternatives, e.g., $|\mu_1 - \mu_2| = 100$, rather than for the entire, complex alternative hypothesis. The power of an experiment for a given alternative hypothesis is the chance that we will get a statistically significant result (reject the null hypothesis) when that alternative is true for any one realization of the experiment. Power varies from α to 1.00 (or $100\alpha\%$ to 100%). Power is at a minimum α because even when the null hypothesis is true, we will still reject the null $100\alpha\%$ of the time—so, any deviation from the null hypothesis should make us reject the null more frequently. The concept of power is related to **Type 2 error**, which is the error we make when we retain the null hypothesis when a particular alternative is true. Usually the *rate* of making Type 2 errors is symbolized by beta (β). Then, power is 1-β or 100-$100\beta\%$. Typically people agree that 80% power (β=20%) for some substantively important **effect size** (specific magnitude of a difference as opposed to the zero difference of the null hypothesis) is a minimal value for good power.

You should use the methods of Chapter 11 to estimate the power of any exper-

iment before running it. Power analyses are very important during the design of an experiment, because it can help determine if an experiment is worth running. For example, let us say that you are considering conducting an experiment, and you find that the power of that experiment is 20-30%. That means that even if you are studying effective treatments, you only have a 20-30% chance of getting a statistically significant result. In many cases, this would mean that running the experiment is not a worthwhile endeavor.

Power is a key part of not only the design stage of an experiment but also the analysis stage. It is important to realize that some statistical analyses will have more power than others. For example, the t-test is meant to detect differences in *means* of two populations; it cannot detect differences in variances or some other characteristic of a population. In short, a t-test has a decent amount of power in detecting differences in means between two populations, but it does not have much power in detecting other differences (unless the means change because of those differences). This is why it is important to think about your scientific questions at hand, because your goal should be to choose the design and analysis that will maximize your power in answering those questions.

Choosing a design and analysis plan that maximizes the power to answer scientific questions of interest is a quality of a good scientist.

For now, the importance of power is how it applies to inference. If you get a small p-value, power becomes irrelevant, and you conclude that you should reject the null hypothesis, always realizing that there is a chance that you might be making a Type 1 error. If you get a large p-value, you "retain" the null hypothesis. If the power of the experiment is small, Maybe change this to, it's difficult to tell if you retained the null because it is true or because you don't have the power to detect an alternative (in which case you've made a Type 2 error). But if you have good power for some reasonably important effect size, then a large p-value is good evidence that no important effect exists, although a Type 2 error is still possible (but less likely).

In a nutshell: All classical statistical inference (e.g., hypothesis tests, p-values, and confidence intervals) is based on the same set of steps in which a sample statistic is compared to the distribution of values we would expect if the null hypothesis is true. Furthermore, any statistical analysis should connect these inferential tools back to the scientific questions at hand.

5.3 Do it in R

To perform a t-test in R, you use the `t.test()` function. Here is the one line of code you would use to perform the t-test for the HCI example:

```
> t.test(time~color, data = background, var.equal = TRUE)

  Two Sample t-test

data:  time by color
t = 0.30284, df = 33, p-value = 0.7639
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -106.9446  144.3498
sample estimates:
mean in group 0 mean in group 1
       679.6471        660.9444
```

Note that the `outcomeVariable groupVariable` is the same syntax we used for the `aggregate()` function and `boxplot()` function. Also note that we set `var.equal = TRUE`, which means the test assumes that the outcome in the two groups (i.e., the reaction time in the yellow group and the cyan group) has the same variance. By default, the `t.test()` function sets `var.equal = FALSE`, but we will talk about that specification later.

There are several key pieces of information we get from the `t.test()` output:

- The top line tells us that the observed t-statistic is 0.30, the degrees of freedom is 33, and the p-value is 0.76. If you remember that the degrees of freedom for the t-test is the number of subjects minus 2, you know from this output that there are 35 subjects. Because the p-value is (much) greater than 0.05, you also know that you fail to reject the null hypothesis.

- We see that the 95% confidence interval *for the mean difference in reaction times between the yellow group and cyan group* is (-106.94, 144.35). Even if you didn't have the p-value, you would still know to fail to reject the null hypothesis, because 0 is contained in this confidence interval.
- We see the same means in the yellow (`color = 0`) and cyan (`color = 1`) groups. This line is very important: Because the yellow group is listed first, we know that the t-test is computing the mean reaction time *for the yellow group* minus the mean reaction time *for the cyan group*. We see that the yellow group had a higher reaction time, which explains why the t-statistic is positive.

It's worthwhile to spend a bit of extra time on the last point, because (in my experience), it's the point where students commonly make mistakes. If the result from the t-test had been statistically significant (i.e., if the p-value had been less than 0.05), what would we have concluded? In that case, it would be correct to say something like, "We reject the null hypothesis, thereby concluding that the mean response time is not equal between the two groups," but this would not be the full picture. If you told that to someone who actually cared about the answer, they would likely say, "Okay...So which group performed better??" From the positive t-statistic, we know that the yellow group had the higher reaction time, which means *the cyan group* performed better (because a lower reaction time is better!)

In general, if you find a statistically significant effect from the t-test, it's very important to remember (1) the direction of the significant effect, and (2) the scientific meaning of that significant effect. Sometimes a positive numeric effect is a "negative" effect (qualitatively speaking), as is the case in this example.

Comparing the sample means within each group show the direction of the effect, and the standard deviations show us the inherent variability of our measurements.

In general, before trusting the results from a statistical test, it is essential to check that the assumptions behind the test are met for the data you're working with. In particular, one of the key assumptions of the t-test is that the outcome measurements in each group are Normally distributed (possibly with equal variances). Thus, a common model-checking step for the t-test is to plot the distribu-

tion of the outcome measurements in the two groups and see if they are Normally distributed.

Figure ?? shows separate histograms of the reaction time in the two groups with overlaid Normal pdfs. (For those curious, the code used to make these histograms is below.) With such a small sample size, we cannot expect perfectly shaped Normal distributions, even if the Normality assumption is perfectly true. The histograms look reasonably consistent with Normal distributions with fairly equal variance, although Normality is hard to judge with such a small sample. With the limited amount of information available, we cannot expect to make definite conclusions about the model assumptions of Normality or equal variance, but we can at least say that we do not see gross violations of these assumptions that would make us suspect that the p-value is misleading. As we will discuss in later chapters, in more complex models, we will usually use a "residual vs. fit" plot and a quantile-normal plot of the residuals for assumption-checking.

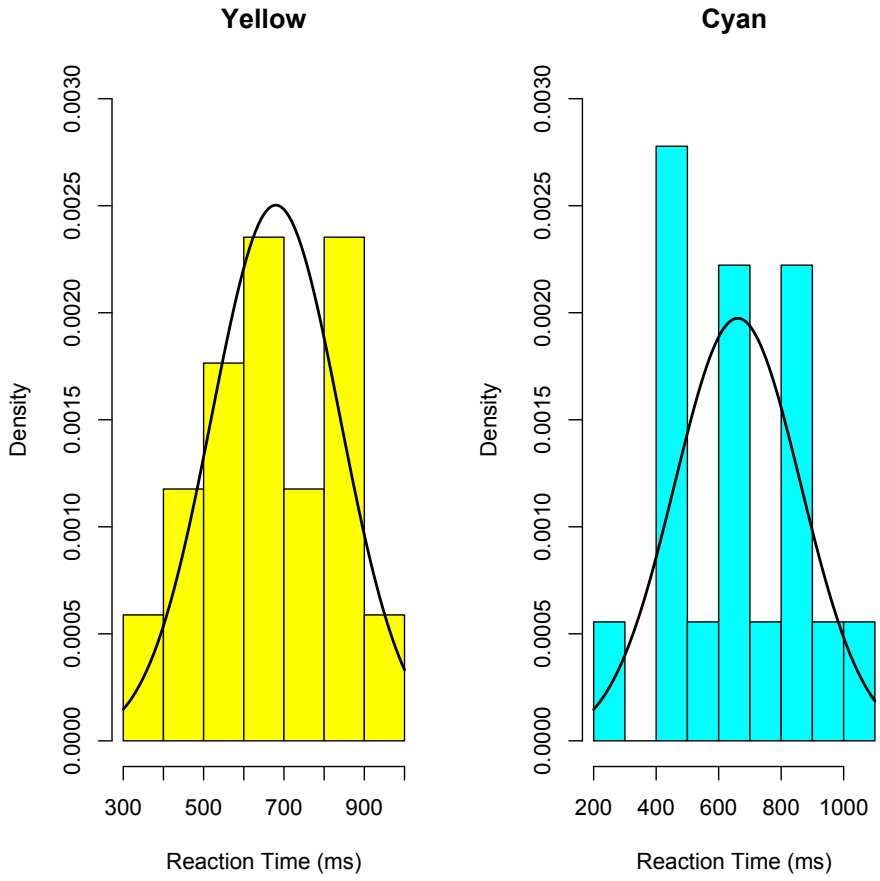


Figure 5.3: Histograms of the reaction times in the yellow and cyan groups, with overlayed Normal density curves.

```
#Are the outcomes in both groups Normally distributed?
#dataset for the yellow group
> background.yellow = subset(background, color == 0)
#dataset for the cyan group
> background.cyan = subset(background, color == 1)
#sample means
> yBar.y = mean(background.yellow$time)
> yBar.c = mean(background.cyan$time)
#sample variances
> s2.y = var(background.yellow$time)
> s2.c = var(background.cyan$time)
```

```
12
13 #make the two histograms with overlayed densities
14 > par(mfrow = c(1,2))
15 > hist(background.yellow$time, prob = TRUE,
16 +   ylim = c(0, 0.003),
17 +   xlab = "Reaction Time (ms)", main = "Yellow",
18 +   col = "yellow")
19 > curve(dnorm(x, mean=yBar.y, sd=sqrt(s2.y)),
20 +       col="black", lwd=2, add=TRUE, yaxt="n")
21 > hist(background.cyan$time, prob = TRUE,
22 +   ylim = c(0, 0.003),
23 +   xlab = "Reaction Time (ms)", main = "Cyan",
24 +   col = "cyan")
25 > curve(dnorm(x, mean=yBar.c, sd=sqrt(s2.c)),
26 +       col="black", lwd=2, add=TRUE, yaxt="n")
```

Now that we know how to perform the t-test in R, let's take a step back to understand how R arrived at the above results. In particular: Even though the yellow group had a slightly higher reaction time within the sample, why did we find a statistically insignificant effect? The key is to remember that the t-test compares the *sample means* within two groups; thus, statistical inference will rely on the *uncertainty* of our sample mean estimates. If there is a lot of uncertainty, confidence intervals will be wide, making it more likely that there is a statistically insignificant result.

The standard error of the mean (SEM) for a sample tells us about how well we have "pinned down" the population mean based on the inherent variability of the outcome and the sample size. It is worth knowing that the estimated SEM is equal to the standard deviation of the sample divided by the square root of the sample size. The less variable a measurement is and the bigger we make our sample, the better we can "pin down" the population mean (what we'd like to know) using the sample (what we can practically study). By "pin down," I mean become more certain in a probabilistic sense (e.g., being able to produce a narrower confidence interval). This is why **confidence intervals** are a key part of many statistical analyses.

When the statistic of interest is the sample mean, as we are focusing on now, we can use the central limit theorem to justify claiming that the (sampling) distribution of the sample mean is Normally distributed with standard deviation equal to $\frac{\sigma}{\sqrt{n}}$ where σ is the true population standard deviation of the measurement. The standard deviation of the sampling distribution of any statistic is called its **standard error**. If we happen to know the value of σ, then we are 95% confident that

the interval $\bar{x} \pm 1.96(\frac{\sigma}{\sqrt{n}})$ contains the true mean, μ. Remember that the meaning of a confidence interval is that if we could repeat the experiment with a new sample many times, and construct a confidence interval each time, they would all be different and 95% (or whatever percent we choose for constructing the interval) of those intervals will contain the single true value of μ.

> Technically, if the original distribution of the data is normally distributed, then the sampling distribution of the mean is normally distributed regardless of the sample size (and without using the CLT). Using the CLT, if certain weak technical conditions are met, as the sample size increases, the shape of the sampling distribution of the mean approaches the Normal distribution regardless of the shape of the data distribution. Typically, if the data distribution is not too bizarre, a sample size of at least 20 is enough to cause the sampling distribution of the mean to be quite close to the Normal distribution.

Unfortunately, the value of σ is not usually known, and we must substitute the sample estimate, s, instead of σ into the standard error formula, giving an estimated standard error. Commonly the word "estimated" is dropped from the phrase "estimated standard error", but you can tell from the context that σ is not usually known and s is taking its place. For example, the estimated standard deviation of the (sampling) distribution of the sample mean is called the standard error of the mean (usually abbreviated SEM), without explicitly using the word "estimated".

Instead of using the quantile of a Normal distribution (1.96 for 95% confidence intervals) times the standard deviation of the sampling distribution to calculate the "plus or minus" for a confidence interval, we must use a different multiplier when we substitute the estimated SEM for the true SEM. The multiplier we use is the quantile of a t-distribution (indeed, this is why it's called a t-test!) This quantile is calculated by the computer (or read off of a table of the t-distribution), but it depends on the number of degrees of freedom of the standard deviation estimate, which in the simplest case is $n - 1$ where n is the number of subjects in the specific experimental group of interest. When calculating 95% confidence intervals, the multiplier can be as large as 4.3 for a sample size of 3, but shrinks towards 1.96 as the sample size grows large. This makes sense: With a smaller

sample size, we are less certain about the true value of σ, and a wider confidence interval reflections that additional uncertainty.

In the context of the yellow-group versus cyan-group comparison, our parameter of interest is the difference in mean reaction time, written as $\mu_Y - \mu_C$, and this is estimated as the difference in the two sample means, written as $\bar{y}_Y - \bar{y}_C$ (i.e., the average response time in the yellow group minus the average response time in the cyan group). As we discussed in Section 5.2.4, the SE of $\bar{y}_Y - \bar{y}_C$ is $\sigma\sqrt{\frac{1}{n_Y} + \frac{1}{n_C}}$, where σ is the *true* standard deviation of the response time (assumed to be the same in both groups), and n_Y and n_C are the sample sizes in the yellow and cyan groups, respectively. As discussed above, σ is replaced with an estimate, and thus we use a t-distribution to compute the confidence interval. Specifically, the degrees of freedom is $n_Y + n_C - 2 = 33$, as was shown in the t-test R output above. Thus, our confidence interval will be $\bar{y}_Y - \bar{y}_C \pm t_{\alpha/2,33} \times \hat{\sigma}\sqrt{\frac{1}{n_Y} + \frac{1}{n_C}}$, where $t_{\alpha/2,33}$ is the $(\alpha/2)$-quantile of the t-distribution with 33 degrees of freedom. The R code below shows how you can manually compute this confidence interval. For the sake of 36-309, you are not expected to memorize this formula or know how to manually compute confidence intervals, but the below code may be useful for some students who want to know what's going on in the background when R conducts the t-test.

```
#dataset for the yellow group
> background.yellow = subset(background, color == 0)
#dataset for the cyan group
> background.cyan = subset(background, color == 1)
#sample means
> yBar.y = mean(background.yellow$time); yBar.c = mean(background.
    cyan$time)
#sample variances
> s2.y = var(background.yellow$time); s2.c = var(background.cyan$
    time)
#sample sizes
> n.y = nrow(background.yellow); n.c = nrow(background.cyan)
#degrees of freedom
> df.y = n.y - 1; df.c = n.c - 1
#the estimated SE for the mean difference is
> se = sqrt( ( (s2.y*df.y + s2.c*df.c)/(df.y + df.c) )*(1/n.y + 1/
    n.c) )
#the multiplier for the confidence interval is
> critVal = qt(p = 0.975, df = 33)
#then, the confidence interval is
> (yBar.y - yBar.c) - critVal*se; (yBar.y - yBar.c) + critVal*se
-106.9446
```

```
144.3498
```

In a nutshell: To analyze a two-group quantitative outcome experiment, first perform EDA to get a sense of the direction and size of the effect, to assess the normality and equal variance assumptions, and to look for mistakes. Then perform a t-test (or equivalently, a one-way ANOVA). If the assumption checks are OK, reject or retain the null hypothesis of equal population means based on a small or large p-value, respectively. However, providing a confidence interval for the difference in means will give a more informative conclusion about the scientific questions at hand than a p-value.

Chapter 6

One-way ANOVA

One-way ANOVA examines equality of population means for a quantitative outcome and a single categorical explanatory variable with any number of levels.

The t-test of Chapter 5 looks at quantitative outcomes with a categorical explanatory variable that has only two levels. The one-way **Analysis of Variance** (**ANOVA**) can be used for the case of a quantitative outcome with a categorical explanatory variable that has two or more levels of treatment. In fact, for $k = 2$ groups, the t-test and one-way ANOVA give identical results; in this sense, one-way ANOVA can be viewed as a generalization of the t-test for assessing if population means differ among $k \geq 2$ groups. The term one-way, also called one-factor, indicates that there is a single explanatory variable ("treatment") with two or more levels, and only one level of treatment is applied at any time for a given subject. In this chapter we assume that each subject is exposed to only one treatment, in which case the treatment variable is being applied "between-subjects". For the alternative in which each subject is exposed to several or all levels of treatment (at different times) we use the term "within-subjects", but that is covered in Chapter 13. We use the term two-way or two-factor ANOVA when the levels of two different explanatory variables are being assigned and each subject is assigned to one level of *each* factor—this is covered in Chapter 8.

It is worth noting that the situation for which we can choose between one-way ANOVA and an independent samples t-test is when the explanatory variable has exactly two levels. In that case we always come to the same conclusions regardless of which method we use.

The term "analysis of variance" is a bit of a misnomer. In ANOVA we use variance-like quantities to study the equality or non-equality of population means. So we are analyzing means, not variances. There are some unrelated methods, such as "variance component analysis" which have variances as the primary focus for inference.

6.1 Moral Sentiment Example

As an example of applying one-way ANOVA in real research, we'll consider the research reported in "Moral sentiments and cooperation: Differential influences of shame and guilt" by de Hooge, Zeelenberg, and M. Breugelmans (Cognition & Emotion,21(5): 1025-1042, 2007).

As background you need to know that there is a well-established theory of Social Value Orientations or SVO (see Wikipedia for a brief introduction and references). SVOs represent characteristics of people with regard to their basic motivations. In this study a questionnaire called the Triple Dominance Measure was used to categorize subjects into "proself" and "prosocial" orientations. In this chapter we will examine simulated data based on the results for the proself individuals.

The goal of the study was to investigate the effects of emotion on cooperation. The study was carried out using undergraduate economics and psychology students in the Netherlands.

The sole explanatory variable is "induced emotion". This is a nominal categorical variable with three levels: control, guilt and shame. Each subject was randomly assigned to one of the three levels of treatment. Guilt and shame were induced in the subjects by asking them to write about a personal experience where they experienced guilt or shame respectively. The control condition consisted of having the subject write about what they did on a recent weekday. (The validity of the emotion induction was tested by asking the subjects to rate how strongly they were feeling a variety of emotions towards the end of the experiment.)

After inducing one of the three emotions, the experimenters had the subjects participate in a one-round computer game that is designed to test cooperation. Each subject initially had ten coins, with each coin worth 0.50 Euros for the subject but 1 Euro for their "partner" who is presumably connected separately to the computer. The subjects were told that the partners also had ten coins, each worth 0.50 Euros for themselves but 1 Euro for the subject. The subjects

decided how many coins to give to the interaction partner, without knowing how many coins the interaction partner would give. In this game, both participants would earn 10 Euros when both offered all coins to the interaction partner (the cooperative option). If a cooperator gave all 10 coins but their partner gave none, the cooperator could end up with nothing, and the partner would end up with the maximum of 15 Euros. Participants could avoid the possibility of earning nothing by keeping all their coins to themselves which is worth 5 Euros plus 1 Euro for each coin their partner gives them (the selfish option). The number of coins offered was the measure of cooperation.

The number of coins offered (0 to 10) is the outcome variable, and is called "cooperation". Obviously this outcome is related to the concept of "cooperation" and is in some senses a good measure of cooperation, but just as obviously, it is not a complete measure of the concept.

Cooperation as defined here is a discrete quantitative variable with a limited range of possible values. As explained below, the Analysis of Variance statistical procedure, like the t-test, is based on the assumption of a Gaussian distribution of the outcome at each level of the (categorical) explanatory variable. In this case, it is judged to be a reasonable approximation to treat "cooperation" as a continuous variable. There is no hard-and-fast rule, but 11 different values might be considered borderline, while, e.g., 5 different values would be hard to justify as possibly consistent with a Gaussian distribution.

Note that this is a randomized experiment. The three levels of "treatment" (emotion induced) are randomized and assigned by the experimenter. If we do see evidence that "cooperation" differs among the groups, we can validly claim that induced emotion *causes* different degrees of cooperation. If we had only measured the subjects' current emotion rather than manipulating it, we could only conclude that emotion is *associated* with cooperation. Such an association could have other explanations than a causal relationship, e.g., poor sleep the night before could cause more feelings of guilt and more cooperation, without the guilt having any direct effect on cooperation. (See Section 7.1 for more on causality.)

The data can be found in MoralSent.dat. The data look like this:

emotion	cooperation
Control	3
Control	0
Control	0
⋮	⋮
Shame	1

Typical exploratory data analyses include a tabulation of the frequencies of the levels of a categorical explanatory variable like "emotion", e.g., by using the `table()` function:

```
> table(moral$emotion)

Control   Guilt   Shame
     39      42      45
```

We see that we have 39 controls, 42 guilt subjects, and 45 shame subjects. It is also typical to tabulate sample statistics of the outcome variable within each level of the explanatory level, e.g., by using the `aggregate()` function:

```
#sample means
> aggregate(cooperation~emotion, data = moral, FUN = mean)
  emotion cooperation
1 Control    3.487179
2   Guilt    5.380952
3   Shame    3.777778
#sample variances
> aggregate(cooperation~emotion, data = moral, FUN = var)
  emotion cooperation
1 Control    9.677463
2   Guilt   10.534262
3   Shame    8.676768
```

Meanwhile, we could also make side-by-side boxplots (code below; graphic in Figure 6.1):

```
#side-by-side boxplots
boxplot(cooperation~emotion, data = moral,
  xlab = "Induced Emotion", ylab = "Cooperation Score")
```

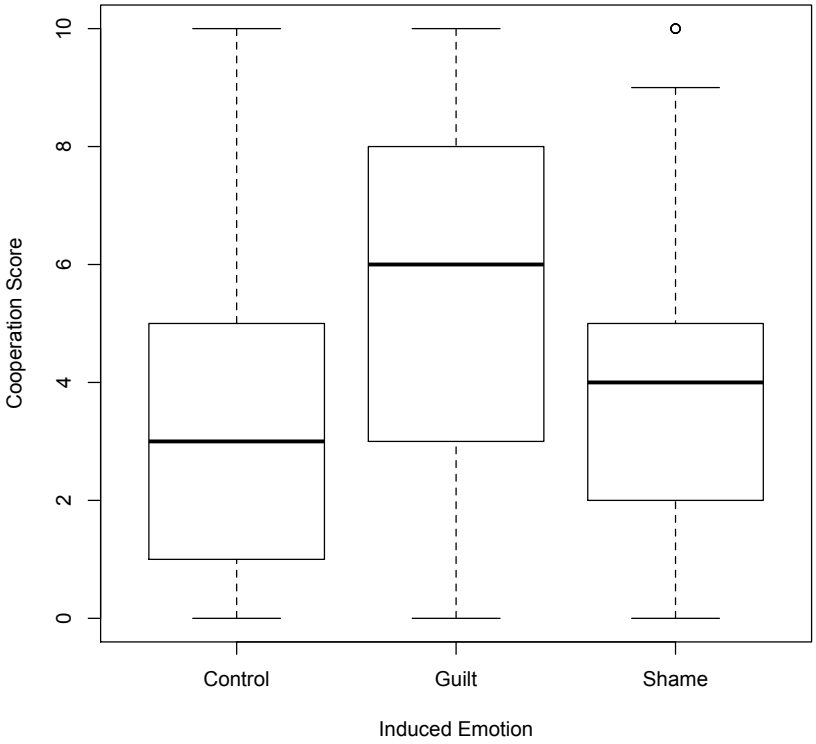


Figure 6.1: Boxplots of cooperation by induced emotion.

Our initial impression is that cooperation is higher for guilt than either shame or the control condition. The mean cooperation for shame is slightly lower than for the control. In terms of assessing model assumptions (which are the same as those of the t-test), the boxplots show fairly symmetric distributions with fairly equal spread (as demonstrated by the comparative IQRs). We see four high outliers for the shame group, but careful thought suggests that this may be unimportant because they are just one unit of measurement (coin) into the outlier region and that region may be "pulled in" a bit by the slightly narrower IQR of the shame group.

6.2 How one-way ANOVA works

The goal of the emotion data example is to assess if there is a significant difference in cooperation between the three treatment groups. Before we conduct an analysis, it will be helpful to go over the formal model for one-way ANOVA; that will help us better understand how to conduct inference using this model.

6.2.1 The model and statistical hypotheses

One-way ANOVA is appropriate when the following model holds. We have a single "treatment" with, say, k levels. "Treatment" may be interpreted in the loosest possible sense as any categorical explanatory variable. There is a population of interest for which there is a true quantitative outcome for each of the k levels of treatment. The population outcomes for each group have mean parameters that we can label μ_1 through μ_k with no restrictions on the pattern of means. The population variances for the outcome for each of the k groups defined by the levels of the explanatory variable are assumed to have the same value, usually called σ^2, with no restriction other than that $\sigma^2 > 0$. For treatment i, the distribution of the outcome is assumed to follow a Normal distribution with mean μ_i and variance σ^2, often written $N(\mu_i, \sigma^2)$. Furthermore, this model assumes that the true deviations of observations from their corresponding group mean parameters, called the "errors", are independent. In this context, independence indicates that knowing one true deviation would not help us predict any other true deviation.

Subjects are randomly selected from the population, and then randomly assigned to exactly one treatment each. The number of subjects assigned to treatment i (where $1 \leq i \leq k$) is called n_i if it differs between treatments or just n if all of the treatments have the same number of subjects. For convenience, define $N = \sum_{i=1}^{k} n_i$, which is the total sample size.

So, we are taking a sample of size N, dividing it into smaller samples of size n_1, n_2, ..., n_k, and then estimating the population mean within each of these k samples. Why would we be confident in relying on the estimation conducted on k-many (possibly small) samples? Because treatment is randomly assigned, the sample mean for any treatment group will (on average) be representative of how the N subjects *in the sample* would respond to that treatment. Furthermore, because the subjects were randomly sampled, the sample mean for any treatment group will (on average) also be representative of how *all subjects in the population*

would respond to that treatment. In most of this book, we focus on experiments where two stages of randomization occur: First, subjects are randomly sampled from a population, and then, sampled subjects are randomly assigned to treatment. Treatment randomization gives us "internal validity" (i.e., that we obtain something representative of subjects *in the sample*) and sampling randomization gives us "external validity" (i.e., that we obtain something representative of *all subjects in the population*). If we take away the randomization from one of these stages, many complications can occur—we'll talk about this in Chapter 7.

Technically, the sample group means are unbiased estimators of the population group means when treatment is randomly assigned and subjects are randomly sampled. The meaning of unbiased here is that the true mean of the sampling distribution of any group sample mean equals the corresponding population mean. Further, under the Normality, independence and equal variance assumptions it is true that the sampling distribution of $\bar{Y}_i$ is $N(\mu_i, \sigma^2/n_i)$, exactly.

The statistical model for which one-way ANOVA is appropriate is that the (quantitative) outcomes for each group are normally distributed with a common variance (σ^2). The errors (deviations of individual outcomes from the population group means) are assumed to be independent. The model places no restrictions on the population group means.

The null hypothesis is a point hypothesis stating that there is no difference in group effects. For one-way ANOVA, the null hypothesis is $H_0 : \mu_1 = \cdots = \mu_k$, which states that all of the population means are equal, without restricting what the common value is. The alternative must include everything else, which can be expressed as "at least one of the k population means differs from all of the others". It is *definitely wrong* to use $H_A : \mu_1 \neq \cdots \neq \mu_k$ because some cases, such as $\mu_1 = 5, \ \mu_2 = 5, \ \mu_3 = 10$, are neither covered by H_0 nor this incorrect H_A. You can write the alternative hypothesis as H_A : Not $\mu_1 = \cdots = \mu_k$ or "the population means are not all equal".

One way to correctly write H_A mathematically is $H_A : \exists\, i, j : \mu_i \neq \mu_j$. This is read as, "There exists an i and j such that μ_i does not equal μ_j."

This null hypothesis is called the "overall" null hypothesis, in the sense that we are testing for equality across all k means, i.e., equality overall. If we have only two levels of our categorical explanatory variable, we are back in the t-test case of Chapter 5, where retaining or rejecting the overall null hypothesis is all that needs to be done in terms of hypothesis testing. However, if we have 3 or more levels ($k \geq 3$) and we reject the overall null hypothesis, then we can conclude that *some* of the group means are unequal, but we do not yet know which ones are unequal. Therefore, follow-up hypothesis tests are usually conducted to get more nuanced assessments of the different population means. This brings up the common issue of multiple hypothesis testing and contrast testing, both of which we will discuss in detail in Chapter 12.

The overall null hypothesis for one-way ANOVA with k groups is $H_0 : \mu_1 = \cdots = \mu_k$. The alternative hypothesis is that "the population means are not all equal".

6.2.2 The F statistic (ratio)

The next step in standard inference is to select a statistic for which we can compute the null sampling distribution and that tends to fall in a different region for the alternative than the null hypothesis. In other words, we want a statistic that is *powerful* in detecting violations in the null hypothesis—i.e., we want a statistic that will change a lot if any of the k group means $\mu_1, \ldots, \mu_k$ differ. We will see that, for the purposes of ANOVA, the "F-statistic" is exactly the kind of statistic we want. The formula for the F-statistic is somewhat complicated, but it is worth mathematically defining what the F-statistic is so we can understand, intuitively, what is going on in one-way ANOVA.

In one-way ANOVA, we want to assess if the k different group means significantly differ from each other. We don't know what the actual population means

$\mu_1, \ldots, \mu_k$ are, so we will have to estimate them using the sample group means $\bar{Y}_1, \ldots, \bar{Y}_k$. (To obtain each of these sample group means, you can imagine dividing up your data into the k different groups, and then looking at the sample mean of some outcome Y in each group). Furthermore, remember that we defined $n_1, \ldots, n_k$ as the sample sizes in each group and N as the total sample size (i.e., $N = n_1 + n_2 + \cdots + n_k$). Finally, let $\bar{Y}$ define the *overall* sample mean (i.e., the mean of Y across all subjects, ignoring the k groups). Intuitively, if indeed the group means are all the same ($\mu_1 = \cdots = \mu_k$), then we expect each $\bar{Y}_1, \ldots, \bar{Y}_k$ to be about the same, and thus they won't vary a lot around the overall mean $\bar{Y}$. However, if some of the group means are different, then we expect the $\bar{Y}_1, \ldots, \bar{Y}_k$ to vary a fair amount around $\bar{Y}$. In other words, the *variance of the group means around the overall mean* is a good proxy for $H_0 : \mu_1 = \cdots \mu_k$; if this variance is low, then maybe H_0 is true, and if this variance is high, then likely this H_0 is false.

However, there is more variation in the data than just the variation of the group means $\bar{Y}_1, \ldots, \bar{Y}_k$. There is also the *variation within each group*. For example, you can imagine looking at all subjects in one of the k groups, and seeing how much the Y measurements vary in that group. Similar to the t-test from Chapter 5, in one-way ANOVA we make the assumption that all of the group variances are equal. So, aggregating the variances we see within each group is a good estimate of the overall variation we expect from each measurement. If our measurements are quite noisy (i.e., have high variance), then we expect the $\bar{Y}_1, \ldots, \bar{Y}_k$ to vary a lot as well, and this should be accounted for when determining if "the $\bar{Y}_1, \ldots, \bar{Y}_k$ vary a fair amount around $\bar{Y}$" (as discussed in the previous paragraph).

In short, the F-statistic assesses if the variation across groups (i.e., the variation of the group means $\bar{Y}_1, \ldots, \bar{Y}_k$) is large *relative to* the overall variance we see in our data. The formula for the F-statistic is:

$$F = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}{N-k}} = \frac{\text{"between group variation"}}{\text{"within group variation"}}$$

Here, Y_{ij} denotes the jth measurement in group i, and remember that n_i is the sample size for group i, N is the overall sample size, there are k groups, $\bar{Y}_i$ denotes the sample mean in group i, and $\bar{Y}$ denotes the overall group mean.

The formula for the F-statistic may be intimidating at first, and that is fine, but it isn't that bad once you start thinking in terms of "variance-like quantities" (remember that, in general, the formula for sample variance of Y is $s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$). To better understand the formula for the F-statistic, first, focus on the numerator

quantity $\frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{k-1}$. Bells should be ringing (even faintly) when you look at this quantity: It looks a lot like a variance, right? In fact, if the n_i weren't there, this would look a lot like the sample variance of $\bar{Y}_1, \ldots, \bar{Y}_k$ (however, the n_i are there to account for the fact that the groups are possibly different sample sizes—bigger groups will contribute more to the overall average $\bar{Y}$, and this needs to be accounted for when computing the variance of the group means). For this reason, this quantity is often called the "between group variation" (because it's measuring the variance between the k group means). Now focus on the denominator quantity $\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}{N-k}$. I hope bells are ringing a bit louder now, because you've been primed to look for variance-like quantities, which is exactly what this denominator term is. For a particular group i, the term $\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = (n_i - 1)s_i^2$; i.e., it is the group size n_i (minus 1) multiplied by the sample variance within group i (denoted s_i^2). So, in the denominator of the F-statistic, we are taking a weighted average of the sample variances within each group (weighted by group sample size). This is why this quantity is often called the "within group variation" (because it's taking a weighted average of the variation we see within each group).

I'm guessing that the past two pages have been a lot to take in mathematically, so maybe this is a good moment to take a breather and practice some self-care (which I usually like to do after encountering a lot of math). Maybe grab a coffee, go for a walk/run around Schenley, text some friends or family, or watch some stand-up (Mitch Hedberg is great for short math cleanses). After you get back from that, try to appreciate the following point:

Decomposing variation in data as "between group variation" and "within group variation" is a key idea in statistical analyses, and this is exactly what one-way ANOVA and the F-statistic are doing.

To better appreciate this point, we're going to take a step back to better understand what exactly these two types of variation are and their purpose in one-way ANOVA.

Remember that a sample variance is calculated as SS/df where SS is "sum of squared deviations from the mean" and df is "degrees of freedom" (see Page 70). As we can see above, the F-statistic is calculated as the ratio of two variance-like quantities, each of which are calculated as SS/df. We will call all of these quantities **mean squares** or MS, i.e., $MS = SS/df$. (Technically, these are

not really means, because the denominator is the df, not n.) So, the F-statistic can also be written as $F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$, where the numerator is the "mean square between-groups" and the denominator is the "mean square within-groups". To get a visual of $\text{MS}_{\text{within}}$, see Figure 6.2, which shows the within-group deviations used in the calculation of $\text{MS}_{\text{within}}$ for a simple two-group experiment with 4 subjects in each group. I hope you can see that the deviations shown (black horizontal lines extending from the colored points to the colored group mean lines) are due to the underlying variation of subjects within a group. We have assumed that the measurements (regardless of the group) have standard deviation σ, so the sample variance within a particular group is a good estimate of σ^2. So, $\text{MS}_{\text{within}}$ is just combining all of the k separate group estimates of σ^2. It is important to know that $\text{MS}_{\text{within}}$ has $N - k$ df.

$\text{MS}_{\text{within}}$ is a good estimate of σ^2 (from our model) regardless of the truth of H_0, because we have assumed a common variance across all groups regardless of H_0. $\text{SS}_{\text{within}}$ (and therefore $\text{MS}_{\text{within}}$) has N-k degrees of freedom with $n_i - 1$ coming from each of the k groups.

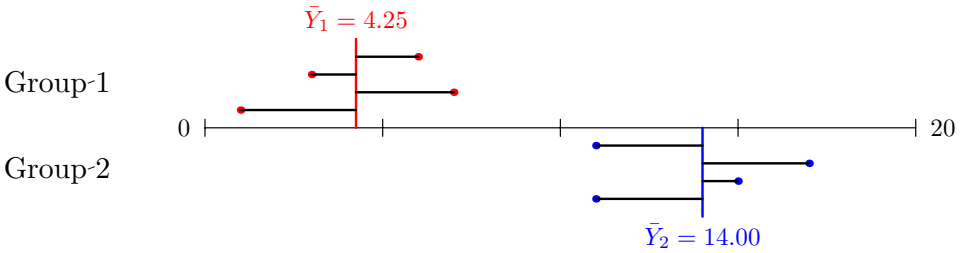


Figure 6.2: Deviations for within-group sum of squares

Now consider Figure 6.3, which represents the between-group deviations used in the calculation of $\text{MS}_{\text{between}}$ for the same little 2-group 8-subject experiment. The single vertical black line is the average of all of the outcomes values in all of the treatment groups, i.e., $\bar{Y}$. The colored vertical lines are still the group means. The horizontal black lines are the deviations used for the between-group calculations. For each subject, we get a deviation equal to the difference from that

subject's group mean to the overall (grand) mean . These deviations are squared and summed to get $\text{SS}_{\text{between}}$, which is then divided by the between-group df, which is $k-1$, to get $\text{MS}_{\text{between}}$.

$\text{MS}_{\text{between}}$ is a good estimate of σ^2 only when the null hypothesis is true. In this case, we expect the group means to be fairly close together and close to the grand mean. When the alternate hypothesis is true, as in our current example, the group means are farther apart and the value of $\text{MS}_{\text{between}}$ tends to be larger than σ^2. (We sometimes write this as "$\text{MS}_{\text{between}}$ is an inflated estimate of σ^2".)

Because of the way $\text{SS}_{\text{between}}$ is defined, $\text{MS}_{\text{between}}$ is a good estimate of σ^2 only if H_0 is true. Otherwise it tends to be larger. $\text{SS}_{\text{between}}$ (and therefore $\text{MS}_{\text{between}}$) has $k-1$ degrees of freedom.

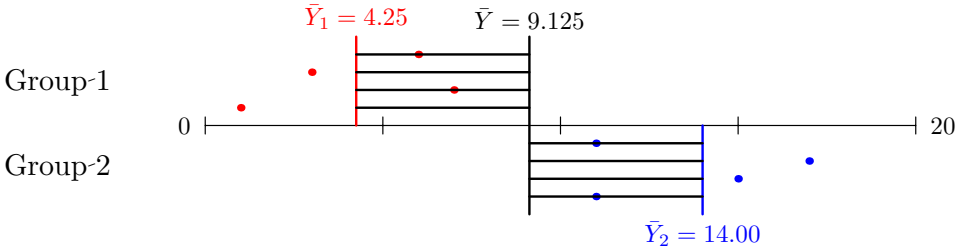


Figure 6.3: Deviations for between-group sum of squares

It might seem that we only need $\text{MS}_{\text{between}}$ to distinguish the null from the alternative hypothesis, but that ignores the fact that we don't usually know the value of σ^2. So instead we look at the ratio

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

to evaluate the null hypothesis. Because the denominator is always (under null and alternative hypotheses) an estimate of σ^2 (i.e., tends to have a value near σ^2), and the numerator is either another estimate of σ^2 (under the null hypothesis) or is inflated (under the alternative hypothesis), it is clear that the (random) values of the F-statistic (from experiment to experiment) tend to fall around 1.0 when

the null hypothesis is true and are *bigger* when the alternative is true. So if we can compute the sampling distribution of the F statistic under the null hypothesis, then we will have a useful statistic for distinguishing the null from the alternative hypotheses, where large values of F argue for rejection of H_0.

The F-statistic, defined by $F = \frac{\mathbf{MS_{between}}}{\mathbf{MS_{within}}}$, tends to be larger if the alternative hypothesis is true than if the null hypothesis is true.

6.2.3 Null sampling distribution of the F statistic

Using the technical condition that the quantities $\text{MS}_{\text{between}}$ and $\text{MS}_{\text{within}}$ are independent, we can apply probability and statistics techniques (beyond the scope of this course) to show that the null sampling distribution of the F statistic is that of the "F-distribution" (see Section 3.9.7). The F-distribution is indexed by two numbers called the numerator and denominator degrees of freedom. This indicates that there are (infinitely) many F-distribution pdf curves, and we must specify these two numbers to select the appropriate one for any given situation.

As we discussed in Section 3.9.7, the F distribution can be represented as the ratio of two independent chi-squared random variables (divided by the ratio of their degrees of freedom). After realizing that variance-like quantities tend to follow chi-squared random variables, it should be intuitive that the F statistic (the ratio of two variance-like quantities divided by their degrees of freedom) indeed follows an F distribution.

Not surprisingly, the null sampling distribution of the F-statistic for any given one-way ANOVA is the F-distribution with numerator degrees of freedom equal to $\text{df}_{\text{between}} = k - 1$ and denominator degrees of freedom equal to $\text{df}_{\text{within}} = N - k$. Note that this indicates that the kinds of F-statistic values we will see if the null hypothesis is true depends only on the number of groups and the numbers of subjects, and not on the values of the population variance or the population

group means. It is worth mentioning that the degrees of freedom are measures of the "size" of the experiment, where bigger experiments (more groups or more subjects) have bigger df.

We can quantify "large" for the F-statistic by comparing the observed value of the F-statistic to its null sampling distribution, which is the specific F-distribution that has degrees of freedom matching the numerator and denominator of the F-statistic.

The F-distribution is a non-negative distribution in the sense that F values, which are squares, can never be negative numbers. The distribution is skewed to the right and continues to have some tiny probability no matter how large F gets. The mean of the distribution is $s/(s-2)$, where s is the denominator degrees of freedom. So if s is reasonably large then the mean is near 1.00, but if s is small, then the mean is larger (e.g., k=2, n=4 per group gives s=3+3=6, and a mean of 6/4=1.5).

Examples of F-distributions with different numerator and denominator degrees of freedom are shown in figure 6.4. These curves are probability density functions, so the regions on the x-axis where the curve is high are the values most likely to occur. And the area under the curve between any two F values is equal to the probability that a random variable following the given distribution will fall between those values. Although very low F values are more likely for, say, the F(1,10) distribution than the F(3,10) distribution, very high values are also more common for the F(1,10) than the F(3,10) values, though this may be hard to see in the figure. The bigger the numerator and/or denominator df, the more concentrated the F values will be around 1.0.

6.2.4 Inference: hypothesis testing

There are two ways to use the null sampling distribution of F in one-way ANOVA: to calculate a p-value or to find the "critical value" (see below).

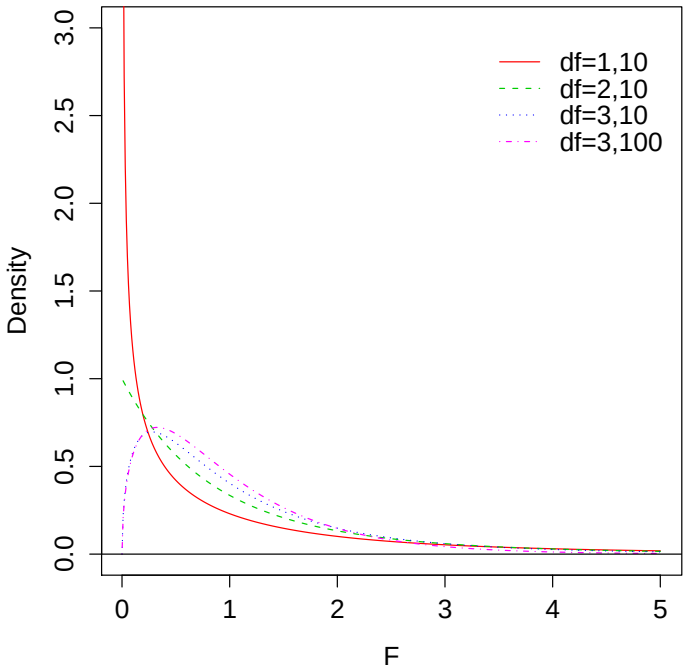


Figure 6.4: A variety of F-distribution pdfs.

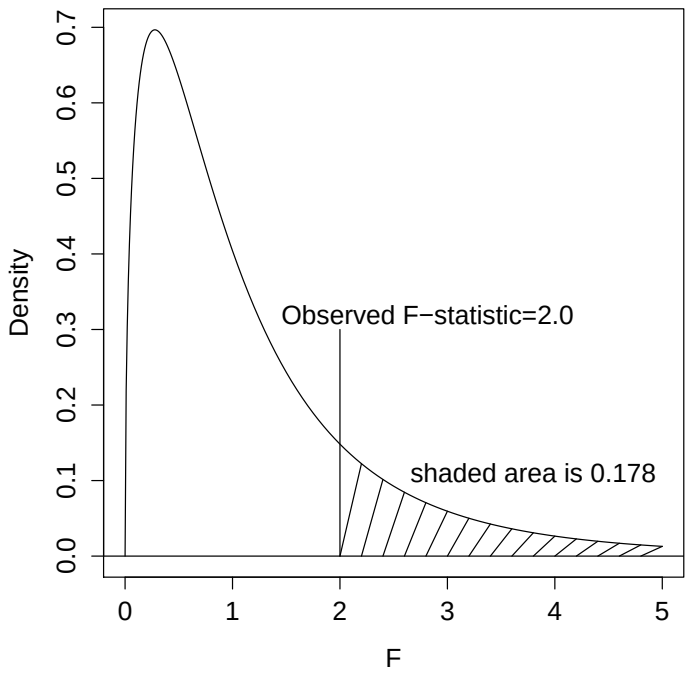


Figure 6.5: The F(3,10) pdf and the p-value for F=2.0.

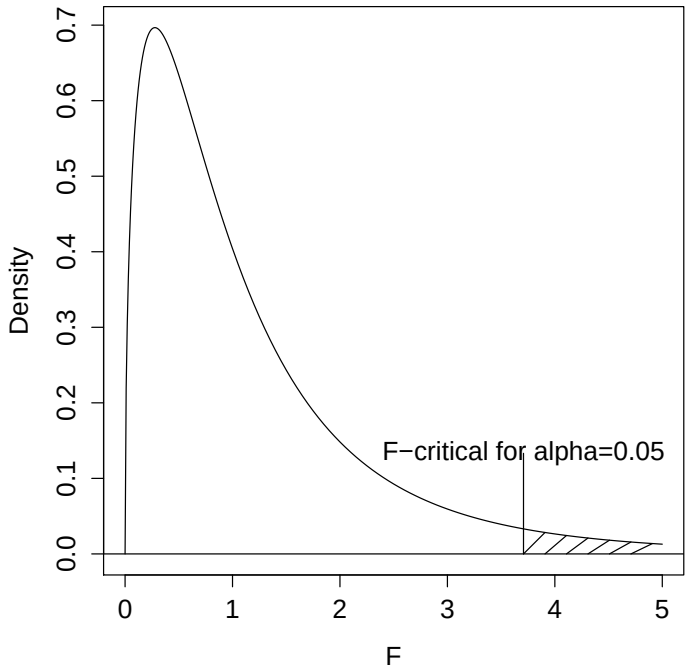


Figure 6.6: The F(3,10) pdf and its alpha=0.05 critical value.

A close up of the F-distribution with 3 and 10 degrees of freedom is shown in Figure 6.5. This is the appropriate null sampling distribution of an F-statistic for an experiment with a quantitative outcome and one categorical explanatory variable (factor) with k=4 levels (each subject gets one of four different possible treatments) and with 14 subjects divided among the 4 groups. A vertical line marks an F-statistic of 2.0 (the observed value from some experiment). The p-value for this result is the chance of getting an F-statistic greater than or equal to 2.0 when the null hypothesis is true, which is the shaded area. The total area is always 1.0, and the shaded area is 0.178 in this example, so the p-value is 0.178 (not significant at the usual 0.05 alpha level).

Figure 6.6 shows another close up of the F-distribution with 3 and 10 degrees of freedom. We will use this figure to define and calculate the **F-critical** value. For a given alpha (significance level), usually 0.05, the F-critical value is the F value above which $100\alpha\%$ of the null sampling distribution occurs. For experiments with 3 and 10 df, and using $\alpha = 0.05$, the figure shows that the F-critical value is 3.71. Note that this value can be obtained from a computer *before* the experiment is run,

as long as we know how many subjects will be studied and how many levels the explanatory variable has. Then when the experiment is run, we can calculate the observed F-statistic and compare it to F-critical. If the statistic is smaller than the critical value, we retain the null hypothesis because the p-value must be bigger than α, and if the statistic is equal to or bigger than the critical value, we reject the null hypothesis because the p-value must be equal to or smaller than α.

6.2.5 Inference: confidence intervals

It is often worthwhile to express what we have learned from an experiment in terms of confidence intervals. In one-way ANOVA, it is possible to make confidence intervals for population group means or for differences in pairs of population group means (or other more complex comparisons). We defer discussion of the latter to Chapter 12. As for the former (making confidence intervals for the population group means), it is quite common for researchers to plot the individual k-many confidence intervals (one for each group mean) and assess if they overlap. However, this kind of assessment is not a formal statistical test and can often give misleading results. For example, if the confidence intervals for two different group means do not overlap, then it is indeed the case that those two group means significantly differ; however, if the two confidence intervals overlap, *the two group means could still significantly differ*. So, merely assessing the overlap among individual confidence intervals is generally not a great way to determine if there are significant differences among group means. We will talk about this further in Chapter 12 as well.

Construction of a confidence interval for a population group means is usually done as an appropriate "plus or minus" amount around a sample group mean. We use $\text{MS}_{\text{within}}$ as an estimate of σ^2, and then for group i, the standard error of the mean is $\sqrt{\text{MS}_{\text{within}}/n_i}$. As discussed in section 5.2.7, the multiplier for the standard error of the mean is the so called "quantile of the t-distribution" which defines a central area equal to the desired confidence level. This comes from a computer or table of t-quantiles. For a 95% CI this is often symbolized as $t_{0.025,df}$ where df is the degrees of freedom of $\text{MS}_{\text{within}}$, $(N-k)$. Construct the CI as the sample mean plus

or minus (SEM times the multiplier).

In a nutshell: In one-way ANOVA we calculate the F-statistic as the ratio $MS_{between}/MS_{within}$. Then the p-value is calculated as the area under the appropriate null sampling distribution of F that is bigger than the observed F-statistic. We reject the null hypothesis if $p \leq \alpha$.

6.3 Do it in R

There are a few ways to run one-way ANOVA in R; the most useful one for our purposes is the `aov()` function in conjunction with the typical `outcomeVariable~groupVariable` syntax we've seen with other functions so far:

```
#run one-way ANOVA
> anovaModel = aov(cooperation~emotion, data = moral)
> summary(anovaModel)
             Df Sum Sq Mean Sq F value Pr(>F)
emotion       2   86.4   43.18   4.495 0.0131 *
Residuals   123 1181.4    9.61
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
```

The `aov()` function implicitly assumes that the outcome variance is equal across groups. If you don't want to make this assumption, you can use the `oneway.test()` function. We will focus on the `aov()` function because it easily tells us the within-group variance and between-group variance decomposition that we have discussed in this chapter. Also, as we'll discuss in Chapter 12, this function can also be used to specify particular planned or unplanned contrasts, which will be especially useful when we consider issues with multiple hypothesis testing.

From the output, we can already spot a p-value that's less than 0.05 (specifically, it's 0.0131), which means that we're going to reject a null hypothesis. But let's take a minute to examine the rest of the output, i.e., the ANOVA table above.

6.4 Reading the ANOVA table

The **ANOVA table** is the main output of an ANOVA analysis. It always identifies the sources of variation (in this case, "emotion" and "Residuals"), as well as the degrees of freedom ("Df"), sum of squares ("Sum Sq"), mean square ("Mean Sq"), F statistic ("F value"), and p-value ("Pr(>F)").

For one-way ANOVA, the F statistic is computed as $\text{MS}_{\text{between}}/\text{MS}_{\text{within}}$. Remember that $\text{MS}_{\text{between}}$ measures the average variation *between* groups, which in this case are defined by the `emotion` variable. Thus, we know from the table that $\text{MS}_{\text{between}} = 43.18$. Meanwhile, the "Residuals" row shows that $\text{MS}_{\text{within}} = 9.61$. Another way to figure out which quantity is which is to observe that the F statistic is greater than one, and thus the bigger MS number must be $\text{MS}_{\text{between}}$.

In the ANOVA table, there is only one p-value, because there is only one (overall) null hypothesis, namely $H_0 : \mu_1 = \cdots = \mu_k$. This p-value comes from comparing the (single) F value to its null sampling distribution. From the "Df" column, we know that the null sampling distribution is the $F_{2,123}$ distribution. Also note that each MS is defined as SS/df; e.g., for the above table, $\text{MS}_{\text{between}} = \frac{\text{SS}_{\text{between}}}{\text{df}_{\text{between}}} = \frac{86.4}{2} = 43.2$.

An ANOVA is a breakdown of the total variation of the data, in the form of SS and df, into smaller independent components. For one-way ANOVA, we break down the deviations of individual values from the overall mean of the data into deviations of the group means from the overall mean (between variation), and then deviations of the individuals from their group means (within variation). The independence of these sources of deviation results in additivity of the SS and df columns (but *not* the MS column). So we note that $\text{SS}_{\text{total}} = \text{SS}_{\text{between}} + \text{SS}_{\text{within}}$ and $\text{df}_{\text{total}} = \text{df}_{\text{between}} + \text{df}_{\text{within}}$.

Note that we can calculate $\text{MS}_{\text{total}} = 1267.78/125 = 10.14$ which is the variance of all of the data (thrown together and ignoring the treatment groups). You can see that MS_{total} is certainly not equal to $\text{MS}_{\text{between}} + \text{MS}_{\text{within}}$.

Another use of the ANOVA table is to learn about an experiment when it is not fully described (or to check that the ANOVA was performed and recorded correctly). Just from this one-way ANOVA table, we can see that there were 3 treatment groups (because $\text{df}_{\text{between}}$ is one less than the number of groups). Also, from df_{total}, we can calculate that there were 125+1=126 subjects in the experiment.

Finally, it is worth knowing that $\text{MS}_{\text{within}}$ is an estimate of σ^2, the variance of outcomes around their group mean. So, we can take the square root of $\text{MS}_{\text{within}}$ to get an estimate of σ, the standard deviation. Then we know that the majority (about $\frac{2}{3}$) of the measurements for each group are within σ of the group mean and most (about 95%) are within 2σ, assuming a Normal distribution. In this example the estimate of the s.d. is $\sqrt{9.61} = 3.10$, so individual subject cooperation values more than 2(3.10)=6.2 coins from their group means would be uncommon.

You should understand the structure of the one-way ANOVA table including that MS=SS/df for each line, SS and df are additive, F is the ratio of between to within group MS, the p-value comes from the F-statistic and its presumed (under model assumptions) null sampling distribution, and the number of treatments and number of subjects can be calculated from degrees of freedom.

6.5 Conclusion about moral sentiments

With $p = 0.013 < 0.05$, we reject the null hypothesis that all three of the group population means of cooperation are equal. We therefore conclude that differences in mean cooperation are caused by the induced emotions, and that among control, guilt, and shame, at least two of the population means differ. However, validly determining *which* groups differ can be tricky, because it can require conducting many different hypothesis tests (especially if there are many groups). This issue is the focus of Chapter 12.

(A complete analysis would also include examination of residuals for additional evaluation of possible non-normality or unequal spread.)

In a nutshell: One-way ANOVA is a generalization of the t-test: It statistically compares $k \geq 3$ group means instead of just 2 group means. To do this, the F-statistic of one-way ANOVA decomposes variation of the data into "between group variation" and "within group variation"; the F-statistic is the ratio of these two measures of variation. If there is a relatively large amount of variation between groups, then the null hypothesis that all the group means are the same is not likely to hold. The p-value of one-way ANOVA is calculated from the F null sampling distribution with matching degrees of freedom. As always, this p-value is only "correct" if the model assumptions are correct. Furthermore, as of now, one-way ANOVA can only conclude whether or not *all* of the group means are equal. If we conclude they are not equal, more nuanced methods will be needed, which we will discuss in depth in later chapters.

Chapter 7

Threats to Your Experiment

Planning to avoid criticism.

One of the main goals of this book is to encourage you to think from the point of view of an experimenter, because other points of view, such as that of a reader of scientific articles or a consumer of scientific ideas, are easy to switch to after the experimenter's point of view is understood, but the reverse is often not true. In other words, to enhance the usability of what you learn, you should pretend that you are a researcher, even if that is not your ultimate goal.

As a researcher, one of the key skills you should be developing is to try, in advance, to think of all of the possible criticisms of your experiment that may arise from the reviewer of an article you write or the reader of an article you publish. This chapter discusses possible complaints about internal validity, external validity, construct validity, Type 1 error, and power.

We are using "threats" to mean things that will reduce the impact of your study results on science, particularly those things that we have some control over.

7.1 Internal validity

In a well-constructed experiment in its simplest form we manipulate variable X and observe the effects on variable Y. For example, outcome Y could be number of people who purchase a particular item in a store over a certain week, and X could be some characteristics of the display for that item, such as use of pictures of people of different "status" for an in-store advertisement (e.g., a celebrity vs. an unknown model). **Internal validity** is the degree to which we can appropriately conclude that the changes in X *caused* the changes in Y.

The study of causality goes back thousands of years, but there has been a resurgence of interest recently. For our purposes we can define **causality** as the state of nature in which an active change in one variable directly changes the probability distribution of another variable. It does not mean that a particular "treatment" is *always* followed by a particular outcome, but rather that some probability is changed, e.g. a higher outcome is more likely with a particular treatment compared to without. A few ideas about causality are worth thinking about now. First, **association**, which is equivalent to non-zero correlation (see section 3.6.1) in statistical terms, means that we observe that when one variable changes, another one tends to change. We cannot have causation without association, but just finding an association is not enought to justify a claim of causation.

Association does not necessarily imply causation.

Why does association not necessarily imply causation? Furthermore, if association does not imply causation, what *does* imply causation? At this point, we should talk about some key distinctions between **experiments** and **observational studies**.

We've discussed experiments throughout this book, so hopefully you are familiar with them at this point. Technically speaking, an experiment (as defined in this book) is a study where subjects are randomly assigned to different levels of treatment (an explanatory variable), with the aim of characterizing how subjects respond to different treatments. If we see a difference in response among the treatments, then we have good reason to claim that the treatments are *causing* the difference in response. Why can we claim this? The reason is that, when we compare the different groups of subjects in the experiment, the *only explanatory*

difference is the treatment to which they were assigned—because we *randomized* subjects to treatment, the different groups should (at least on average) look the same in terms of all other explanatory variables (even explanatory variables we may not have recorded—this is the huge benefit of randomization). Thus, the treatment variable is literally the only explanatory variable that could explain the difference in response among the groups (if there is a significant difference).

Meanwhile, an observational study is exactly the same setup as an experiment, except treatment is *not* randomly assigned, but rather is self-selected by the subjects or assigned by some process that is unknown to us. As a result, the different treatment groups can look very different from each other in terms of many explanatory variables (both observed and unobserved). As a result, it becomes difficult to tell if it is the treatment that is causing the difference in response or some other explanatory variables that happen to be associated with treatment. In this scenario, it is common to say that the treatment is **confounded** with other explanatory variables ("confounded" meaning "mixed up in a confusing way"). There is a famous example of this phenomenon that you should know about—well, it's famous enough to be discussed on the "Correlation does not imply causation" Wikipedia page, at least, and this is the example I was first taught when I learned about observational studies. In the 1990s, epidemiologists kept finding very convincing observational study evidence that hormone replacement therapy (HRT) reduced coronary heart disease (CHD) among women—so convincing, in fact, that the Food and Drug Administration (FDA) approved that HRT can be used to reduce CHD. Then, in the late 1990s, very convincing experimental evidence showed the exact opposite: HRT *increased* CHD among women. As it turns out, in non-experimental settings, women who took HRT tended to be from higher socioeconomic backgrounds than women who did not take HRT, and it was this socioeconomic difference that was causing the difference in CHD. Meanwhile, in the randomized experiments, researchers ensured that the HRT group and non-HRT groups were similar in terms of socioeconomic status (and other factors), such that HRT was the only factor that could explain the difference in CHD between treatment groups.

This brings us to the concept of **internal validity**. Internal validity boils down to differences in explanatory variables among treatment groups (other than the treatment variable): If there are virtually no differences, then there is a high level of internal validity, and if the treatment groups look very different in terms of explanatory variables, then there is low internal validity. The latter scenario has low internal validity because we cannot determine if it is the treatment variable that is causing changes in subjects' responses or the many other explanatory

variables that differ across treatment groups that are causing changes. Because **randomization** tends to make treatment groups look identical (on average) on *all* explanatory variables (observed *and* unobserved!), randomized experiments tend to have a high level of internal validity, and are thus considered the gold standard of scientific inquiry.

However, this does not mean that randomized experiments are fool-proof: For any given randomization, there is always a chance that you get an "unlucky randomization." To go back to the HRT-CHD example, imagine you are running the randomized clinical trial testing the effect of HRT on CHD. Because you've heard that randomized experiments are great, you randomize half the women in your study to HRT and half the women in your study to non-HRT (a placebo). After doing this, you notice that all of the women you assigned to HRT are from higher socioeconomic backgrounds than the women you assigned to non-HRT (there's always a chance that this unlucky randomization happens). Would you continue with the study? If you did, you would conclude that HRT *reduced* CHD (even though, in truth, it increases it), and you'd even have a randomized clinical trial to back up that claim. However, this randomized clinical trial would have low internal validity—it would be just as bad as the aforementioned observational study scenario.

Luckily, we are not at the mercy of unlucky randomizations—we can prevent them by design. If some variables, such as gender or socioeconomic status, are known to be related to the outcomes under study, we can utilize **block randomization**, in which randomization among treatments is performed separately for each level of the other explanatory factors. In short, this makes sure that unlucky randomizations do not happen, and it increases the internal validity of our study. As a result, it makes us more certain that, at the end of the study, we will be able to make a *causal* claim about the treatments we are experimenting upon. (Block randomization is the main design technique we will consider in this course for increasing internal validity, but it becomes less straightforward when there are many explanatory variables to consider in an experiment. One of my main research areas is about increasing the internal validity of experiments in these situations, so I'd love to talk about this if you are interested.)

Randomization is normally done using computerized random number generators. Ideally, all subjects are identified before the experiment begins and assigned numbers from 1 to N (the total number of subjects), and then a computer's random number generator is used to assign treatments to the subjects via these numbers.

For block randomization this can be done separately for each block. If all subjects cannot be identified before the experiment begins, some way must be devised to assure that each subject has an equal chance of getting each treatment (if equal assignment is desired). One way to do this is as follows. If there are k levels of treatment, then collect the subjects until k (or 2k or 3k, etc.) are available, then use the computer to randomly assign treatments among the available subjects.

It is important to note that confounding can still occur even in randomized experiments where treatment is the only explanatory variable—this usually happens because of a bad choice in the treatments to study. As an example of designed confounding, consider the treatments "drug plus psychotherapy" vs. "placebo" for treating depression. If a significant difference in depression is found, then we will not know whether the success of the treatment is due to the drug, the psychotherapy, or their combination. If no difference is found, then that may be due to the effect of drug canceling out the effect of the psychotherapy. If the drug and the psychotherapy are known to individually help patients with depression and we really do want to study the combination, it would probably be better to run a two-way factor experiment, where there are four treatment groups (placebo, only drug, only psychotherapy, and drug and psychotherapy) so that we can assess if the drug adds a benefit to psychotherapy and vice versa. We will talk more about these types of experiments in Chapter 8. As another example, consider a test of the effects of a mixed herbal supplement on memory. Again, a success tells us that something in the mix helps memory, but a follow-up trial is needed to see if all of the components are necessary. And again we have the possibility that one component would cancel another out, causing a "no effect" outcome when one component really is helpful. But we must also consider the possibility that the mix itself is effective while the individual components are not, so this might be a good experiment. In short, when you pick the treatments you want to study in your experiment, you are also constraining the types of questions you can answer, so you should pick your treatments carefully.

Finally, it is important to note that conducting a randomized experiment on your treatments of interest is not always possible. For example, let us say that you are interested in studying the effects of smoking on lung health. Ideally (for your research), you would randomly force some people to smoke for 20+ years, and randomly force others to *not* smoke for all those years, and then record the results. For good or bad, society has deemed such an experiment unethical, so you cannot conduct an experiment like that. However, haven't we as a society declared that smoking *causes* deteriorated lung health? How have we been able to do this,

despite not being able to run a randomized experiment? In short, there are many methods for ascertaining causal effects in observational studies (even in the above HRT-CHD example), but they are outside the scope of this class and will not be discussed further here. (However, estimating causal effects in observational studies is my other main research area, and so I would love to talk with you if you are interested.)

In experiments—as opposed to observational studies—treatment assignment among subjects is under experimenters' control. Ideally, treatment is randomized (or possibly block randomized) to increase internal validity of the study. It is also important to consider what types of treatment should be studied in the experiment.

Blinding (also called masking) is another key factor in internal validity. Blinding indicates that the subjects are prevented from knowing which (level of) treatment they have received. If subjects know which treatment they are receiving and believe that it will affect the outcome, then we may be measuring the effect of the belief rather than the effect of the treatment. In psychology this is called the **Hawthorne effect**. In medicine it is called the **placebo effect**. As an example, in a test of the causal effects of acupuncture on pain relief, subjects may report reduced pain because they believe the acupuncture should be effective. Some researchers have made comparisons between acupuncture with needles placed in the "correct" locations versus similar but "incorrect" locations. When using subjects who are not experienced in acupuncture, this type of experiment has much better internal validity because patient belief is not confounding the effects of the acupuncture treatment. In general, you should attempt to prevent subjects from knowing which treatment they are receiving, if that is possible and ethical, so that you can avoid the placebo effect (prevent confounding of belief in effectiveness of treatment with the treatment itself), and ultimately prevent valid criticisms about the interval validity of your experiment. On the other hand, when blinding is not possible, you must always be open to the possibility that any effects you see are due to the subjects' beliefs about the treatments.

Double blinding refers to blinding the subjects and also assuring that the *experimenter* does not know which treatment the subject is receiving. For example, if the treatment is a pill, a placebo pill can be designed such that neither the subject nor the experimenter knows what treatment has been randomly assigned

to each subject. This prevents confounding in the form of difference in treatment application (e.g., the experimenter could subconsciously be more encouraging to subjects in one of the treatment groups) or in assessment (e.g, if there is some subjectivity in assessment, the experimenter might subconsciously give better assessment scores to subjects in one of the treatment groups). Of course, double blinding is not always possible, and when it is not used you should be open to the possibility that any effects you see are due to differences in treatment application or assessment by the experimenter.

Triple blinding refers to not letting the person doing the statistical analysis know which treatment labels correspond to which actual treatments. Although rarely used, it is actually a good idea because there are several places in most analyses where there is subjective judgment involved, and a biased analyst may subconsciously make decisions that push the results toward a desired conclusion. The label "triple blinding" is also applied to blinding of the rater of the outcome in addition to the subjects and the experimenters (when the rater is a separate person).

In a nutshell: Blinded, randomized experiments where treatment precedes the outcome is usually the best way to limit confounding and thus increase the internal validity of a study. Statistically significant association *can* imply causation if there is a high level of internal validity in the study.

7.2 External validity

External validity is synonymous with **generalizability**. When we perform an ideal experiment, we randomly choose subjects (in addition to randomly assigning treatment) from a population of interest. Examples of populations of interest are all college students, all reproductive aged women, all teenagers with type I diabetes, all 6 month old healthy Sprague-Dawley rats, all workplaces that use

Microsoft Word, or all cities in the Northeast with populations over 50,000. If we randomly select our experimental units from the population such that each unit has the same chance (or with special statistical techniques, a fixed but unequal chance) of ending up in our experiment, then we may appropriately claim that our results apply to that population. In many experiments, we do not truly have a random sample of the population of interest. In so-called "convenience samples", e.g., "as many of my classmates as I could attract with an offer of a free slice of pizza", the population these subjects represent may be quite limited.

After you complete your experiment, you will need to write a discussion of your conclusions, and one of the key features of that discussion is your set of claims about external validity. First, you need to consider what population your experimental units truly represent. In the pizza example, your subjects may represent students who like free food (and/or are low-income) and don't mind participating in experiments. In short: While internal validity is about the similarity of different treatment groups, external validity is about the similarity of your sample and your population of interest. If your sample is highly similar to (or representative of) your population of interest, then your study has a high level of external validity; if the sample and population are very dissimilar, there is a low level of external validity. Note that it is possible to have a high level of internal validity but a low level of external validity: In this case, you have a lot of evidence for causal claims *about the sample*, but those causal claims may not be generalizable to your population of interest. However, even with low external validity, researchers can make arguments that their results are generalizable (e.g., pregnant women are often not included in clinical trials, but there is reason to believe many health effects observed in clinical trials would apply to pregnant women), but such arguments hinge on scientific arguments rather than statistical evidence. In contrast, observational studies can often be viewed as having a high level of external validity but a low level of internal validity—i.e., we can estimate population-level associations, but not necessarily population-level causal effects. To estimate causal effects in observational studies, researchers often have to trade external validity for internal validity by focusing on a subsample that is similar across treatment groups but not necessarily representative of the population of interest.

There are three phenomena that often lead to non-generalizability (poor external validity) and are thus worth more discussion. First is

non-participation. If you randomly select subjects, e.g., through phone records, or college e-mail, then some subjects may decline to participate. You should always consider the very real possibility that the decliners are different in one or more ways from the participators, and thus your results do not really apply to the population of interest.

A second problem is dropout, which is when subjects who start a study do not complete it. Dropout can affect both internal and external validity. Dropout affects internal validity if certain treatments have negative side effects on certain subpopulations. For example, if a treatment has negative interactions with alcohol, then those who tend to drink may drop out of the treatment group but not the placebo group. The simplest form of dropout that affects external validity is when subjects who are too busy or less committed drop out because of the length or burden of the experiment. For example, subjects who have to take the bus to get to the experiment site may dropout due to inconvenience. This type of dropout reduces the population to which generalization can be made, and in experiments that collect many measurements over time—such as those studying the effects of ongoing behavioral therapy on adjustment to a chronic disease—this can be a critical blow to external validity.

The third special form of non-generalizability relates to the terms efficacy and effectiveness in the medical literature. In this case, the lack of generalizability refers to the environment and the details of treatment application rather than the subjects. If a well-designed clinical trial is carried out under high controlled conditions in a tertiary medical center, and finds that drug X cures disease Y with 80% success (i.e., it has high efficacy), then we are still unsure whether we can generalize this to real clinical practice in a doctor's office (i.e, whether the treatment has high effectiveness). Even outside the medical setting, it is important to consider expanding spheres of environmental and treatment application variability.

External validity (generalizability) relates to the breadth of the population we have sampled and how well we can justify extending our results to an even broader population.

7.3 Construct validity

Once we have made careful operational definitions of our variables and classified their types, we still need to think about how useful they will be for testing our hypotheses. **Construct validity** is a characteristic of devised measurements that describes how well the measurement can stand in for the scientific concepts or "constructs" that are the real targets of scientific learning and inference.

Construct validity addresses criticisms like "you have shown that changing X causes a change in measurement Y, but I don't think you can justify the claims you make about the causal relationship between concept W and concept Z", or "Y is a biased and/or unreliable measure of concept Z".

The classic paper on construct validity is *Construct Validity in Psychological Tests* by Lee J. Cronbach and Paul E. Meehl, first published in Psychological Bulletin, 52, 281-302 (1955). Construct validity in that article is discussed in the context of four types of validity. For the first two, it is assumed that there is a "gold standard" against which we can compare the measure of interest. The simple correlation (see section 3.6.1) of a measure with the gold standard for a construct is called either concurrent validity if the gold standard is measured at the same time as the new measure to be tested or predictive validity if the gold standard is measured at some future time. Content validity is a bit ambiguous but basically refers to picking a representative sample of items on a multi-item test. Here we are mainly concerned with construct validity, and Cronbach and Meehl state that it is pertinent whenever the attribute or quality of interest is not "operationally defined". That is, if we define happiness to be the score on our happiness test, then the test is a valid measure of happiness by definition. But if we are referring to a concept without a direct operational definition, we need to consider how well our test stands in for the concept of interest. This is the construct validity. Cronbach and Meehl discuss the theoretical basis of construct validity for psychology, and this should be applicable to

other social sciences. They also emphasize that there is no single measure of construct validity, because it is a complex, often judgment-laden set of criteria.

To assess contruct validity, you should be sure that your measure correlates with other measures for which it should correlate if it is a good measure of the concept of interest. For example, if there is a "gold standard" in your field, then your measure should have a high correlation with that standard, at least in the kinds of situations where you will be using it. And it should not be correlated with measures of other unrelated concepts.

It is worth noting that good construct validity doesn't mean much if your measure is not also reliable. A good measure should not depend strongly on who is administering the test (called high inter-rater reliability), and repeat measurements should have a small statistical "variance" (called test-retest reliability).

Most of what you will be learning about construct validity must be left to reading and learning in your specific field, but a few examples are given here. In public health studies, a measure of obesity is often desired. In particular, researchers are interested in a measure of obesity that is a good predictor of health outcomes—otherwise, the measure will not be useful for making claims relevant to public health, i.e., it won't have good construct validity for the concept of "health". The United States Center for Disease Control (CDC) has classifications for obesity based on the Body Mass Index (BMI), which is a formula involving only height and weight. The BMI is a simple substitute that has reasonably good concurrent validity for more technical definitions of body fat such as percent total body fat, which can be better estimated by more expensive and time-consuming methods such as a buoyancy method. But even total body fat percent may be insufficient, because some health outcomes may be better predicted by information about the amount of fat at specific locations. Beyond these problems, the CDC assigns labels (underweight, health weight, at risk of overweight, and overweight) to specific

ranges of BMI values, but the cutoff values—while partially based on scientific methods—are also partly arbitrary. And surely the "best" cutoff for predicting outcomes will vary depending on the outcome, e.g., heart attack, stroke, teasing at school, or poor self-esteem. So, although there is some degree of validity to these categories of BMI (e.g., as shown by different levels of disease for people in different categories and correlation with buoyancy tests) there is also some controversy about the construct validity.

As another classic example of construct validity: Is the Stanford-Binet "IQ" test a good measure of "intelligence"? Many gallons of ink have gone into discussion of this topic. Low variance for individuals tested multiple times shows that the test has high test-retest validity, and as the test is self-administered and objectively scored there is no issue with inter-rater reliability. There have been numerous studies showing good correlation of IQ with various outcomes that "should" be correlated with intelligence, such as future performance on various tests. On the other hand, the test has been severely criticized for cultural and racial bias. And other critics claim there are multiple dimensions to intelligence, not just a single "intelligence" factor. In summation, the IQ test as a measure of the construct "intelligence" is considered by many researchers to have low construct validity.

Construct validity is important because it makes us think carefully whether the measures we use really stand in well for the concepts that label them.

7.4 Maintaining Type 1 error

Type 1 error is related to the statistical concept that in the real world of natural variability we cannot be certain about our conclusions from an experiment. A Type 1 error is a claim that a treatment is effective, i.e., we decide to reject the null hypothesis, when the null hypothesis is actually true. Obviously in any single real situation, we cannot know whether or not we have made a Type 1 error: if we knew the absolute truth, we would not make the error.

As explained in more detail in several other chapters, statistical inference is the process of making appropriately qualified claims in the face of uncertainty.

Type 1 error deals with the probabilistic validity of those claims. When we make a statement such as "we reject the hypothesis that the mean outcome is the same for both the placebo and the active treatments with alpha equal to 0.05" we are claiming that the procedure we used to arrive at our conclusion only leads to false positive conclusions 5% of the time *when the truth happens to be that there is no difference in the effect of treatment on outcome*. This is *not at all* the same as the claim that there is only a 5% chance that any "reject the null hypothesis decision" will be the wrong decision!

Maintaining Type 1 error means doing all we can to assure that the false positive rate really is set to whatever nominal level (usually 5%) we have chosen. This will be discussed much more fully in future chapters, but it basically involves choosing an appropriate statistical procedure and assuring that the assumptions of our chosen procedure are reasonably met. Part of the latter is verifying that we have chosen an appropriate model for our data (see Section 5.2.2).

A special case of not maintaining Type 1 error is "data snooping". For example, it is very common to perform many different analyses of the same dataset, each with a nominal Type 1 error rate of 5%, and then report just the one(s) with p-values less than 0.05, but this is not the appropriate way to analyze an experiment and maintain Type 1 error. As seen in Section 12.4, this approach to data analysis results in a much larger chance of making false conclusions.

Using models with broken assumptions and/or data snooping tend to result in an increased chance of making false claims in the presence of ineffective treatments (i.e., an increased chance of making a Type 1 error).

7.5 Power

The **power** of an experiment refers to the probability that we will correctly conclude that the treatment caused a change in the outcome. If some particular true non-zero difference in outcomes is caused by the active treatment, and you have low power to detect that difference, you will probably make a Type 2 error (have a "false negative" result) in which you conclude that the treatment was ineffective,

when it really was effective. The Type 2 error rate, often called "beta" (β), is the fraction of the time that a conclusion of "no effect" will be made (over repeated similar experiments) when some true non-zero effect is really present. The power is equal to $1-\beta$. It is very important to note that power (and the Type 2 error rate) depends on the alternative—e.g., the size of treatment effects that are present. Intuitively, large effect sizes will be easy to detect (i.e., you will have a lot of power in detecting them) while small effect sizes will be difficult to detect.

Before the experiment is performed, you have some control over the power of your experiment, so you should estimate the power for various reasonable effect sizes and, whenever possible, adjust your experiment to achieve reasonable power (e.g., at least 80%). If you perform an experiment with low power, you are just wasting time and money! See Chapter 11 for details on how to calculate and increase the power of an experiment.

Power is the chance of getting a statistically significant result when a particular real treatment effect exists. Including a sufficient number of subjects in an experiment is the most well-known way to assure sufficient power.

In addition to sample size, the main (partially) controllable experimental characteristic that affects power are sources of variability. If you can reduce the overall variability in a study, you can increase power. Different **sources of variation**. are measurement, environmental, treatment application, and subject-to-subject variation. We describe these different sources of variation below.

Measurement variation refers to differences in repeat measurement values when they should be the same. (Sometimes repeat measurements should change, for example the diameter of a balloon with a small hole in it in an experiment of air leakage.) Measurement variability is usually quantified as the standard deviation of many measurements of the same thing. Researchers also often use the term **precision**, which is the inverse of variance—i.e., precision equals 1/variance. So, high precision implies a low variance (and thus standard deviation). It is worth knowing that a simple and usually cheap way to improve measurement precision is to make repeated measurements and take the mean; this mean is less variable than an individual measurement. Another inexpensive way to improve precision, which should almost always be used, is to have good explicit procedures for making

the measurement and good training and practice for whoever is making the measurements. Other than possibly increased cost and/or experimenter time, there is no down-side to improving measurement precision, so it is an excellent way to improve power.

Controlling environmental variation is another way to reduce the variability of measurements, and thus increase power. For each experiment you should consider what aspects of the environment (broadly defined) can and should be controlled (fixed or reduced in variation) to reduce variation in the outcome measurement. For example, if we want to look at the effects of a hormone treatment on rat weight gain, controlling the diet, the amount of exercise, and the amount of social interaction (such as fighting) will reduce the variation of the final weight measurements, making any differences in weight gain due to the hormone easier to see. Other examples of environmental sources of variation include temperature, humidity, background noise, lighting conditions, etc. As opposed to reducing measurement variation, there is often a down-side to reducing environmental variation. There is usually a trade-off between reducing environmental variation which increases power but may reduce external validity (see above). In short, controlling environmental factors may reduce variation in a study, environmental factors are not controlled for in the wild, and thus a well-controlled study may have less generalizability.

The trade-off between power and external validity also applies to treatment application variation. While some people include this in environmental variation, it is worth noting that treatment application can be controlled in an experiment. In a particular experiment, treatment may vary in its quality or quantity among subjects assigned to the same (nominal) treatment. For example, consider an experiment where one treatment group gets 100mg of a drug. If two drug manufacturers have different production quality such that all of the pills from the first manufacturer have a mean of 100 mg and s.d. of 5 mg, while the second has a mean of 100 mg and s.d. of 20 mg, the increased variability of the second manufacturer will result in decreased power to detect any true differences between the 100 mg dose and any other doses studied. For treatments like "behavioral therapy," decreasing variability is done by standardizing the number of sessions and having good procedures and training. On the other hand, there may be a concern that too much control of variation in a treatment like behavioral therapy might make the experiment unrealistic (i.e., reduce external validity) for the same reasons discussed in the previous paragraph on environmental variation.

Finally, there is subject-to-subject variability. Remember that ideally we choose a population from which we draw our participants for our study (as opposed to using a "convenience sample"). If we choose a broad population like "all Americans" there is a lot of variability in age, gender, height, weight, intelligence, diet, etc. some of which are likely to affect our outcome (or even the difference in outcome between the treatment groups). If we choose to limit our study population for one or several of these traits, we reduce variability in the outcome measurement (for each treatment group) and improve power, but always at the expense of generalizability. As in the case of environmental and treatment application variability, you should make an intelligent, informed decision about trade-offs between power and generalizability in terms of choosing your study population.

On the other hand, researchers often use methods like regression adjustment to account for variability due to explanatory variables among subjects, and—when appropriately used—this can be a great way to increase the power of an experiment in the analysis stage rather than the design stage. Regression adjustment is discussed in depth in Chapter 9.

But there are also experimental design choices that address subject-to-subject variation in ways that improve power without reducing generalizability. For example, the use of a **within-subjects design**—in which each subject receives two or more treatments—is often an excellent way to improve power, although it is not applicable in all cases. See Chapter 13 for more details. Remember that you must change your analysis procedures to ones which do not assume independent errors if you choose a within-subjects design.

> Using the language of Section 3.6, it is useful to think of all measurements as being conditional on whatever environmental and treatment variables we choose to fix, and marginal over those that we let vary.

Reducing variability improves power. In some circumstances this may be at the expense of decreased generalizability. Reducing measurement error and/or use of within-subjects designs usually improves power without sacrificing generalizability.

As mentioned at the beginning of this section, the strength of your treatments (actually, the difference in true outcomes between treatments) strongly affects power. If you are studying very weak treatments—e.g., the effects of one ounce of beer on driving skills, or 1 microgram of vitamin C on catching colds, or one treatment session on depression severity—then you will not have much power to detect the resulting treatment effects (if they indeed exist). For this reason, researchers often recommend that you study strong treatments, because the resulting effects will be easier to detect. That said, it depends what is of scientific interest: If it is indeed your primary interest to study treatments that may have small but non-negligible effect sizes, that is fine, but then a lot of resources (e.g., a large sample) must go into an experiment in order to have the power to detect those small effects.

Power depends on the alternative hypothesis: In short, strong treatments are easier to detect; thus, increasing treatment strength increases power.

Another way to improve power without reducing generalizability is to employ **blocking**. Blocking involves using subject-matter knowledge to select one or more categorical factors whose effects are not of primary importance, but whose levels define homogeneous groups called "blocks". Homogeneity of the blocks is key: A variable will be a good blocking factor if there is not much variation in outcomes within a particular block but there is a lot of variation across blocks. In other words, a good blocking factor correctly identifies sources of variation in an experiment. Common blocking factors may be gender, education level, age, or socioeconomic status, but it will depend on the application at hand. When employing blocking, the blocking factor is "adjusted for" or "controlled for"—i.e., the variation across blocks is accounted for—in the analysis stage. For example, in ANOVA, block can be included as an additional explanatory variable beyond the primary treatment of interest, and inclusion of the block factor tends to improve power if the blocks are more homogeneous within blocks than between blocks. Blocking does not sacrifice generalizability because it does not restrict the types of subjects that are used for analysis; rather, it categorizes the subjects into different blocks, and takes that categorization into account in the analysis stage.

Also, when you hear "blocking," this should ring a bell about a term you heard earlier in this chapter: Block randomization (Page 157). Block randomization

divides subjects into blocks and randomizes treatment within each block; so, the blocking factor is accounted for in the design stage of the experiment, rather than the analysis stage. Without getting too technical, it is preferable to account for blocks in the design stage via block randomization, but if this cannot be done, using blocking in the analysis stage is a good alternative.

So, blocking is a way to account for categorical explanatory variables in the analysis of an experiment. But what about other explanatory variables (e.g., continuous ones)? A natural extension of blocking is a model that allows for multiple **control variables** (also called covariates) that can be any type (categorical, continuous, etc.) The most basic form of this kind of model is **linear regression**, and we will discuss that method in Chapter 9.

Blocking and use of control variables are good ways to improve power without sacrificing generalizability.

7.6 Missing explanatory variables

Another threat to your experiment is not including important explanatory variables. For example, if a treatment raises the mean outcome in males and lowers it in females, then not including gender as an explanatory variable (i.e., including its interaction with treatment) will give misleading results. (See Chapters 10 and 8 for more on interaction.) In other cases—where there may be no interaction between treatment and other variables—ignoring important explanatory variables can nonetheless decrease power (rather than directly causing misleading results). Explanatory variables should especially be taken into account when there are large imbalances among treatment groups—but at the same time, large imbalances lead to less internal validity, which cannot always be circumvented by adjusting for imbalances during the analysis stage. (Randomizing treatment will balance explanatory variables on average, but for any particular randomization, there are likely *some* imbalances that could be taken into account during the analysis stage.)

An extreme case of a missing variable is **Simpson's paradox**. Described by Edward H. Simpson and others, this term describes the situation where the observed effect is in opposite directions for all subjects as a single group (defined based

	Small Stones		Large Stones		Combined	
Treatment A	81/87	**0.93**	192/263	**0.79**	273/350	**0.78**
Treatment B	234/270	**0.87**	55/80	**0.69**	289/350	**0.83**

Table 7.1: Simpson's paradox in medicine

on a variable other than treatment) vs. separately for each group. It only occurs when the fraction of subjects in each group differs markedly between the treatment groups. A nice medical example comes comes from the 1986 article *Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy* by C. R. Chang, et al. (Br Med J 292 (6524): 879-882) as shown in table 7.1.

The data show the number of successes divided by the number of times the treatment was tried for two treatments for gall stones. The "paradox" is that for "all stones" (combined) Treatment B is the better treatment (has a higher success rate), but if the patients' gall stones are classified as either "small" or "large", then Treatment A is better. There is nothing artificial about this example; it is based on the actual data. And there is really nothing "statistical" going on (in terms of randomness); we are just looking at the definition of "success rate". If stone size is omitted as an explanatory variable, then Treatment B looks to be the better treatment, but for each stone size, Treatment A was the better treatment. Which treatment would you choose? If you have small stones or if you have large stones (the only two kinds), you should choose treatment A. Dropping the important explanatory variable gives a misleading ("marginal") effect, when the "conditional" effect is more relevant. Ignoring the confounding (also called lurking) variable "stone size" leads to misinterpretation. What's going on in this example is that there is a lower success rate for large stones than for small stones (which is intuitive—larger stones are more difficult to address), and it just so happens that the majority of Treatment A cases were large stones. Meanwhile, the majority of Treatment B cases were small stones. Thus, when we look at the marginal success rate between Treatment A and B, this is confounded by the marginal success rates of small stones and large stones.

It's worth mentioning that we can go too far in including explanatory variables. This is both in terms of the "multiple comparisons problem" and something called the "bias-variance trade-off". The former artificially raises our Type 1 error if uncorrected but lowers our power if corrected. The latter, in this context, can

come into play when we account for too many unimportant explanatory variables, which can reduce power. We will talk about these issues in Chapter 9 and 12.

Missing explanatory variables can decrease power and/or cause misleading results.

7.7 Threat summary

After you have completed and reported your experiment, your critics may complain that some confounding factors may have destroyed the internal validity of your experiment; that your experiment does not really tell us about the real world concepts of interest because of poor construct validity; that your experimental results are only narrowly applicable to certain subjects or environments or treatment application setting; that your statistical analysis did not appropriately control Type 1 error (if you report "positive" results); or that your experiment did not have enough power (if you report "negative" results). In fact, after reading this chapter, you may be this critic when you are reading papers and reports about experimental results! Whether you are conducting an experiment or simply reading about one, you should always consider possible threats to experimental results. Much of the rest of this book discusses how to deal with—and balance solutions to—these threats.

In a nutshell: If you learn about the various categories of threats to experiments, you will be in a better position to make choices that balance competing risks when designing an experiment. You will also be better able to interpret experimental results in a critical and constructive way.

Chapter 8

Two-Way ANOVA

An analysis method for a quantitative outcome and two categorical explanatory variables.

Two-way ANOVA is the most common analysis for experiments that have a quantitative outcome and two categorical explanatory variables, where each experimental unit (subject) can be exposed to any combination of one level of one explanatory variable and one level of the other explanatory variable. For example, let's say that one explanatory variable is categorical with k possible levels, and the other explanatory variable is also categorical with m possible levels. Thus, there are $k \times m$ possible treatment combinations, and each subject is randomized to experience only one of these combinations.

There are many alternative names for two-way ANOVA, but they all represent settings that involve categorical explanatory variables. One common naming convention for a model incorporating a k-level categorical explanatory variable and an m-level categorical explanatory variable is "k by m ANOVA" or "$k \times m$ ANOVA". ANOVA with more than two explanatory variables is often called **multi-way ANOVA**. An experiment with multiple categorical treatment variables is often called a "factorial experiment," where each treatment is considered a "factor," and multi-way ANOVA is often used to analyze such experiments.

Similar to t-tests and one-way ANOVA, the statistical model in two-way ANOVA assumes that errors are Normally distributed with equal variance for all subjects. Again, we will call that common variance σ^2. Similar to t-tests and one-way ANOVA, we assume independent errors.

Two-way (or multi-way) ANOVA is an appropriate analysis method for a study with a quantitative outcome and two (or more) categorical explanatory variables. The usual assumptions of Normality, equal variance, and independent errors apply.

Because there are two explanatory variables, an explanatory variable's effect on the outcome may either (a) not depend on the level of the other explanatory variable (an additive model), or (b) depend on the level of the other explanatory variable (an interactive model). The structural model for two-way ANOVA *with* interaction is that each combination of levels of the explanatory variables has its own population mean with no restrictions on the patterns. One common notation is to call the population mean of the outcome for subjects with level a of the first explanatory variable and level b of the second explanatory variable as μ_{ab}. The interaction model says that any pattern of μ's is possible. In contrast, the no-interaction (additive) model does have a restriction on the population means of the outcomes. Specifically, the **additive model** assumes that the effect of one explanatory variable on the outcome is the same for every level of the other explanatory variable. Mathematically, this means that the additive model assumes $\mu_{ac} - \mu_{bc} = \mu_{ad} - \mu_{bd}$ for any levels a, b, c, and d. In English, the previous equation states, "The effect of a-versus-b for the first factor is the same regardless of whether the second factor is set to level c or level d." By adding and subtracting terms, we can see that this also implies $\mu_{ac} - \mu_{ad} = \mu_{bc} - \mu_{bd}$ (i.e., the effect of c-versus-d for the second factor is the same regardless of whether the first factor is set to level a or b).

In case mathematical and English explanations do not resonate with you, Figure 8.1 gives a visual explanation of the additive model for two-way ANOVA. Figure 8.1 plots population means for different combinations of two factors, and the same information is shown in both panels of the figure. In each panel, the mean outcome is shown on the y-axis, the levels of one factor are shown on the x-axis, and separate colors are used for the second factor. The second panel reverses the roles of the factors from the first panel. Each point is a population mean of the outcome for a combination of one level from factor A and one level from factor B. The lines are shown as dashed because the explanatory variables are categorical, so interpolation "between" the levels of a factor makes no sense. The parallel nature of the dashed lines is what tells us that these means have a relationship that can be called additive. Also the choice of which factor is placed on the x-axis does not

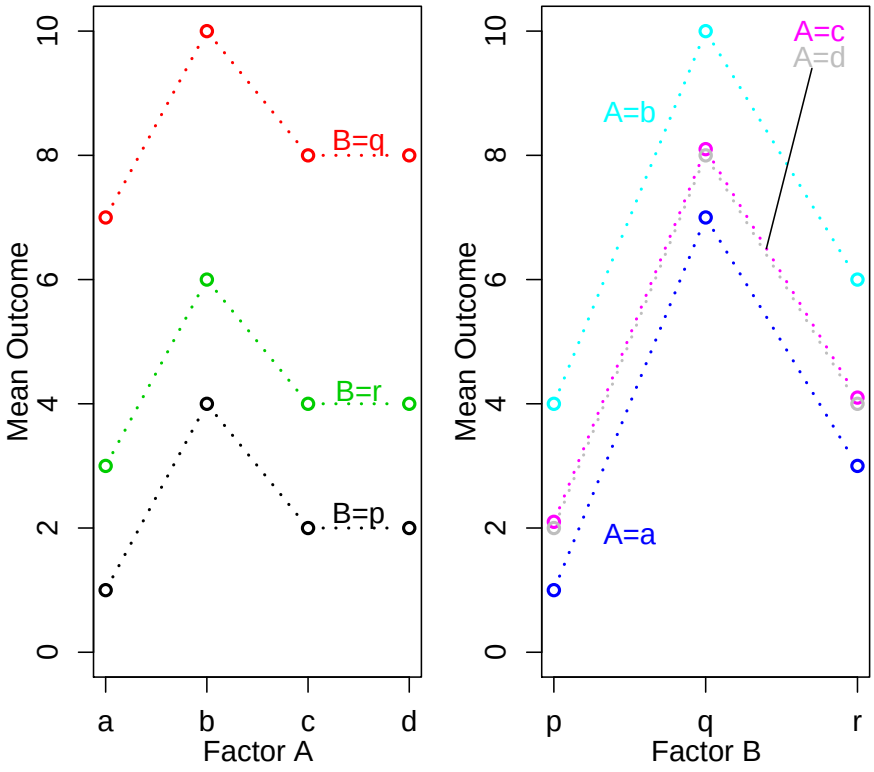


Figure 8.1: Population means for a no-interaction two-way ANOVA example.

affect the interpretation, but commonly the factor with more levels is placed on the x-axis. Using this figure, you should now be able to understand the equations of the previous paragraph. In either panel, the change in outcome (vertical distance) is the same if we move between any two horizontal points along any dotted line. In other words, the lines are parallel, meaning that there is no interaction between the factors.

The two possible means models for two-way ANOVA are the additive model and the interaction model. The additive model assumes that the effects of one explanatory variable on the outcome *does not* depend on the level of the other explanatory variable. An interaction model is needed when the effects of one explanatory variable *does* depend on the level of the other explanatory variable.

A **profile plot**, also called an **interaction plot**, is very similar to Figure 8.1, but instead the points represent the *estimates* of the population means for some data rather than the (unknown) true values. Because we can fit models with or without an interaction term, the same data will show different profile plots depending on which model we use. It is very important to realize that a profile plot from fitting a model without an interaction always shows the best possible parallel lines for the data, regardless of whether an additive model is adequate for the data, so this plot should not be used as EDA for choosing between the additive and interaction models. On the other hand, the profile plot from a model that includes the interaction shows the actual sample means, and is useful EDA for choosing between the additive and interaction models.

A profile plot is a way to look at outcome means for two factors simultaneously. The lines on this plot are meaningless, and only are an aid to viewing the plot. A plot drawn with parallel lines (or for which, given the size of the error, the lines could be parallel) suggests an additive model, while non-parallel lines suggests an interaction model.

8.1 Pollution Filter Example

This example comes from a statement by Texaco, Inc. to the Air and Water Pollution Subcommittee of the Senate Public Works Committee on June 26, 1973. Mr. John McKinley, President of Texaco, cited an automobile filter developed by Associated Octel Company as effective in reducing pollution. However, questions had been raised about the effects of filters on vehicle performance, fuel consumption, exhaust gas back pressure, and silencing. On the last question, he referred to the data in CarNoise.dat as evidence that the silencing properties of the Octel filter were at least equal to those of standard silencers.

This is an experiment in which the treatment "filter type" with levels "standard" and "octel" are randomly assigned to the experimental units, which are cars. Three types of experimental units are used, a small, a medium, or a large car, presumably representing three specific car models. The outcome is the quantitative (continuous) variable "noise". The categorical experimental variable "size" could be considered a blocking variable, but it is also reasonable to consider it to be an additional variable of primary interest, although of limited generalizability due to the use of a single car model for each size. We could also view this experiment as three mini experiments (one for each car size).

A reasonable (initial) statistical model for these data is that for any combination of size and filter type, the noise outcome is normally distributed with equal variance. We also can assume that the errors are independent if there is no serial trend in the way the cars are driven during the experiment or if there is no "drift" in the accuracy of the noise measurement over the duration of the experiment.

The means part of the structural model is either the additive model or the interaction model. We could either use EDA to pick which model to try first, or we could check the interaction model first, then switch to the additive model if the interaction term is not statistically significant.

Because we have two categorical variables (size and filter type), one useful form of EDA is a contingency table (i.e., cross-tabulation) using the `table()` function:

```
> table(carNoise$SIZE, carNoise$TYPE)

        octel standard
large       6        6
medium      6        6
small       6        6
```

The cross-tabulation shows us the number of subjects in our data that belong to each combination of size and treatment. In this case, the same number of subjects belong to each combination; this is called a **balanced design**. One of the key features of this experiment which tells us that it is okay to use the assumption of independent errors is that a different subject (car) is used for each test (row in the data). (Thus, 36 cars were used in this study, as can be seen by adding up the individual entries in the above table.) This is called a **between-subjects design**, and is the same as all of the studies described up to this point in the book, as contrasted with a within-subjects design in which each subject is exposed to multiple treatments (levels of the explanatory variables). For this experiment an appropriate within-subjects design would be to test each individual car with both types of filter, in which case a different analysis (called within-subjects ANOVA) would be needed.

Now let's make some EDA involving the outcome variable, noise. Given a quantitative outcome variable and two categorical variables, one useful form of EDA is side-by-side boxplots, where there is one box for each combination of the two categorical variables (see Figure 8.2). Here's the code used to make Figure 8.2:

```
#side-by-side boxplot
> boxplot(NOISE ~ SIZE + TYPE, data = carNoise,
+     col = rep( c("red", "purple", "blue"), 2 ))
```

Note that we changed the `col` argument within the `boxplot()` function such that boxes corresponding to the same `SIZE` had the same color, thereby making comparisons in the plot easier.

The boxplots in Figure 8.2 show that the small and medium sized cars have more noise than the large cars. It appears that the Octel filter reduces the median noise level for medium sized cars and is equivalent to the standard filter for small and large cars. We also see that, for all three car sizes, there is less car-to-car variability in noise when the Octel filter is used.

One last useful form of EDA is an error bar plot, as shown in Figure 8.3. An error bar plot displays the sample mean and confidence interval (for the outcome variable) for each combination of the two categorical explanatory variables. The standard deviations and sample sizes for each of the six groups are used separately to construct the confidence intervals, but this is less than ideal if the equal variance assumption is met, in which case a pooled standard deviation is better. When interpreting this plot, remember that non-overlapping confidence intervals

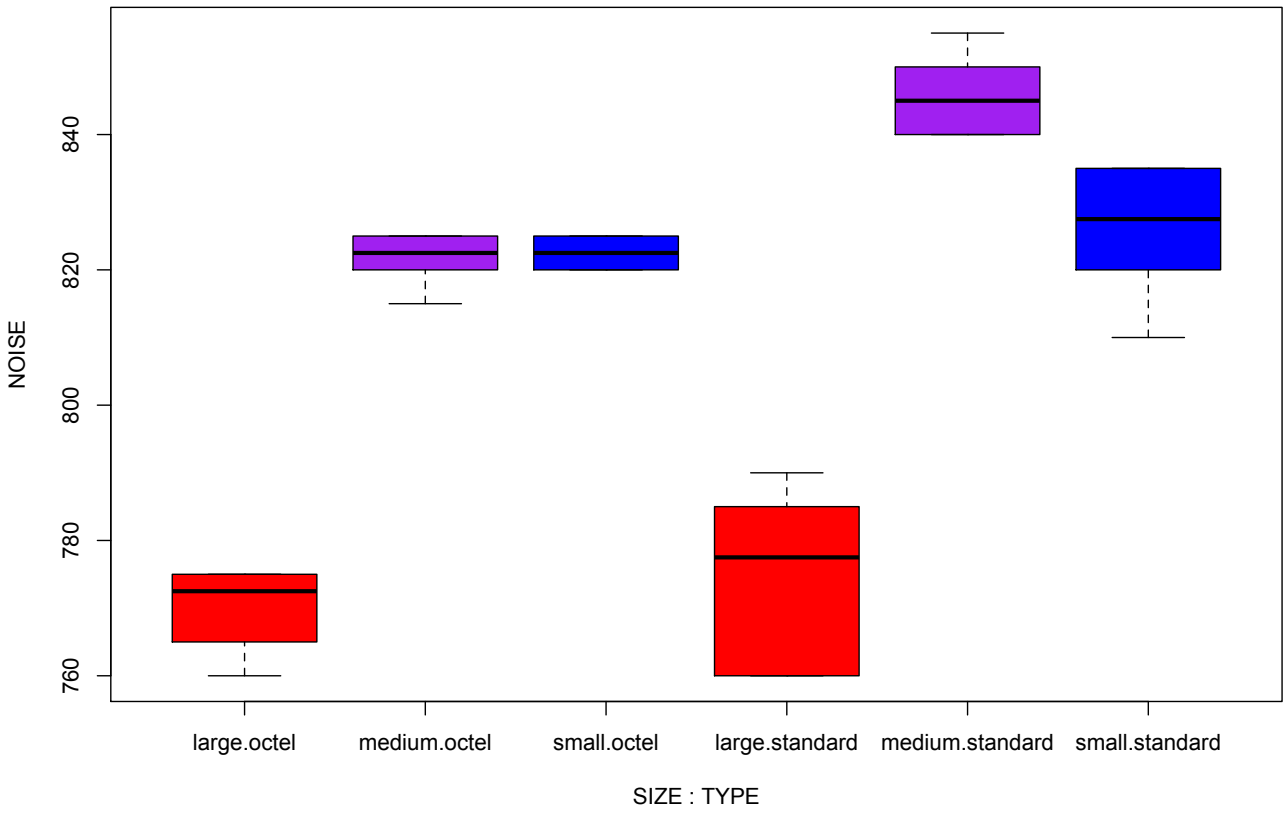


Figure 8.2: Side-by-side boxplots for car noise example.

denote statistically significant differences, but overlapping confidence intervals do not necessarily denote non-significant differences. For example, within each filter type, there appears to be a significant difference between (1) large size, and (2) medium or low size. However, it is important to emphasize that this is just EDA - now we should turn to a formal analysis that we can use to finalize our conclusions.

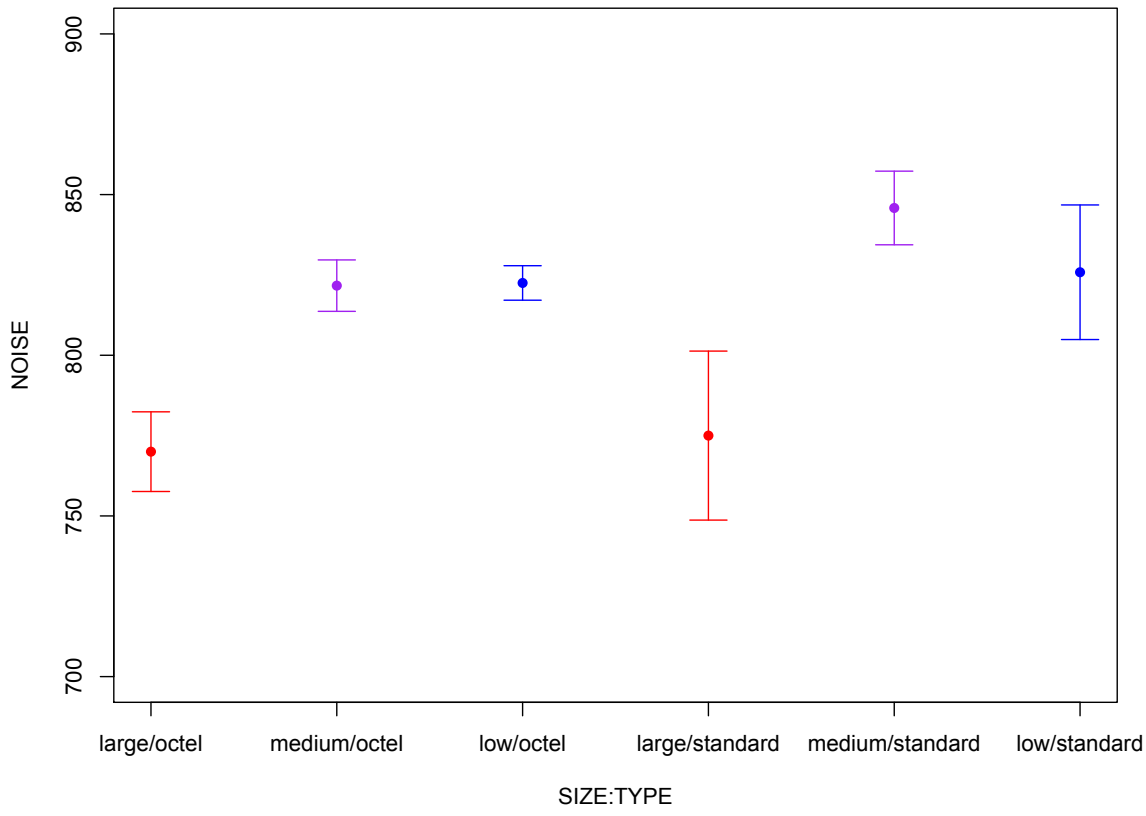


Figure 8.3: Error bar plot for car noise example. Dots denote sample means; bars denote 95% confidence intervals.

8.2 Interpreting the two-way ANOVA results

Now we will perform a two-way ANOVA analysis for the car noise example. Remember that, in Chapter 6 we used the `anova()` function to perform one-way ANOVA. You use this same function to run two-way ANOVA. Here is the code to run a two-way ANOVA analysis, where `NOISE` is the outcome, and the explanatory variables are `SIZE`, `TYPE`, and their interaction:

```
#Run two-way ANOVA
> twoWay.car = aov(NOISE ~ SIZE * TYPE, data = carNoise)
#Examine output
> summary(twoWay.car)
             Df Sum Sq Mean Sq F value   Pr(>F)
SIZE          2  26051   13026 199.119  < 2e-16 ***
TYPE          1   1056    1056  16.146 0.000363 ***
SIZE:TYPE     2    804     402   6.146 0.005792 **
Residuals    30   1963      65
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
        0.1          1
```

Notice that the syntax is extremely similar to the syntax for multiple linear regression that we saw in the previous chapter: By writing `SIZE * TYPE`, we are include the main effects for `SIZE` and `TYPE` as well as their interaction. As we will discuss shortly, you should always start a two-way ANOVA analysis by including the main effects of each explanatory variable *and* their interaction; you will exclude the interaction only if you find that the interaction is statistically insignificant.

Now let's look at the four-row ANOVA table (i.e., the output from `aov()`). This table is structured just like the one-way ANOVA table. The "Sum Sq" column represents the sum of squared deviations (we will elaborate on what each of these four deviations are shortly). Each sum of squares (SS) has a corresponding df (degrees of freedom) which is a measure of the number of independent pieces of information used to compute the corresponding SS (see Section 3.10 for a review of details about degrees of freedom). Meanwhile, each "Mean Sq" is the "Sum Sq" column divided by the "Df" column. Each "Mean Sq" (MS) is a variance estimate or a variance-like quantity, and as such its units are the squares of the outcome units.

Similar to when we first saw ANOVA in Chapter 6, each F-statistic (in the "F value" column) is the ratio of two MS values. For the between-groups ANOVA discussed in this chapter, the denominators are all the MS in the "Residuals" row,

i.e., 65 in this example. This MS (sometimes called "Mean Squared Error," or MSE) corresponds exactly to $\text{MS}_{\text{within}}$ of the one-way ANOVA table. MSE is a "pure" estimate of σ^2, the common group variance, in the sense that it is unaffected by whether or not null hypotheses are true. Just like in one-way ANOVA, each SS is computed using the deviations of group means from an overall mean. For example, the "SIZE" SS is defined by first computing the mean noise for small cars, the mean noise for medium cars, and the mean noise for large cars, and then taking the sum of squared deviations of these three sample means from the overall mean. Thus, the SS for `SIZE` essentially represents the variance of these three sample means; this also explains why the df for `SIZE` is equal to 2, because the df of a variance quantity is always the number of units used to compute that variance minus 1. Relatedly, the sum of all the df in the ANOVA table will be equal to the total number of subjects minus 1; thus, we see that there are 36 subjects in this experiment, because $2 + 1 + 2 + 30 = 35$.

Meanwhile, each F-statistic is compared against its null sampling distribution to compute a p-value. Interpretation of each of the p-values depends on knowing the null hypothesis for each F-statistic, which corresponds to the situation for which the numerator MS has an expected value σ^2. For the main effects (in this example, `SIZE` and `TYPE`), the null hypotheses are that the group means are equal; for example, the null hypothesis for the "SIZE" row is $H_0 : \mu_S = \mu_M = \mu_L$, where μ_S, μ_M, and μ_L are the population means for small, medium, and large cars, respectively. Similarly, the null hypothesis for the "TYPE" row is $H_0 : \mu_{oct} = \mu_{stand}$, where μ_{oct} and μ_{stand} are the population means for octel and standard filters, respectively. Notice that these null hypotheses are the same kind of null hypotheses we saw in one-way ANOVA. Meanwhile, the null hypothesis for the interaction row is that there is no interaction between the two explanatory variables. It is misleading to interpret the main effects rows if a significant interaction is present, as we will discuss next.

The ANOVA table has lines for each main effect, the interaction (if included) and the error. Each of these lines demonstrates MS=SS/df. For the main effects and interaction, there are F values (which equal that line's MS value divided by the error MS value) and corresponding p-values.

In the presence of an interaction, the p-value for the interaction may be the

most important, because then some changes in *both* explanatory variables must have an effect on the outcome, regardless of the main effect p-values. If the p-value for the interaction is less than alpha, then we have a statistically significant interaction, and we have evidence that any non-parallelness seen on a profile plot is "real" rather than due to random error.

A typical example of a statistically significant interaction with statistically non-significant main effects is where we have three levels of factor A and two levels of factor B, and the pattern of effects of changes in factor A is that the means are in a "V" shape for one level of B and an inverted "V" shape for the other level of B. Then the main effect for A is a test of whether at all three levels of A the mean outcome, averaged over both levels of B are equivalent. No matter how "deep" the V's are, if the V and inverted V are the same depth, then the mean outcomes averaged over B for each level of A are the same values, and the main effect of A will be non-significant. But this is usually misleading, because changing levels of A has big effects on the outcome for either level of B, but the effects differ depending on which level of B we are looking at. See Figure 8.4.

If the interaction p-value is statistically significant, then we conclude that the effect on the mean outcome of a change in one factor *depends* on the level of the other factor. More specifically, for at least one pair of levels of one factor the effect of a particular change in levels for the other factor depends on which level of the first pair we are focusing on. However, similar to the issues seen in one-way ANOVA in Chapter 6, a significant interaction p-value does not tell us *which* levels of the two factors are significantly interacting with each other. To do this, we need to do some form of follow-up testing and appropriately account for multiple hypothesis testing issues—we discuss this in Chapter 12.

In our current car noise example, we explain the statistically significant interaction as telling us that the population means for noise differ between standard and Octel filters for at least one car size. Equivalently, we could say that the population means for noise differ among the car sizes for at least one type of filter. A natural question to ask is, "But what particular car sizes and/or filter types are driving (pun intended) the significant interaction?" Looking back at the error bar plot in Figure 8.3 suggests (but does not prove) that the important difference is

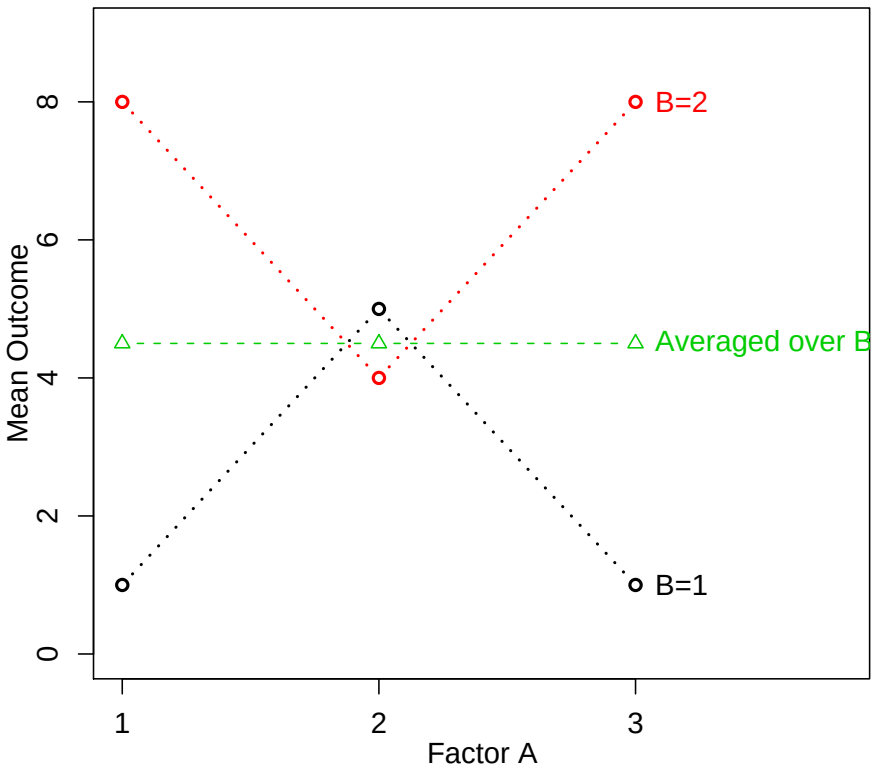


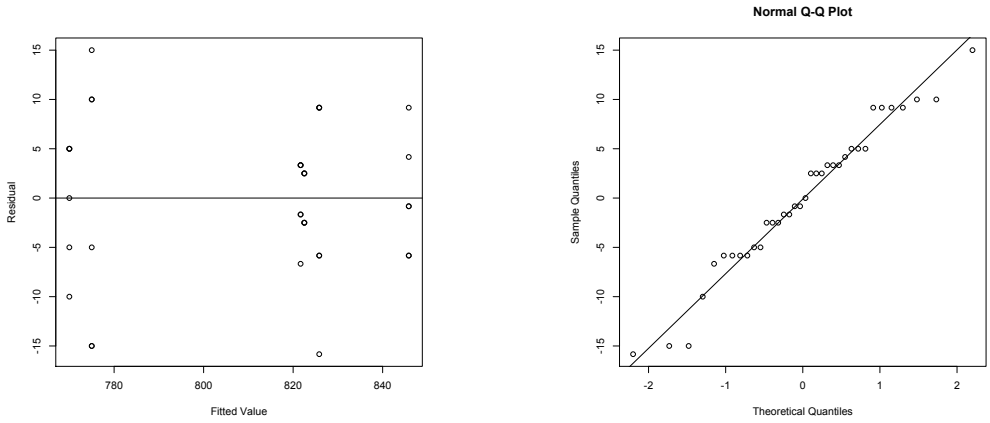
Figure 8.4: Significant interaction with misleading non-significant main effect of factor A.

that the noise level is higher for the standard filter than the Octel filter for medium sized cars, but the filters have equivalent effects for small and large cars.

If the interaction p-value is not statistically significant, then in most situations analysts would re-run the ANOVA without the interaction, i.e., as a main effects only, additive model. The interpretation of the main effects F-statistics is quite similar to what we have already seen in one-way ANOVA in Chapter 6. Each main effect p-value corresponds to the null hypothesis that population means of the outcome are equal for all levels of the factor ignoring the other factor, e.g., for a factor with three levels, the null hypothesis is that $H_0 : \mu_1 = \mu_2 = \mu_3$, and the alternative is that at least one population mean differs from the others. Because the population means for one factor are averaged over the levels of the other factor, unbalanced sample sizes can give misleading p-values. Furthermore, for similar reasons, estimates for main effects may give misleading results if there are significant interaction effects. If there are only two levels, then we can and should immediately report which one is "better" by looking at the sample means. However, if there are more than two levels, then additional analyses in the form of "contrast testing" must be conducted, as shown in Chapter 12, to determine which levels are statistically significantly different.

Inference for two-way ANOVA tables involves first checking the interaction p-value to see if we can reject the null hypothesis that the additive model is sufficient. If that p-value is smaller than α then the adequacy of the additive model can be rejected, and you should conclude that both factors affect the outcome, and that the effect of changes in one factor *depends* on the level of the other factor, i.e., there is an interaction between the explanatory variables. If the interaction p-value is larger than α, then you can conclude that the additive model is adequate, and you should re-run the analysis without an interaction term, and then interpret each of the p-values as in one-way ANOVA, realizing that the effects of changes in one factor are assumed to be the same at every fixed level of the other factor.

Before we complete our analysis, we should note that - similar to one-way ANOVA and linear regression - it is important to do some residual diagnostics that assess if our assumptions for two-way ANOVA are reasonable using residual-versus-fit plots and quantile-normal plots. Here is the code to make these plots



(a) Residual-versus-fit plot. (b) Quantile-normal plot.

Figure 8.5: Residual diagnostics for the car noise example.

for the car noise example (this code follows the exact same format as the linear regression chapters):

```
#residual and fitted values
> twoWay.car.res = residuals(twoWay.car)
> twoWay.car.fit = predict(twoWay.car)
#residual-versus-fit plot
> plot(twoWay.car.fit, twoWay.car.res,
+      xlab = "Fitted Value",
+      ylab = "Residual")
> abline(h = 0)
> #quantile-normal plot
> qqnorm(twoWay.car.res)
> qqline(twoWay.car.res)
```

The resulting residual-versus-fit plot and quantile-normal plot are shown in Figure 8.5. For the residual-versus-fit plot, there are six vertical bands of residual because there are six combinations of filter level and size level, giving six possible predictions. Check the equal variance assumption in the same way as for a regression problem. Meanwhile, verifying that the means for all of the vertical bands are at zero is a check that the mean model is okay. Finally, the quantile-normal plot does not suggest any severe violations of non-Normality.

Residual checking for two-way ANOVA is very similar to regression and one-way ANOVA.

8.3 Math example

In the previous section, we went through a two-way ANOVA analysis where we found that the interaction between two explanatory variables was significant, and thus we did not interpret the main effects and only concluded that there was an interaction between the two variables. In this section, we'll go through another example of two-way ANOVA; but this time, the interaction will be insignificant, and thus this section will demonstrate how main effects should be interpreted in that case.

The data in mathActivity.txt are from an observational study carried out to investigate the relationship between the ACT Math Usage Test and the explanatory variables level of mathematics coursework taken (1=algebra only, 2=algebra+geometry, 3=through calculus) and whether or not someone is in extracurricular activities (1 = no, 2 = yes) for 861 high school seniors. The outcome, ACT score, ranges from 0 to 36 with a median of 15 and a mean of 15.33. (As a sidenote: This dataset and the dataset from the previous section can be called a 3x2 ("three by two") ANOVA because both datasets consist of one categorical variable with 3 levels and another with 2 levels.) Because this is an observational study (i.e., coursework and activity participation were not randomized), the results presented here are obviously not of any kind of causal nature and are merely associations. Furthermore, this example obviously takes a binary view of gender and is limited as such. Nonetheless, this example will demonstrate how to perfom two-way ANOVA on a dataset using factors commonly seen in education datasets.

The rows of the data table (experimental units) are individual students. There is some concern about independent errors if the 861 students come from just a few schools, with many students per school, because then the errors for students from the same school are likely to be correlated. In that case, the p-values and confidence intervals will be unreliable, and we should use an alternative analysis such as mixed models (Chapter 14), which accounts for the clustering into schools. For the analysis below, we assume that students are randomly sampled throughout

the country so that including two students from the same school would only be a rare coincidence.

First, note that the *categorical* variables are nonetheless coded *numerically*. Thus, analyses in `R` will be very incorrect if we don't first convert these variables to categorical variables using the `factor()` function. To demonstrate this, let's see what happens if we try to run two-way ANOVA as soon as we read in the dataset:

```
# INCORRECT two-way ANOVA output
> twoWay.math = aov(score ~ courses*activity, data = mathAct)
> summary(twoWay.math)
                   Df Sum Sq Mean Sq F value   Pr(>F)
courses             1  14057   14057 540.028  < 2e-16 ***
activity            1    685     685  26.333 3.56e-07 ***
courses:activity    1      0       0   0.002    0.964
Residuals         857  22307      26
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1          1
```

We know that the `courses` variable has 3 levels, so its df should be 2; however, we see that the df is listed as 1 in the above table, so we should already know something is wrong. Furthermore, there is something off about the interaction row - the sum of squares (SS) is listed as 0. We won't go into the details as to what's happening here; the point is that `R` will report incorrect results if you don't first take the care to make sure that all categorical variables are indeed classified as such.

The below code converts the `courses` and `activity` variables to categorical variables using the `factor()` function, as we've seen before. Note, however, that the variables will still be labeled with the not-very-meaningful (1, 2, 3) and (1, 2). (Sure, you can try to remember that 1=algebra only, 2=algebra+geometry, 3=through calculus for `courses` and 1 = no, 2 = yes for `activity`, but isn't that annoying?) The below code gives these variables more sensible labels using the `levels()` function, which we haven't seen before:

```
#Note that we have the following for each variable:
# courses: 1 = alg only, 2 = alg + geom, 3 = calc
# activity: 1 = no, 2 = yes
>
#convert both variables to factors
> mathAct$courses = factor(mathAct$courses)
> mathAct$activity = factor(mathAct$activity)
#add more meaningful labels to the factors
```

```
> levels(mathAct$courses) = c("alg", "algGeom", "calc")
> levels(mathAct$activity) = c("no", "yes")
```

Now we can perform two-way ANOVA correctly using the previous code on this appropriately-converted dataset. But first, let's do some EDA. The below code uses the `table()` function to produce a contingency table for the two categorical variables:

```
> table(mathAct$courses, mathAct$activity)

           no yes
  alg      82  48
  algGeom 387 223
  calc     54  67
```

Note that the contingency table is much easier to read because we took the time to relabel the variables. Unlike the car noise example from the previous section, this is not a balanced ANOVA, because it has unequal cell sizes. Indeed, this is an observational study where these two explanatory variables were outside our control - it would be very unusual if this were a balanced ANOVA.

Now let's make a profile plot. Unlike the previous section, we'll show the code for making a profile plot - it's a bit tedious and for the sake of 36-309 you are not expected to code this yourself, but we'll show it here for demonstration. The resulting profile plot is in Figure 8.6.

```
#compute sample means for six groups
> sampleMeans = aggregate(score ~ courses*activity, data = mathAct
    , FUN = mean)
#subset by activitiy (the binary variable)
> sampleMeans.no = subset(sampleMeans, activity == "no")
> sampleMeans.yes = subset(sampleMeans, activity == "yes")
#profile plot
> plot(1:3, sampleMeans.no$score, type = "l", ylim = c(5,30),
+      xlab = "Courses", ylab = "Score",
+      xaxt = "n")
> lines(1:3, sampleMeans.yes$score, col = "red", lty = 2)
#add points
> points(1:3, sampleMeans.no$score, pch = 16)
> points(1:3, sampleMeans.yes$score, pch = 17, col = "red")
#add x-axis labels
> axis(side = 1, at = 1:3, labels = sampleMeans.no$courses)
#add legend
> legend("bottomright",
```

```
+      legend = c("Didn't Participate in Activities", "Participated
    in Activities"),
+      col = c("black", "red"), lty = c(1, 2))
```

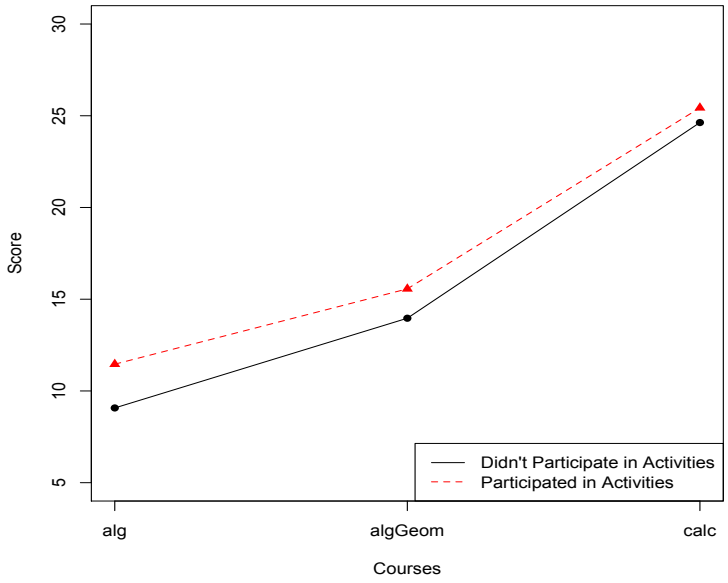


Figure 8.6: Six sample means for the math example.

The first impression is that students who take more courses have higher scores, students who participate in activities have slightly higher scores than those who do not, and perhaps the activity difference is smaller for students who take more courses.

Now let's perform two-way ANOVA on this dataset. We'll include `activity`, `courses`, and their interaction. We'll use the `aov()` code again, but now our variables are correctly classified as categorical variables, so we will be conducting the correct analysis!

```
#After the variables are coded as categorical,
#this is the correct two-way ANOVA:
> twoWay.math = aov(score ~ courses*activity, data = mathAct)
> summary(twoWay.math)
```

```
                   Df  Sum Sq Mean Sq F value   Pr(>F)
courses             2   15619    7809 319.826  < 2e-16 ***
activity            1     517     517  21.158 4.87e-06 ***
courses:activity    2      38      19   0.771    0.463
Residuals         855   20877      24
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1          1
```

Note that the interaction line of the table (courses:activity) has 2 df because the difference between an additive model (with a parallel pattern of population means) and an interaction model (with arbitrary patterns) can be thought of as taking the parallel pattern, then moving any two points for any one of the activity levels. More generally, the interaction df is $(k-1)(m-1)$ for any k by m ANOVA. In this example, $k = 3$ and $m = 2$, so $(k-1)(m-1) = 2 \times 1 = 2$.

The main point of this ANOVA table is that the interaction between the explanatory variables is not significant (p=0.463), so we have no evidence to reject the additive model, and we conclude that course effects on the outcome are the same for both activity levels, and activity effects on the outcome are the same for all three levels of coursework. Therefore it is appropriate to re-run the ANOVA with a different means model, i.e., with an additive rather than an interactive model:

```
#two-way ANOVA without an interaction:
> twoWay.math = aov(score ~ courses + activity, data = mathAct)
> summary(twoWay.math)
             Df  Sum Sq Mean Sq F value   Pr(>F)
courses       2   15619    7809  320.00  < 2e-16 ***
activity      1     517     517   21.17 4.84e-06 ***
Residuals   857   20914      24
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1          1
```

Our conclusion, using a significance level of $\alpha = 0.05$ is that both courses and activities affect test score. Specifically, because `activity` has only two levels (1 df), we can directly check the profile plot to see that students who participate in activities have a higher mean. Then we can conclude based on the small p-value that participating in activities is associated with a higher math ACT score compared to not participating, for each level of courses. This is not in conflict with the observation that some students who don't participate in activities have better scores than some students who do participate, because it is only a statement about

Case	light	sound	interaction
A	<0.0005	0.971	0.802
B	0.787	0.380	0.718
C	<0.0005	<0.0005	<0.0005
D	<0.0005	<0.0005	0.995
E	0.506	<0.0005	0.250

Table 8.1: P-values for various light/sound experiment cases.

means. This result also does not suggest that participating in activities *causes* a student to achieve higher test scores; again, what we have found here is only an associative effect, not a causal effect.

Looking at the p-value for courses, we see that at least one level of courses differs from the other two, and this is true separately for both levels of `activity`, because the additive model is an adequate model. But we cannot make further important statements about *which* levels of courses are significantly different without additional analyses, which are discussed in Chapter 12.

We can also note that the residual (within-group) variance is 24, so our estimate of the population standard deviation for each group is $\sqrt{24} = 4.9$. Therefore, about 95% of test scores for any level of coursework and activity participation are within 9.8 points of that group's mean score.

8.4 More on profile plots, main effects and interactions

Consider an experiment looking at the effects of different levels of light and sound on some outcome. Five possible outcomes are shown in the profile plots of Figures 8.7, 8.8, 8.9, 8.10, and 8.11 which include plus or minus 2 SE error bars (roughly 95% CI for the population means). We will briefly go through each of these five cases, which will demonstrate how to interpret results from different two-way ANOVAs.

Table 8.1 shows the p-values from two-way ANOVA's of these five cases.

In Case A we can see that the lines are roughly parallel, so an additive model

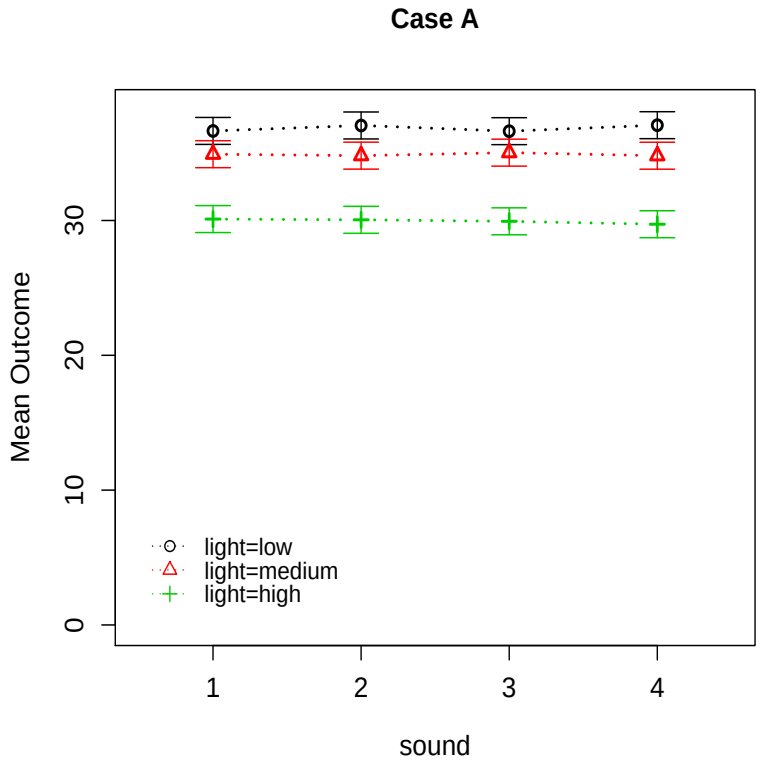


Figure 8.7: Case A for light/sound experiment.

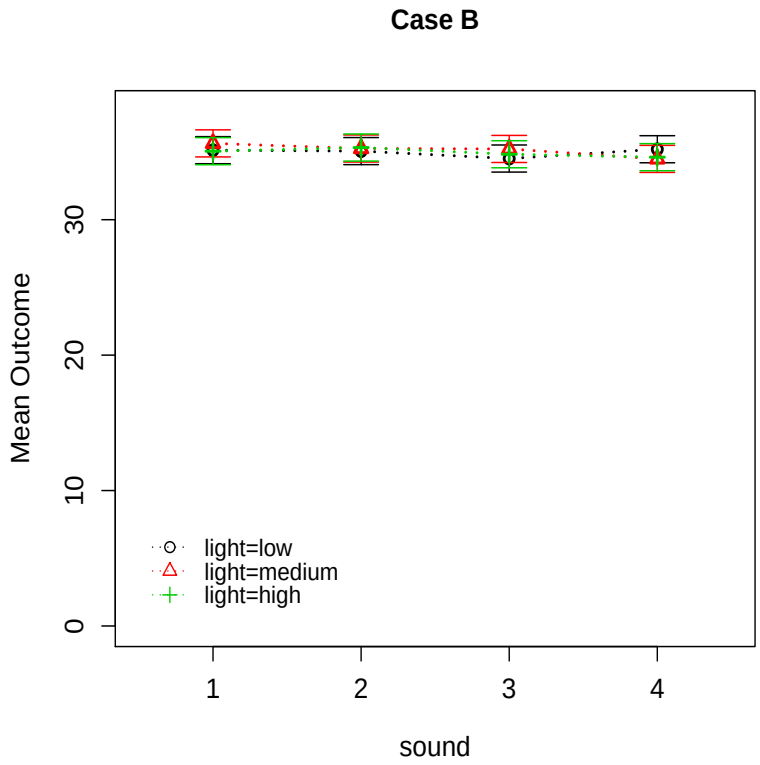


Figure 8.8: Case B for light/sound experiment.

should be OK, and indeed the interaction p-value is 0.802. We should re-fit a model without an interaction term. We can also see that, for any particular level of light (i.e., for any one of the colored lines in the figure), as we change sound levels (move left or right) the mean outcome (y-axis value) does not change much, indicating that sound level does not significantly affect the outcome; and indeed, we get a non-significant p-value for this main effect (0.971). But changing light levels (moving from one colored line to another, at any sound level) does change the mean outcome, e.g., high light gives a low outcome, so we expect a significant p-value for light, and indeed it is <0.0005.

In case B, as in case A, the lines are nearly parallel, suggesting that an additive, no-interaction model is adequate, and we should re-fit a model without an interaction term. We also see that changing sound levels (moving left or right on the plot) has no effect on the outcome (vertical position), so sound is not a significant

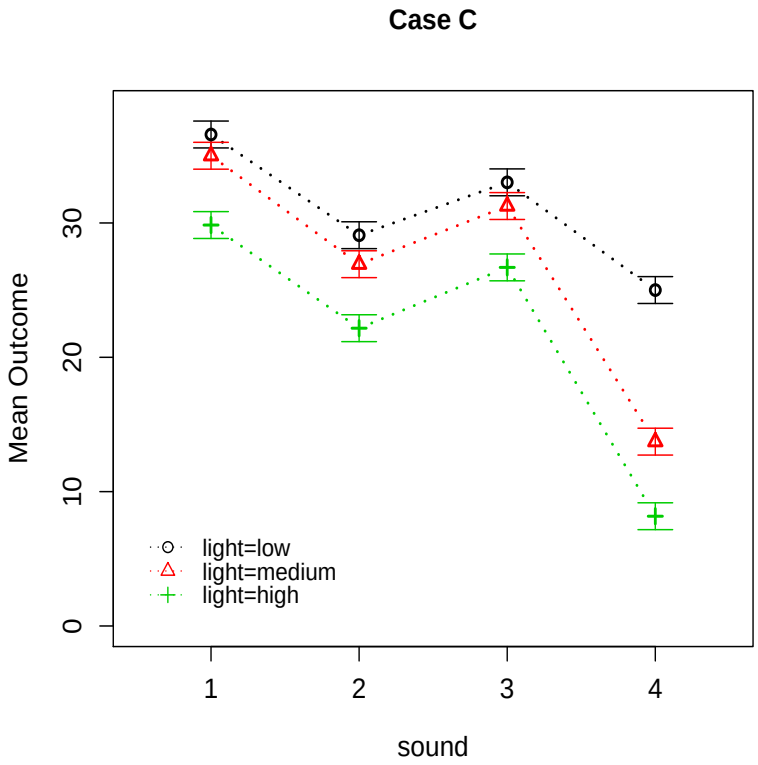


Figure 8.9: Case C for light/sound experiment.

explanatory variable. Also changing light level (moving between the colored lines) has no effect. So all the p-values are non-significant (>0.05).

In case C, there is a single cell, low light with sound at level 4, that must be moved much more than the size of the error bars to make the lines parallel. This is enough to give a significant interaction p-value (<0.0005), and require that we stay with this model that includes an interaction term, rather than using an additive model. The p-values for the main effects (though significant) now have no real interest. However, we get significant main effects p-values because, for any light level, it looks like increasing sound tends to decrease the outcome; likewise, for any sound level, increasing light tends to decrease the outcome. However, with the significant interaction, we know that these marginal effects actually depend on the level of the other factor. For example, although we need contrast testing to be sure, it is quite obvious that changing from low to high light level for any sound

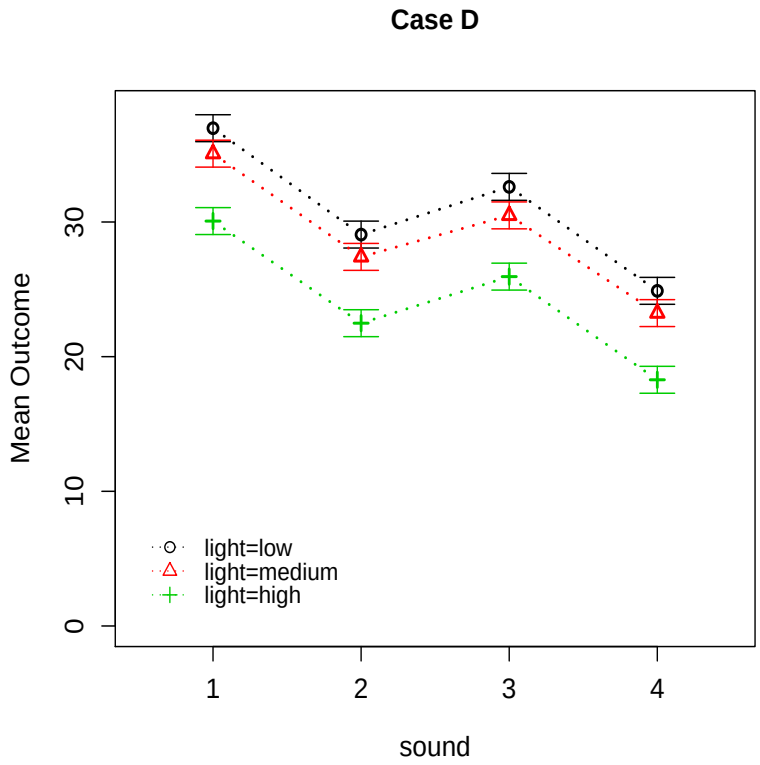


Figure 8.10: Case D for light/sound experiment.

level lowers the outcome, and changing from sound level 3 to 4 for any light level lowers the outcome.

Case D shows no interaction (p=0.995) because on the scale of the error bars, the lines are parallel. Both main effects are significant because (similar to Case C) we can clearly see that marginally increasing the light or sound tends to decrease the outcome.

Case E shows no interaction. The light factor is not statistically significant as shown by the fact that for any sound level, changing light level (moving between colored lines) tends to not change the outcome. However, the sound factor is found to be statistically significant, because, for any light level, changing the sound levels tends to affect the outcome (overall, there appears to be a negative trend).

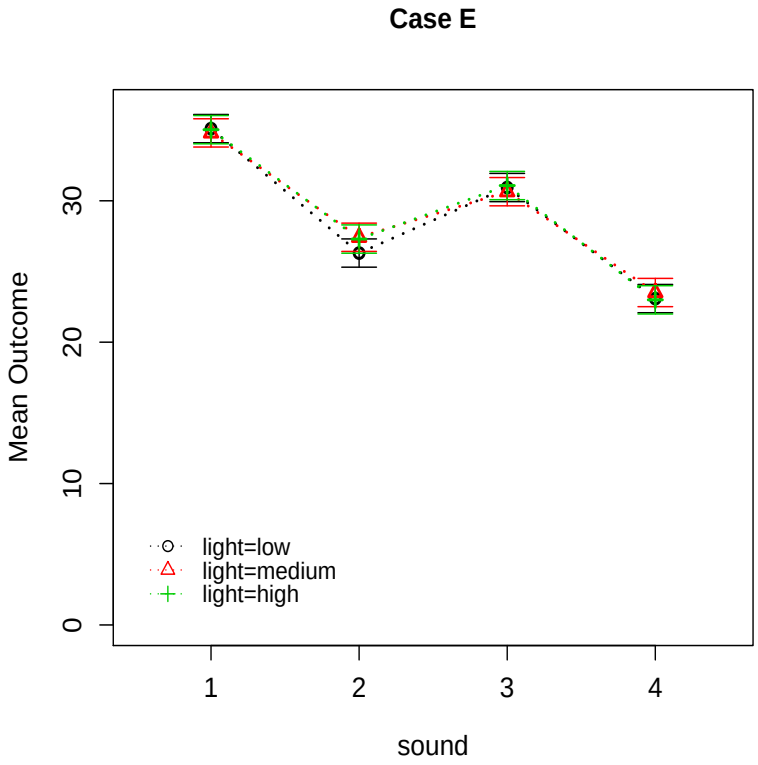


Figure 8.11: Case E for light/sound experiment.

Taking error into account, in most cases you can get a good idea which p-values will be significant just by looking at a profile plot.

8.5 Do it in R

Throughout this chapter we have given example code to conduct two-way ANOVA analyses in `R`, but we will provide a quick review here.

You use the `aov()` to run a two-way ANOVA analysis, and you use the `summary()` function to look at the output from that analysis. These are the same functions we used to run one-way ANOVA in Chapter 6. Meanwhile, we use very similar "formula syntax" (e.g., we write things like `y ~ a*b`), just like we did when using the `lm()` function for multiple linear regression in Chapter 10. In particular, when you first implement two-way ANOVA, you should always include main effects as well as interactions using the "*" symbol (e.g., you'll first run something like `aov(y ~ a*b)`). If you find that the interaction is significant, then you conclude that there is a significant interaction but do not interpret the main effects (as it would be misleading to do so). If you find that the interaction is not significant, then you run the two-way ANOVA analysis again but without the interaction (e.g., you'll run something like `aov(y ~ a+b)`); then, you can proceed to interpret the main effects.

For two-way ANOVA, it's still important to assess assumptions with residual diagnostics. The `residuals()` function is used to obtain residuals from a two-way ANOVA model defined with `aov()`, and the `predict()` function is used to obtain fitted values from a two-way ANOVA model. Then, a residual-versus-fit plot can be made. Furthermore, a quantile-normal plot of the residuals can be made using `qqnorm()` and `qqline()`.

Lastly, because two-way ANOVA involves two categorical variables, it is especially important that you make sure that your variables are classified in `R` as categorical variables before you proceed to the analysis stage. You can use the `class()` function to check the class of your individual variables. If you find that one of your explanatory variables is classified as numeric, you can redefine it as a factor (i.e., categorical variable) using the `factor()` function. Furthermore, you can use the `levels()` function to relabel the levels of your categorical variable

with more meaningful/interpretable names (which is often helpful).

Chapter 9

Simple Linear Regression

An analysis appropriate for a quantitative outcome and a single quantitative explanatory variable.

So far we have discussed how to statistically compare outcome variables among treatment groups (in particular, ANOVA, where t-tests are a special case of ANOVA when there are only two treatment groups). ANOVA only uses two types of variables: The outcome variable and the treatment variable. It does not use any other explanatory variables. This can be problematic, because, as we discussed in Chapter 7, if treatment groups differ in terms of other explanatory variables, then an experiment is said to have low internal validity, i.e., causal conclusions cannot readily be made from analyses like ANOVA. This is why it is important to use design techniques like block randomization and blinding to ensure that treatment groups are similar to each other in terms of explanatory variables.

However, it's often impossible to ensure that treatment groups look exactly the same. For example, let's say we are conducting a treatment-versus-control experiment where we have a quantitative, continuous measure of income for each subject. Ideally, the treatment and control groups have exactly the same levels of income, but that will probably never be the case (since basically no two people have exactly the same income levels). Of course, by randomizing subjects to treatment and control, income levels should look the same on average. And better yet, we can divide subjects into blocks like "low income," "medium income," and "high income" to ensure that the treatment and control groups look similar to each other in terms of this categorized version of income. In any case, there will always

be *some* differences between the treatment groups in terms of this quantitative, continuous explanatory variable. Does this mean we are doomed to accept that there is nothing we can do about these inevitable differences?

Luckily, no. There are statistical models beyond ANOVA that "adjust" for differences in explanatory variables (in particular, continuous ones). We will elaborate on what we mean by "adjust" over the next few chapters, but it more or less means that these models account for the relationship between the outcome variable and other explanatory variables when estimating treatment effects. Linear regression is by far the most popular method for "adjusting" for other explanatory variables—it is widely used in virtually every field of (social) science. In this chapter, we discuss how linear regression models the relationship between the outcome and a single quantitative explanatory variable. Then, in Chapter 10, we discuss how linear regression can be used to adjust for multiple explanatory variables and estimate treatment effects in experiments.

9.1 The model behind linear regression

When we are examining the relationship between a quantitative outcome and a single quantitative explanatory variable, simple linear regression is the most commonly considered analysis method. (The "simple" part tells us we are only considering a single explanatory variable.) In linear regression, we usually have many different values of the explanatory variable, and we usually assume that values between the observed values of the explanatory variable are also possible values (in other words, the explanatory variable is continuous). Linear regression, unsurprisingly, postulates a linear relationship between the population mean of the outcome and the value of the explanatory variable. If we let Y be some outcome, and x be some explanatory variable, then we can express the structural model using the equation

$$E(Y|x) = \beta_0 + \beta_1 x$$

where E(), which is read "expected value of", indicates a population mean; $Y|x$, which is read "Y given x", indicates that we are looking at the possible values of Y when x is restricted to some single value; β_0, read "beta zero", is the intercept parameter; and β_1, read "beta one," is the slope parameter. A common term for any parameter or parameter estimate used in an equation for predicting Y from x is **coefficient**. Often the "1" subscript in β_1 is replaced by the name of the explanatory variable or some abbreviation of it.

When we did ANOVA, the structural model simply stated that the outcome means may differ among the treatment groups, leading to parameters $\mu_1, \ldots, \mu_k$. However, now we have a continuous explanatory variable, and we are assuming that the outcome means may differ among values of the explanatory variable. In particular, the structural model for linear regression says that the outcome mean for all subjects who have a particular value "x" for their explanatory variable can be calculated using the simple linear expression $\beta_0 + \beta_1 x$. (In reality, we never know what β_0 and β_1 are—these are parameters, i.e., "secrets of nature"—similar to how we can never know the true treatment group means $\mu_1, \ldots, \mu_k$ in ANOVA. However, we can make estimates of these parameters and substitute the estimates into the $\beta_0 + \beta_1 x$ equation. We will discuss how to estimate β_0 and β_1 soon.)

In real life we know that although the equation makes a prediction of the true mean of the outcome for any fixed value of the explanatory variable, it would be unwise to use **extrapolation** to make predictions *outside* of the range of x values that we have available for study. On the other hand it *is* reasonable to **interpolate**, i.e., to make predictions for unobserved x values in between the observed x values. The structural model is essentially the assumption of "linearity", at least within the range of the observed explanatory data.

It is important to realize that the "linear" in "linear regression" does *not* imply that only linear relationships can be studied. For example, instead of placing x in the linear regression, we could just have easily placed x^2, thereby exploring a quadratic relationship. In general, any transformation of x can be placed in the linear regression, meaning that any *nonlinear* relationship can actually be explored using linear regression (which sounds counter-intuitive!)

The structural model (i.e., the form of the mean outcome) underlying a linear regression analysis is that the explanatory and outcome variables are linearly related such that the population mean of the outcome for any x value is $\beta_0 + \beta_1 x$.

The error model that we use is that for each particular x, if we have or could collect many subjects with that x value, their distribution around the population mean is Gaussian with a spread, say σ^2, that is the same value for each value of x (and corresponding population mean of y). Put it mathematical terms, this means

that simple linear regression posits the following statistical model:

$$Y \overset{iid}{\sim} N(\beta_0 + \beta_1 x, \sigma^2) \tag{9.1}$$

This is exactly the same as ANOVA, but now the mean is denoted by $\beta_0 + \beta_1 x$ instead of $\mu_1, \ldots, \mu_k$. Furthermore, this model is making the same assumptions as ANOVA: Normality, equal variance, and independent errors. (Also similar to ANOVA, the value of σ^2 is an unknown parameter, and we can estimate it from the data.) However, in addition to these three assumptions, there is also the assumption of "fixed-x" in simple linear regression. The "fixed-x" assumption is that the explanatory variable is measured without error. Sometimes this is possible, e.g., if it is a count, such as the number of legs on an insect, and the fixed-x assumption is often quite reasonable in social science datasets containing information like demographic data. However, sometimes there is some error in the measurement of the explanatory variable. In practice, we need to be sure that the size of the error in measuring x is small compared to the variability of Y at any given x value. For more on this topic, see the section on robustness, below.

The error model underlying a linear regression analysis includes the assumptions of fixed-x, Normality, equal spread, and independent errors.

In addition to the three error model assumptions just discussed, we also assume "independent errors". This assumption comes down to the idea that the **error** (deviation of the true outcome value from the population mean of the outcome for a given x value) for one observational unit is not predictable from knowledge of the error for another observational unit. For example, in predicting time to complete a task from the dose of a drug suspected to affect that time, knowing that the first subject took 3 seconds longer than the mean of all possible subjects with the same dose should not tell us anything about how far the next subject's time should be above or below the mean for their dose. This assumption can be trivially violated if we happen to have a set of identical twins in the study, in which case it seems likely that if one twin has an outcome that is below the mean for their assigned dose, then the other twin will also have an outcome that is below the mean for their assigned dose (whether the doses are the same or different).

A more interesting cause of correlated errors is when subjects are trained in groups, and the different trainers have important individual differences that affect

the trainees' performance. Then knowing that a particular subject does better than average gives us reason to believe that most of the other subjects in the same group will probably perform better than average because the trainer was probably better than average.

Another important example of non-independent errors is **serial correlation** in which the errors of adjacent observations are similar. This includes adjacency in both time and space. For example, if we are studying the effects of fertilizer on plant growth, then similar soil, water, and lighting conditions would tend to make the errors of adjacent plants more similar. In many task-oriented experiments, if we allow each subject to observe the previous subject perform the task which is measured as the outcome, this is likely to induce serial correlation. And worst of all, if you use the same subject for every observation, just changing the explanatory variable each time, serial correlation is extremely likely. Breaking the assumption of independent errors does not indicate that no analysis is possible, only that linear regression is an inappropriate analysis. Other methods such as time series methods or mixed models are appropriate when errors are correlated; we will discuss these methods in Chapter 14.

The worst case of breaking the independent errors assumption in regression is when the observations are repeated measurements on the same experimental unit.

Before going into the details of linear regression, it is worth thinking about the variable types for the explanatory and outcome variables and the relationship of ANOVA to linear regression. For both ANOVA and linear regression we assume a Normal distribution of the outcome for each value of the explanatory variable. Implicitly this indicates that the outcome should be a continuous quantitative variable, because the Normal distribution is a continuous, quantitative distribution. As we've discussed previously, practically speaking, real measurements are usually rounded and therefore some of their continuous nature is not available to us (for example, age is often recorded in years with no decimal places, but this does not mean that it is not a continuous variable), meaning that usually we can only hope that measurements are approximately Normally distributed rather than literally Normally distributed. Fortunately, regression and ANOVA are both quite robust to deviations from the Normality assumption, and it is okay to use discrete or continuous outcomes that have at least a moderate number of different values, e.g.,

10 or more. In some circumstances, it can even be reasonable to use regression or ANOVA when the outcome is ordinal with a fairly small number of levels.

The explanatory variable in ANOVA is categorical and nominal. Imagine we are studying the effects of a drug on some outcome and we first do an experiment comparing control (no drug) vs. drug (at a particular concentration). Regression and ANOVA would give equivalent conclusions about the effect of drug on the outcome, but regression seems inappropriate. Two related reasons are that there is no way to check the appropriateness of the linearity assumption, and that after a regression analysis it is appropriate to interpolate between the x (dose) values, and that is inappropriate here.

Now consider another experiment with 0, 50 and 100 mg of drug. ANOVA and regression will give different answers, because ANOVA makes no assumptions about the relationships of the three population means, but regression assumes a linear relationship. If there truly is a linear relationship between dosage and the outcome, the regression will have a bit more power than ANOVA. If the truth is non-linearity, regression will make inappropriate predictions, but at least regression will have a chance to detect the non-linearity. ANOVA also loses some power because it incorrectly treats the doses as nominal when they are at least ordinal. As the number of doses increases, it is more and more appropriate to use regression instead of ANOVA, and we will be able to better detect any non-linearity and correct for it, e.g., with a data transformation.

Figure 9.1 shows a way to think about and remember most of the regression model assumptions. The four little Normal curves represent the Normally distributed outcomes (Y values) at each of four fixed x values. The fact that the four Normal curves have the same spreads represents the equal variance assumption. And the fact that the four means of the Normal curves fall along a straight line represents the linearity assumption. Only the fifth assumption of independent errors is not shown on this mnemonic plot.

9.2 Statistical hypotheses

It is often of interest to assess whether there is some kind of true linear relationship between the outcome and the explanatory variable, and this is where hypothesis testing can be useful. For simple linear regression, the chief null hypothesis is $H_0 : \beta_1 = 0$, and the corresponding alternative hypothesis is $H_1 : \beta_1 \neq 0$. If this

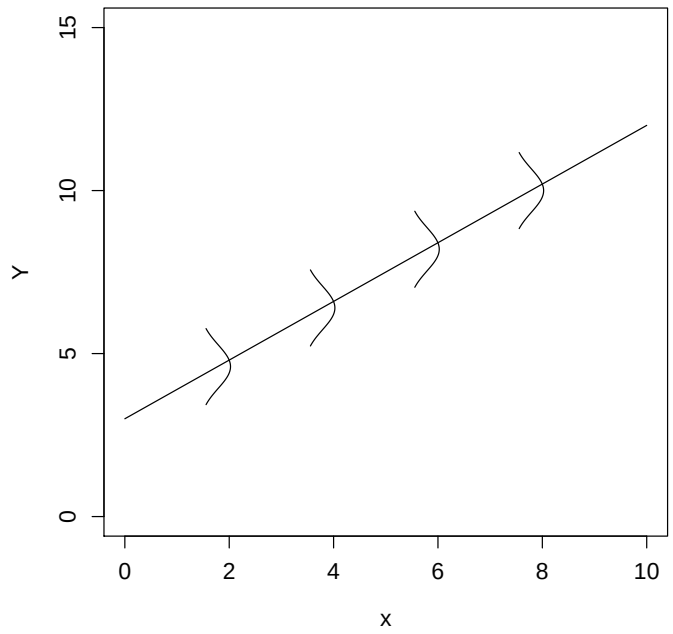


Figure 9.1: Mnemonic for the simple regression model.

null hypothesis is true, then, from $E(Y) = \beta_0 + \beta_1 x$ we can see that the population mean of Y is β_0 for *every* x value, which tells us that x has no linear association with Y. The alternative is that changes in x are associated with changes in Y (or changes in x cause changes in Y in a randomized experiment).

Sometimes it is reasonable to choose a different null hypothesis for β_1. For example, if x is some **gold standard** for a particular measurement, i.e., a best-quality measurement often involving great expense, and y is some cheaper substitute, then the obvious null hypothesis is $\beta_1 = 1$ with alternative $\beta_1 \neq 1$. For example, if x is percent body fat measured using the cumbersome whole body immersion method, and Y is percent body fat measured using a formula based on a couple of skin fold thickness measurements, then we expect either a slope of 1, indicating equivalence of measurements (on average) or we expect a different slope indicating that the skin fold method proportionally over- or under-estimates body fat.

Sometimes it also makes sense to construct a null hypothesis for β_0, usually $H_0 : \beta_0 = 0$. This should only be done if each of the following is true:

1. There are data that span $x = 0$, or at least there are data points near $x = 0$.

2. The statement "the population mean of Y equals zero when $x = 0$" both makes scientific sense and the difference between equaling zero and not equaling zero is scientifically interesting.

The first point again emphasizes that it is usually not appropriate to extrapolate results from linear regression. As an example of a violation of the second point, let us say that we use linear regression to assess the relationship between people's weight (the outcome) and height (the explanatory variable). It is impossible for someone to have literally 0 height, and so the intercept from such a linear regression would not be of scientific interest. We will discuss these points further in the interpretation section (Section 9.5).

The usual regression null hypothesis is $H_0 : \beta_1 = 0$. Sometimes it is also meaningful to test $H_0 : \beta_0 = 0$ or $H_0 : \beta_1 = 1$.

9.3 Simple linear regression example

As a (simulated) example, consider an experiment in which corn plants are grown in pots of soil for 30 days after the addition of different amounts of nitrogen fertilizer. The data are in corn.dat, which is a space delimited text file with column headers. Corn plant final weight is in grams, and amount of nitrogen added per pot is in mg.

EDA, in the form of a scatterplot, is shown in Figure 9.2. Here is the `R` code used to make the scatterplot:

```
plot(corn$nitrogen, corn$weight,
  xlab="Soil Nitrogen (mg/pot)",
    ylab="Final Weight (gm)",
    cex.axis=1.2, cex.lab=1.2)
```

In general, to make a scatterplot, you write `plot(data$x, data$y)`, where `data` is the name a dataset, `x` is the x-axis variable you want to plot, and `y` is the y-axis variable you want to plot. The other arguments within the `plot()` function above make the scatterplot more readable, which you should always do when making graphs that will be seen by someone other than you!

We want to use EDA to check that the assumptions are reasonable before trying a regression analysis. We can see that the assumptions of linearity seems plausible because we can imagine a straight line from bottom left to top right going through the center of the points. Also the assumption of equal spread is plausible because for any narrow range of nitrogen values (horizontally), the spread of weight values (vertically) is fairly similar. These assumptions should only be doubted at this stage if they are drastically broken. The assumption of Normality is not something that human beings can test by looking at a scatterplot. But if we noticed, for instance, that there were only two possible outcomes in the whole experiment, we could reject the idea that the distribution of weights is Normal at each nitrogen level.

The assumption of fixed-x cannot be seen in the data. Usually we just think about the way the explanatory variable is measured and judge whether or not it is measured precisely (with small spread). Here, it is not too hard to measure the amount of nitrogen fertilizer added to each pot, so we accept the assumption of fixed-x. In some cases, we can actually perform repeated measurements of x on the same case to see the spread of x and then do the same thing for y at each of a few values, then reject the fixed-x assumption if the ratio of x to y variance is

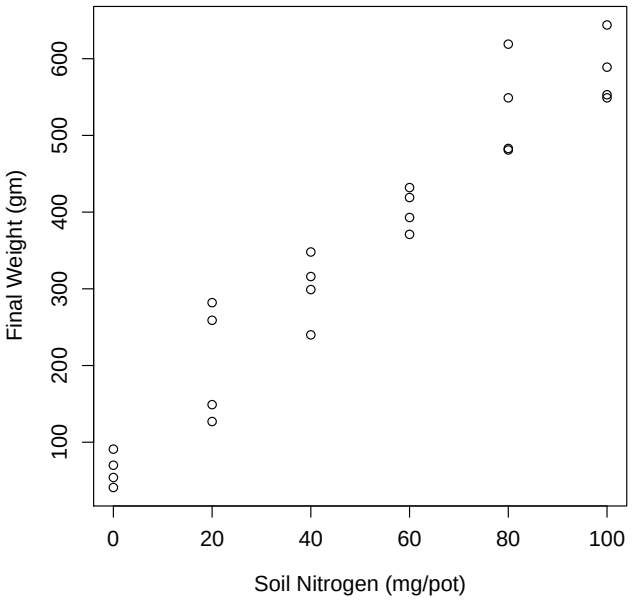


Figure 9.2: Scatterplot of corn data.

larger than, e.g., around 0.1.

The assumption of independent error is usually not visible in the data and must be judged by the way the experiment was run. But if serial correlation is suspected, there are tests such as the Durbin-Watson test that can be used to detect such correlation.

Once we make an initial judgement as to whether linear regression is a sensible thing to do for our data, based on plausibility of the model after examining our EDA, we perform the linear regression analysis, then further verify the model assumptions with residual checking.

9.4 Regression calculations

The basic regression analysis uses fairly simple formulas to get estimates of the parameters β_0, β_1, and σ^2. We denote these estimates by $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, respectively. (For example, $\hat{\beta}_0$ is pronounced as "beta 0 hat". In general, estimates of parameters are denoted with "hats".) These estimates can be derived from either

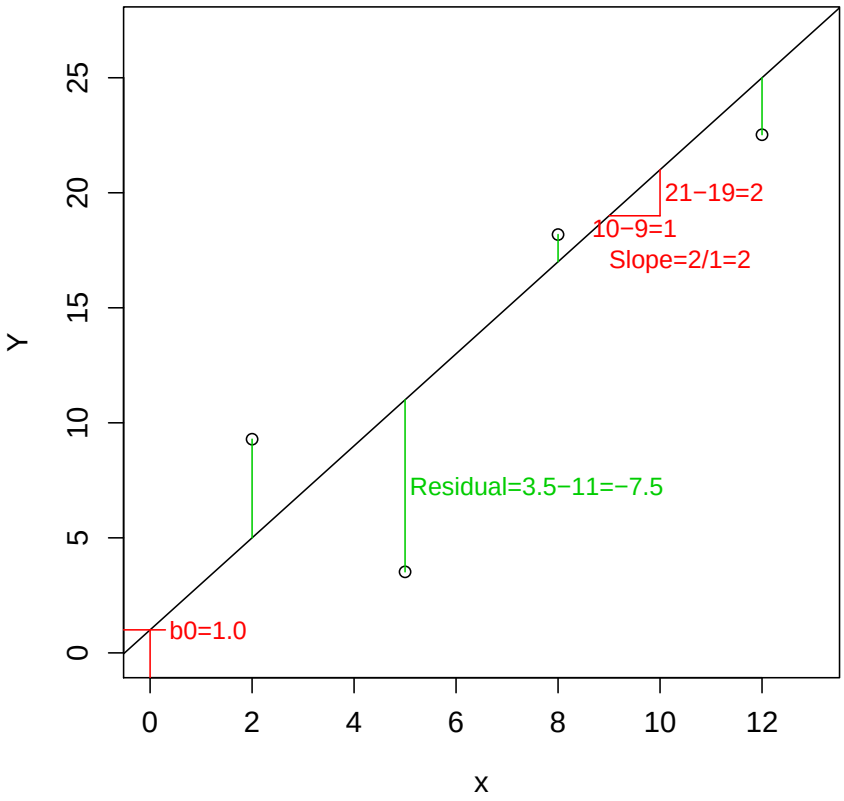


Figure 9.3: Least square principle.

of two basic approaches which lead to identical results. We will not discuss the more complicated maximum likelihood approach here. The least squares approach is fairly straightforward. It says that we should choose as the best-fit line the line which minimizes the sum of the squared residuals, where the **residuals** are the vertical distances from individual points to the best-fit "regression" line.

The principle is shown in Figure 9.3. The plot shows a simple example with four data points. The diagonal line shown in black is close to, but not equal to the "best-fit" line. (Bells should be ringing here. This is the same kind of visual that we used in Chapter 6 for within and between variation, i.e., Figures 6.2 and 6.3. In Chapter 6, we discussed variation from sample means. Here, we are discussing variation from regression means. There is an equivalence here that will be discussed in Chapter 10.)

Any line can be characterized by its intercept and slope. The intercept is the y value when x equals zero, which is 1.0 in the example. *Be sure to look carefully at the x-axis scale; if it does not start at zero, you might read off the intercept incorrectly.* The slope is the change in y for a one-unit change in x. Because the line is straight, you can read this off anywhere. Also, an equivalent definition is the change in y divided by the change in x for *any* segment of the line. In the figure, a segment of the line is marked with a small right triangle. The vertical change is 2 units and the horizontal change is 1 unit, therefore the slope is 2/1=2. Using $\hat{\beta}_0$ for the intercept and $\hat{\beta}_1$ for the slope, the equation of the line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

By plugging different values for x into this equation, we can find the corresponding y values that are on the line drawn. For any given $\hat{\beta}_0$ and $\hat{\beta}_1$ we get a potential best-fit line, and the vertical distances of the points from the line are called the **residuals**. We can use the symbol $\hat{y}_i$, pronounced "y i hat", to indicate the fitted or predicted value of outcome y for subject i. For subject i, who has explanatory variable x_i, the prediction is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and the residual is $y_i - \hat{y}_i$. The least square principle says that the best-fit line is the one with the smallest sum of squared residuals. As is the case for the residuals of any statistical model, the sum of the residuals (not squared) is zero for the least-squares best-fit line.

In practice, we don't really try every possible line in an attempt to find the one that minimizes of the sum of squared residuals. Instead, we use calculus to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that give the minimum sum of squared residuals. You don't need to memorize or use these equations, but here they are in case you are interested.

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Note that $\hat{\beta}_1$ and $\hat{\beta}_0$ are the *estimates* for the unknown parameters β_1 and β_0. In this way, they are analogous to sample means that estimate parameters like μ. Also, the best estimate of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

(Note that we divide by $n-2$ instead of $n-1$ here; in other words, the degrees of freedom is $n-2$. The reason is that we are "using up" two degrees of freedom to obtain $\hat{y}$—one to estimate β_0 and one to estimate β_1.) Whenever we ask a

computer to perform simple linear regression, it uses these equations to find the best fit line, then shows us the parameter estimates.

Here are the derivations of the coefficient estimates. SSR indicates sum of squared residuals, the quantity to minimize.

$$
\begin{aligned}
SSR &= \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 && (9.2)\\
&= \sum_{i=1}^{n}\left(y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2\right) && (9.3)\\
\frac{\partial SSR}{\partial \beta_0} &= \sum_{i=1}^{n}(-2y_i + 2\beta_0 + 2\beta_1 x_i) && (9.4)\\
0 &= \sum_{i=1}^{n}\left(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i\right) && (9.5)\\
0 &= -n\bar{y} + n\hat{\beta}_0 + \hat{\beta}_1 n\bar{x} && (9.6)\\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} && (9.7)\\
\frac{\partial SSR}{\partial \beta_1} &= \sum_{i=1}^{n}\left(-2x_i y_i + 2\beta_0 x_i + 2\beta_1 x_i^2\right) && (9.8)\\
0 &= -\sum_{i=1}^{n} x_i y_i + \hat{\beta}_0\sum_{i=1}^{n} x_i + \hat{\beta}_1\sum_{i=1}^{n} x_i^2 && (9.9)\\
0 &= -\sum_{i=1}^{n} x_i y_i + (\bar{y} - \hat{\beta}_1\bar{x})\sum_{i=1}^{n} x_i + \hat{\beta}_1\sum_{i=1}^{n} x_i^2 && (9.10)\\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} && (9.11)
\end{aligned}
$$

A little algebra shows that this formula for $\hat{\beta}_1$ is equivalent to the one shown above because $c\sum_{i=1}^{n}(z_i - \bar{z}) = c \cdot 0 = 0$ for any constant c and variable z.

In multiple regression, the matrix formula for the coefficient estimates is $(X'X)^{-1}X'y$, where X is the matrix with all ones in the first column (for

the intercept) and the values of the explanatory variables in subsequent columns.

Because the intercept and slope estimates are statistics, they have sampling distributions, and these are determined by the true values of β_0, β_1, and σ^2, as well as the positions of the x values and the number of subjects at each x value. If the model assumptions are correct, the sampling distributions of the intercept and slope estimates both have means equal to the true values, β_0 and β_1, and are Normally distributed with variances that can be calculated according to fairly simple formulas which involve the x values and σ^2.

In practice, we have to estimate σ^2 with s^2. This has two consequences. First we talk about the standard errors of the sampling distributions of each of the betas instead of the standard deviations, because, by definition, SE's are estimates of s.d.'s of sampling distributions. Second, the sampling distribution of $b_j - \beta_j$ (for j=0 or 1) is now the t-distribution with $n - 2$ df (see Section 3.9.5), where n is the total number of subjects. (Loosely we say that we lose two degrees of freedom because they are used up in the estimation of the two beta parameters.) Using the null hypothesis of $\beta_j = 0$ this reduces to the null sampling distribution $b_j \sim t_{n-2}$.

The computer will calculate the standard errors of the betas, the t-statistic values, and the corresponding p-values (for the usual two-sided alternative hypothesis). We then compare these p-values to our pre-chosen alpha (usually $\alpha = 0.05$) to make the decisions whether to retain or reject the null hypotheses.

To demonstrate, let's consider the corn experiment originally discussed in Figure 9.2. In that experiment, the outcome variable is corn weight and the explanatory variable is amount of nitrogen added to the soil. To run a linear regression in `R`, you use the `lm()` function with the format `lm(y~x, data = dataset)`, where `y` is the name of the outcome variable, `x` is the name of the explanatory variable, and `dataset` is the name of the dataset:

```
#run linear regression for the corn experiment
> linReg = lm(weight~nitrogen, data = corn)
#regression output
> summary(linReg)

Call:
lm(formula = weight ~ nitrogen, data = corn)
```

```
8
9  Residuals:
10     Min      1Q  Median      3Q     Max
11 -63.19  -33.41  -11.38   31.38  112.69
12
13 Coefficients:
14             Estimate Std. Error t value Pr(>|t|)
15 (Intercept)  84.8214    18.1158   4.682 0.000114 ***
16 nitrogen      5.2686     0.2992  17.610 1.87e-14 ***
17 ---
18 Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
          0.1         1
19
20 Residual standard error: 50.06 on 22 degrees of freedom
21 Multiple R-squared:  0.9338,  Adjusted R-squared:  0.9307
22 F-statistic: 310.1 on 1 and 22 DF,  p-value: 1.869e-14
```

There's a lot of information you get from a linear regression output in R; for now, we're going to focus on the following things you can learn from the above output:

- The "Estimate" column gives us the estimates of the coefficients in the linear regression. From this column, we can see that the estimate for the intercept β_0 is 84.82. (This can be also be called "estimated intercept" or "intercept estimate.") Typically, this is written as $\hat{\beta}_0 = 84.82$. Note that it is incorrect to write $\beta_0 = 84.82$, because β_0 is a fixed, unknown parameter value that is a "secret of nature." Meanwhile, we see that $\hat{\beta}_1 = 5.27$ is our estimated slope coefficient. (Sometimes symbols such as $\hat{\beta}_{\text{nitrogen}}$ or $\hat{\beta}_N$ are used to denote more meaningful notation, especially when dealing with multiple explanatory variables, which we'll discuss in the next chapter.)

- The "Std. Error" column gives the standard errors for the regression coefficients. For example, the SE of 0.30 for $\hat{\beta}_N$ gives an idea of the variability of our estimated slope. In this case, our estimated slope is 5.27, but it will vary with a standard deviation of approximately 0.30 around the true, unknown value of β_N if we repeat the whole experiment many times.

- The "t value" column represents the t-statistic for the null hypothesis $H_0 : \beta_j = 0$ (i.e., the first row denotes the statistic for $H_0 : \beta_0 = 0$ and the second row denotes the statistic for $H_0 : \beta_N = 0$). These statistics are computed using the general t-statistic formula is $t_j = \frac{\hat{\beta}_j - \text{hypothesized value of } \beta_j}{\text{SE}(\hat{\beta}_j)}$.

- The "Pr($>|t|$)" column gives us the p-value for the null hypothesis $H_0 : \beta_j = 0$. These p-values are computed using the t-statistics in the "t value" column. Specifically, the computer uses the null sampling distribution of the t-distribution with $n-2$ df to compute 2-sided p-values as the areas under the null sampling distribution more extreme (farther from zero) than the coefficient estimates for this experiment. (Here, the df is $n-2$ because we have estimated two parameters, the intercept and slope. Note that near the bottom of the output, we see "22 degrees of freedom," which means that there were 24 subjects in this study.)

The formulas for the standard errors come from the formula for the variance covariance matrix of the joint sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ which is $\sigma^2(X'X)^{-1}$, where X is the matrix with all ones in the first column (for the intercept) and the values of the explanatory variable in the second column. This formula also works in multiple regression where there is a column for each explanatory variable. The standard errors of the coefficients are obtained by substituting s^2 for the unknown σ^2 and taking the square roots of the diagonal elements.

For simple regression this reduces to

$$SE(\hat{\beta}_0) = s\sqrt{\frac{\sum_{i=1}^n x^2}{n\sum_{i=1}^n (x^2) - (\sum_{i=1}^n x)^2}}$$

and

$$SE(\hat{\beta}_1) = s\sqrt{\frac{n}{n\sum_{i=1}^n (x^2) - (\sum_{i=1}^n x)^2}}.$$

In simple regression the p-value for the null hypothesis $H_0 : \beta_1 = 0$ comes from the t-test for $\hat{\beta}_1$. If applicable, a similar test is made for β_0.

Note that the `summary()` output from `lm()` doesn't provide confidence intervals for the regression coefficients, which are often also of interest. To get confidence

intervals for linear regression coefficients, you use the `confint()` function after you define your linear regression model:

```
#95% CIs for regression coefficients
> confint(linReg)
                 2.5 %     97.5 %
(Intercept) 47.251480 122.391377
nitrogen     4.648124   5.889019
```

The confidence intervals give us additional information that we can't get from just the `summary()` output. For example, now we can say, "We are 95% confident that β_N is between 4.65 and 5.89." Previously, we could only confidently conclude that $\beta_N \neq 0$ (because we rejected the null hypothesis $H_0 : \beta_N = 0$, since the p-value was less than 0.05), but now we have a better idea of the magnitude of β_N. Similar to the confidence intervals we saw in ANOVA, the confidence intervals for the regression coefficients are computed using the appropriate t distribution and the SEs we saw in the `summary()` output. Although you do not need to know how to compute the confidence intervals for regression coefficients, you do need to know how to interpret them. For example, if 0 is contained in the 95% confidence interval for β_N, then we fail to reject the null hypothesis that $H_0 : \beta_N = 0$ (with 95% confidence).

The confidence interval for a regression parameter (e.g., β_1) gives a meaningful measure of the location of the parameter and our uncertainty about that location, regardless of whether or not the null hypothesis is true.

9.5 Interpreting regression coefficients

It is very important that you learn to correctly and completely interpret the coefficient estimates. From $E(Y|x) = \beta_0 + \beta_1 x$ we can see that $\hat{\beta}_0$ represents our estimate of the mean outcome when $x = 0$. Before making an interpretation of $\hat{\beta}_0$, first check the range of x values covered by the experimental data. If there is no x data near zero, then the intercept is still needed for calculating $\hat{y}$ and residual values, but it should be intepreted with a great deal of caution because it is an extrapolated value.

If there are x values near zero, then to interpret the intercept you must express it in terms of the actual meanings of the outcome and explanatory variables. For the example of this chapter, we would say that $\hat{\beta}_0$ (84.82) is the estimated corn plant weight (in grams) when no nitrogen is added to the pots (which is the meaning of $x = 0$). This point estimate is of limited value, because it does not express the degree of uncertainty associated with it. Often it is better to report the CI for $\hat{\beta}_0$. In this case, we say that we are 95% confident that the mean weight for corn plants with no added nitrogen is between 47 and 122 gm, which is quite a wide range.

After interpreting the *estimate* of $\hat{\beta}_0$ and its CI, you should consider whether the *null hypothesis*, $\beta_0 = 0$ makes scientific sense. For the corn example, the null hypothesis is that the mean plant weight equals zero when no nitrogen is added. Because it is unreasonable for plants to weigh nothing, we should stop here and not interpret the p-value for the intercept. As another example, consider a regression of weight gain in rats over a 6-week period as it relates to dose of an anabolic steroid. Because it is possible that rats may have 0 weight gain over 6 weeks (and because it is possible to give the rats 0 dosage of the drug), it might make sense to test $H_0 : \beta_0 = 0$. If the null hypothesis is rejected, then we conclude that the weight gain is non-zero when the dose is zero (the control group). This means that other factors (e.g., time) are affecting weight gain and not just dose of the drug, which is helpful information to know.

Interpret the estimate, $\hat{\beta}_0$, only if there are data near zero and setting the explanatory variable to zero makes scientific sense. The meaning of $\hat{\beta}_0$ is the estimate of the mean outcome when $x = 0$, and should always be stated in terms of the actual variables of the study. The p-value for the intercept should be interpreted (with respect to retaining or rejecting $H_0 : \beta_0 = 0$) only if both the equality and the inequality of the mean outcome to zero when the explanatory variable is zero are scientifically plausible.

For interpretation of a slope coefficient, this section will assume that the setting is a randomized experiment, and conclusions will be expressed in terms of causation. Be sure to substitute association if you are looking at an observational study or the slope coefficient of an explanatory variable that is not a treatment variable. The general meaning of a slope coefficient is the change in Y caused by a one-unit

increase in x. It is very important to know in what units x are measured, so that the meaning of a one-unit increase can be clearly expressed. For the corn experiment, the slope is the change in mean corn plant weight (in grams) caused by a one mg increase in nitrogen added per pot. If a one-unit change is not substantively meaningful, the effect of a larger change should be used in the interpretation. For the corn example we could say the a 10 mg increase in nitrogen added causes a 52.7 gram increase in plant weight on average. We can also interpret the CI for β_1 in the corn experiment by saying that we are 95% confident that the change in mean plant weight caused by a 10 mg increase in nitrogen is 46.8 to 58.9 gm.

Be sure to pay attention to the sign of $\hat{\beta}_1$. If it is positive, then $\hat{\beta}_1$ represents the increase in outcome caused by each one-unit increase in the explanatory variable. If $\hat{\beta}_1$ is negative, then each one-unit increase in the explanatory variable is associated with a *decrease* in the outcome of magnitude equal to the absolute value of $\hat{\beta}_1$.

A significant p-value indicates that we should reject the null hypothesis that $\beta_1 = 0$. For the corn experiment example, we can express this as evidence that plant weight is affected by changes in nitrogen added. If the null hypothesis is retained, we should express this as having no good evidence that nitrogen added affects plant weight.

The interpretation of $\hat{\beta}_1$ is the change (increase or decrease depending on the sign) in the average outcome when the explanatory variable increases by one unit. This should always be stated in terms of the actual variables of the study. Retention of the null hypothesis $H_0 : \beta_1 = 0$ indicates no evidence that a change in x is associated with (or "causes" if it is a treatment variable for a randomized experiment) a change in y. Rejection indicates that changes in x is associated with (or, again, "causes," if appropriate) changes in y.

9.6 Residual checking

Every regression analysis should include a residual analysis as a further check on the adequacy of the chosen regression model. Remember that there is a residual value for each data point, and that it is computed as the difference $y_i - \hat{y}_i$. A pos-

itive residual indicates a data point higher than expected, and a negative residual indicates a point lower than expected.

A residual is the deviation of an outcome from the predicated mean value for all subjects with the same value for the explanatory variable.

Remember that the statistical model for simple linear regression is $Y \overset{iid}{\sim} N(\beta_0 + \beta_1 x, \sigma^2)$. Because of the variance σ^2, we fully expect deviations from the linear trend $\beta_0 + \beta_1 x$, where some of the deviations will be positive and others will be negative. Furthermore, this model (like the model for ANOVA) makes the assumptions of Normality, equal variance, and independent errors. A plot of all residuals on the y-axis vs. the predicted values on the x-axis, called a **residual-versus-fit plot**, is a good way to check the linearity and equal variance assumptions. A quantile-normal plot of all of the residuals is a good way to check the Normality assumption.

To analyze a residual-versus-fit plot, such as any of the examples shown in Figure 9.4, you should mentally divide it up into about 5 to 10 vertical stripes. Then each stripe represents all of the residuals for a number of subjects who have a similar predicted values. For simple regression, when there is only a single explanatory variable, similar predicted values is equivalent to similar values of the explanatory variable. But be careful, if the slope is negative, low x values are on the right, because lower x values will lead to higher fitted values in this case. (Note that sometimes the x-axis is set to be the values of the explanatory variable, in which case each stripe directly represents subjects with similar x values.)

To check the linearity assumption, consider that for each x value, if the mean of Y falls on a straight line, then the residuals for that particular x value should have a mean of zero. If we incorrectly fit a straight line to a curve, then some or most of the predicted means are incorrect, and this causes the residuals for at least specific ranges of x (or the predicated Y) to be non-zero on average. In other words, the residuals should have no kind of trend (linear or nonlinear) across fitted values if the linearity assumption is true. Specifically, if the data follow a simple curve, we will tend to have either a pattern of high then low then high residuals or the reverse. So the technique used to detect non-linearity in a residual vs. fit plot is to find the (vertical) mean of the residuals for each vertical stripe, then actually

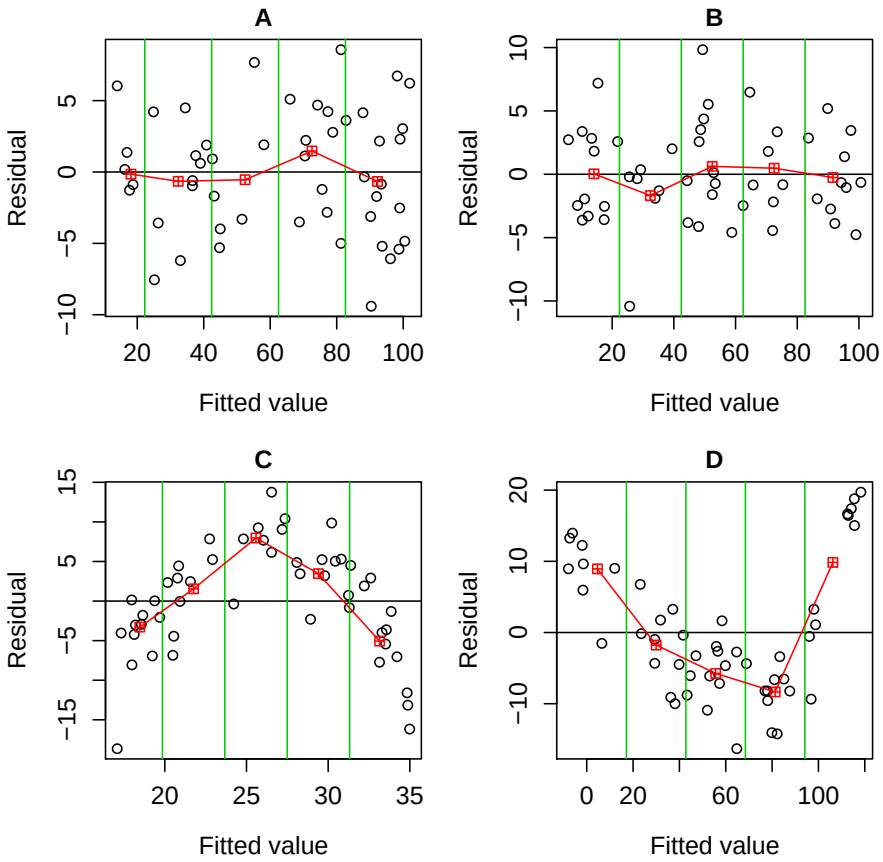


Figure 9.4: Sample residual-versus-fit plots for testing linearity.

or mentally connect those means, either with straight line segments, or possibly with a smooth curve. If the resultant connected segments or curve is close to a horizontal line at 0 on the y-axis, then we have no reason to doubt the linearity assumption. If there is a clear curve, most commonly a "smile" or "frown" shape, then we suspect non-linearity.

Four examples are shown in Figure 9.4. In each band the mean residual is marked, and lines segments connect these. Plots A and B show no obvious pattern away from a horizontal line other than the small amount of expected "noise". Plots C and D show clear deviations from normality, because the lines connecting the mean residuals of the vertical bands show a clear frown (C) and smile (D) pattern, rather than a flat line. Untransformed linear regression is inappropriate for the data that produced plots C and D. With practice you will get better at reading

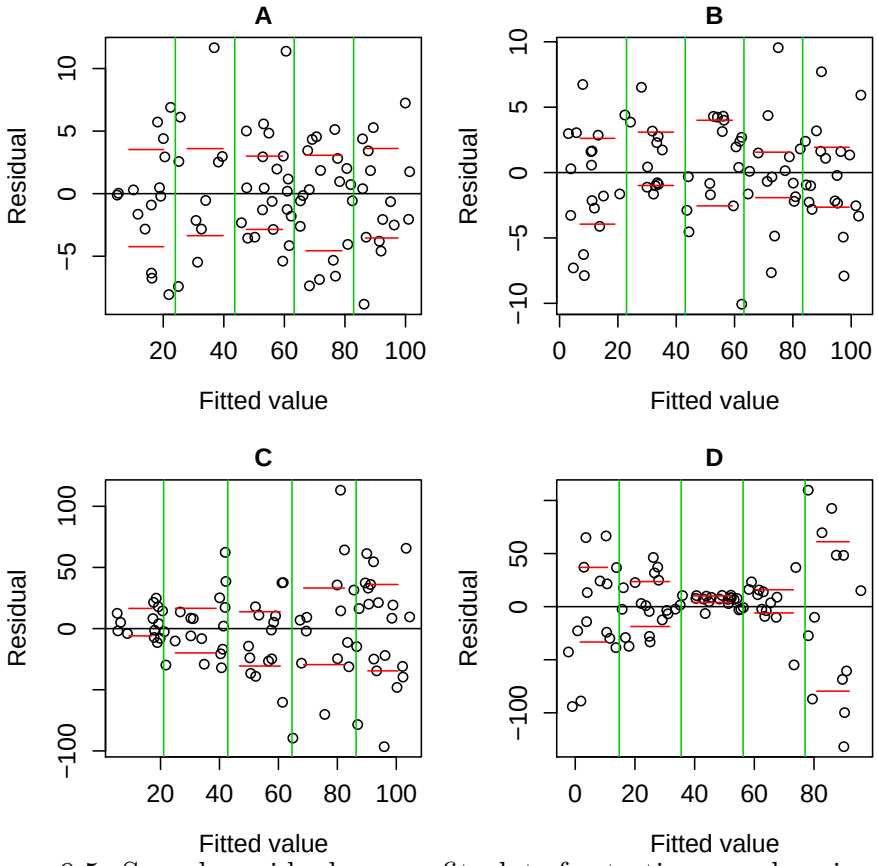


Figure 9.5: Sample residual-versus-fit plots for testing equal variance.

these plots.

To detect unequal spread, we use the vertical bands in a different way. Ideally the vertical spread of residual values is equal in each vertical band. This takes practice to judge in light of the expected variability of individual points, especially when there are few points per band. The main idea is to realize that the minimum and maximum residual in any set of data is not very robust, and tends to vary a lot from sample to sample. We need to estimate a more robust measure of spread such as the IQR. This can be done by eyeballing the middle 50% of the data. Eyeballing the middle 60 or 80% of the data is also a reasonable way to test the equal variance assumption.

Figure 9.5 shows four residual-versus-fit plots, each of which shows good linearity. The red horizontal lines mark the central 60% of the residuals. Plots A and

B show no evidence of unequal variance; the red lines are a similar distance apart in each band. In plot C you can see that the red lines increase in distance apart as you move from left to right. This indicates unequal variance, with greater variance at high predicted values (high x values if the slope is positive). Plot D shows a pattern with unequal variance in which the smallest variance is in the middle of the range of predicted values, with larger variance at both ends. Again, this takes practice, but you should at least recognize obvious patterns like those shown in plots C and D. And you should avoid over-reading the slight variations seen in plots A and B.

The residual-versus-fit plot can be used to detect non-linearity and/or unequal variance.

The check of normality can be done with a quantile normal plot, as seen in Figure 9.6. Plot A shows no problem with Normality of the residuals because the points show a random scatter around the reference line (see Section 4.3.3). Plot B is also consistent with Normality, perhaps showing slight skew to the left. Plot C shows definite skew to the right, because at both ends we see that several points are higher than expected. Plot D shows a severe low outlier as well as heavy tails (positive kurtosis) because the low values are too low and the high values are too high.

A quantile normal plot of the residuals of a regression analysis can be used to detect non-Normality.

As a demonstration of how to make these diagnostics in R, we will show how to make a residual-versus-fit plot and quantile-normal plot for the corn experiment example. To obtain the residuals from a linear regression model, you use the `residuals()` function, and to obtain the fitted values, you use the `predict()` function. In mathematical notation, `residuals()` gives you the $(y_i - \hat{y}_i)$, and `predict()` gives you the $\hat{y}_i$. Here's the code for making the residual-versus-fit and quantile-normal plots for the corn experiment:

```
#run linear regression for the corn experiment
> linReg = lm(weight~nitrogen, data = corn)
```

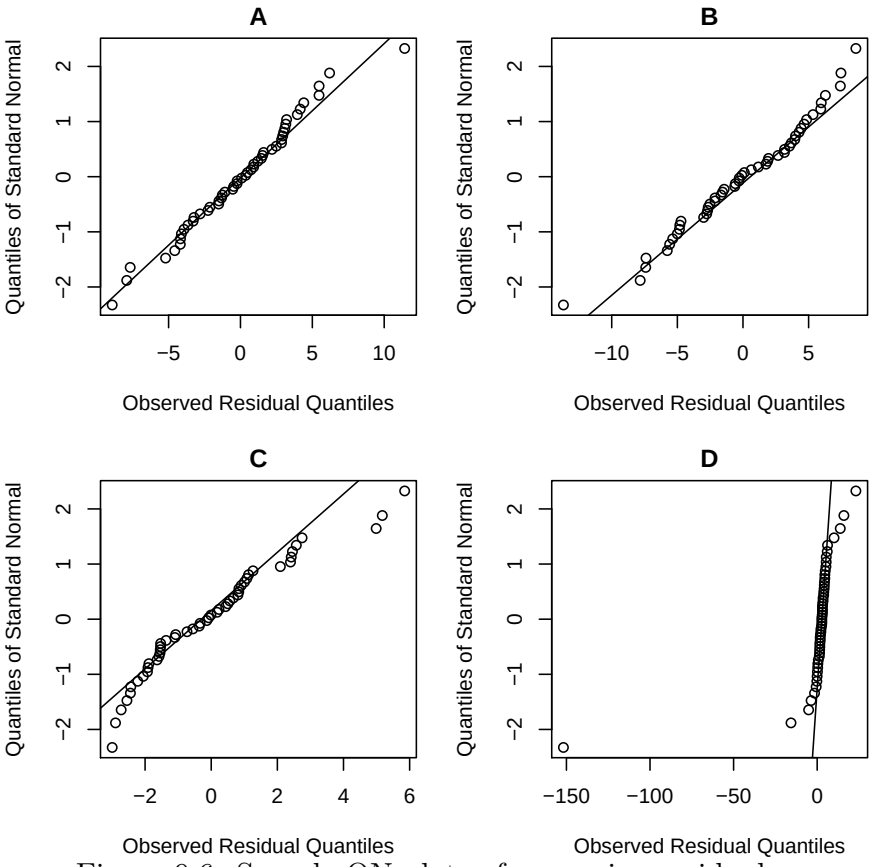
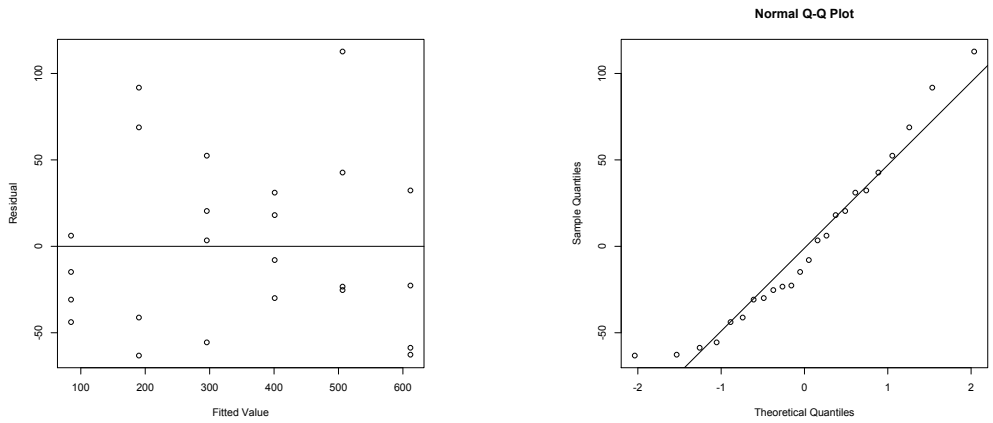


Figure 9.6: Sample QN plots of regression residuals.

```
#residuals
> linReg.res = residuals(linReg)
#fitted values
> linReg.fit = predict(linReg)
#residual versus fit plot
> plot(linReg.fit, linReg.res,
+      xlab = "Fitted Value", ylab = "Residual")
> abline(h = 0)
#residual versus fit plot
> qqnorm(linReg.res)
> qqline(linReg.res)
```

Note that for the residual-versus-fit plot, we added a horizontal vertical line at 0 using `abline(h = 0)`, which helps us see if there are any notable trends away from 0. Meanwhile, we use the `qqnorm()` and `qqline()` functions to make a quantile-normal plot; note that we make a quantile-normal plot of the *residuals*. The residual-versus-fit plot is shown in Figure 9.7a, and the quantile-normal plot is shown in Figure 9.7b. Neither plot shows any clear violations of the linear model assumptions: In the residual-versus-fit plot, there are no clear trends around 0 (indicating that the linearity assumption is plausible) and there is an even spread across the plot (indicating that the equal variance assumption is plausible); and in the quantile-normal plot, the residuals seem to fall relatively well along the diagonal line (indicating that the normality assumption is plausible).



(a) Residual-versus-fit plot.

(b) Quantile-normal plot.

Figure 9.7: Residual checking for the corn experiment.

9.7 Robustness of simple linear regression

No model perfectly represents the real world. It is worth learning how far we can "bend" the assumptions without breaking the value of a regression analysis.

The more the linearity assumption is violated, the larger degree to which the linear regression loses its meaning. The most obvious way this happens is in the interpretation of $\hat{\beta}_1$. We interpret $\hat{\beta}_1$ as the change in the mean of Y for a one-unit increase in x. If the relationship between x and Y is curved, then the change in Y for a one-unit increase in x *varies* at different parts of the curve, invalidating the interpretation. Luckily it is fairly easy to detect non-linearity through EDA (scatterplots) and/or residual analysis. If non-linearity is detected, you should try to address it by transforming the x and/or y variables. Common transformations (usually for the outcome variable) are log and square root. Alternatively, it is common to *add* additional explanatory variables in the form of a square, cube, etc. of the original x variable one at a time until the residual-versus-fit plot shows that linearity is a reasonable assumption. For data that can only lie between 0 and 1, it is worth knowing (but not memorizing) that the square root of the arcsine of y is often a good transformation.

You should not feel that transformations are "cheating". In fact, a lot of data is already recorded as some kind of transformation. For example, pH for acidity, decibels for sound, and the Richter earthquake scale are all typically measured on log scales. When reporting results, transformed values are usually transformed back to the original scale (but it must also be reported that the analysis was on a transformed scale).

Regression is reasonably robust to the equal variance assumption. Moderate degrees of violation, e.g., the band with the widest variation is up to twice as wide as the band with the smallest variation, tend to cause minimal problems. For more severe violations, the p-values are incorrect in the sense that their null hypotheses tend to be rejected more that $100\alpha\%$ of the time when the null hypothesis is true. The confidence intervals (and the SE's they are based on) are also incorrect. For worrisome violations of the equal variance assumption, try transformations of the y variable (because the assumption applies at each x value, transformation of x will be ineffective).

Regression is quite robust to the Normality assumption. You only need to worry about severe violations. For markedly skewed or kurtotic residual distributions, we need to worry that the p-values and confidence intervals are incorrect. In that

case, try transforming the y variable. For example, data on income are often right-skewed, and for this reason it is very standard to take a log transformation of income data. Also, in the case of data with less than a handful of different y values or with severe truncation of the data (values piling up at the ends of a limited width scale), regression may be inappropriate due to non-Normality.

The fixed-x assumption is actually quite important for regression. If the variability of *a single* x measurement is of similar or larger magnitude to the variability of *a single* y measurement, then regression is inappropriate. Regression will tend to give smaller than correct slopes under these conditions, and the null hypothesis on the slope will be retained far too often. Alternate techniques are required if the fixed-x assumption is broken, including so-called Type 2 regression or "errors in variables regression".

The independent errors assumption is also critically important to regression. A slight violation, such as a few twins in the study doesn't matter, but other mild to moderate violations destroy the validity of the p-value and confidence intervals. In that case, use alternate techniques such as the paired t-test, repeated measures analysis, mixed models, or time series analysis, all of which model correlated errors rather than assume zero correlation.

Regression analysis is not very robust to violations of the linearity, fixed-x, and independent errors assumptions. It is somewhat robust to violation of equal variance, and moderately robust to violation of the Normality assumption.

9.8 Additional interpretation of regression output

Regression output usually includes a few additional components beyond the slope and intercept estimates and their t and p-values. Let's again look at the regression output from the corn experiment example:

```
#run linear regression for the corn experiment
> linReg = lm(weight~nitrogen, data = corn)
#regression output
```

```
> summary(linReg)

Call:
lm(formula = weight ~ nitrogen, data = corn)

Residuals:
   Min     1Q Median     3Q    Max
-63.19 -33.41 -11.38  31.38 112.69

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.8214    18.1158   4.682 0.000114 ***
nitrogen      5.2686     0.2992  17.610 1.87e-14 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 50.06 on 22 degrees of freedom
Multiple R-squared:  0.9338,  Adjusted R-squared:  0.9307
F-statistic: 310.1 on 1 and 22 DF,  p-value: 1.869e-14
```

First, let's look at the output under the coefficients table. The "Residual standard error" (50.06) is the standard error of the residuals from the linear regression, and it it is the best estimate of σ from our model. Thus, it represents how far we expect data to fall from the estimated regression line, on the scale of the outcome variable. For example, for the corn analysis above, we estimate that only about 5% of the data falls more than 1.96(50.06)=98.12 gm away from the estimated regression line.

Meanwhile, the "Multiple R-squared" value - also known as the R^2 or **multiple correlation coefficient** - is equal to the square of the correlation between x and y in simple regression but not in multiple regression (as we will discuss in the next chapter). In either case, R^2 can be interpreted as the fraction of the total variation in the outcome that is "accounted for" by regressing the outcome on the explanatory variable. An $R^2 = 1$ means that the data falls perfectly along a straight line; we see that $R^2 = 0.93$, meaning that the data falls nearly along a straight line (look back at Figure 9.2 - this is indeed the case). We also see the "Adjusted R-squared"; this quantity is strictly less than R^2 and penalizes the use of many explanatory variables in the regression. Because there is only one variable used in this regression, the R^2 and Adjusted R^2 are nearly identical. We will discuss Adjusted R^2 in Chapter 10.

A little math can help us better understand what R^2 is. The total variance, var(Y), in a regression problem is the sample variance of y ignoring x, which comes from the squared deviations of y values around the mean of y. Since the mean of y is the best guess of the outcome for any subject if the value of the explanatory variable is unknown, we can think of total variance as measuring how well we can predict y without knowing x.

If we perform regression and then focus on the residuals, these values represent our residual error variance when predicting y while *using* knowledge of x. The estimate of this variance is called mean squared error or MSE and is the best estimate of the quantity σ^2 defined by the regression model.

If we subtract total minus residual error variance (var(Y)-MSE), we can call the result the "explained error". It represents the amount of variability in y that is *explained* away by regressing on x. Then we can compute R^2 as

$$R^2 = \frac{\text{explained variance}}{\text{total variance}} = \frac{\text{var}(Y) - \text{MSE}}{\text{var}(Y)} = 1 - \frac{\text{MSE}}{\text{var}(Y)}.$$

So R^2 is the portion of the total variation in Y that is explained away by using the x information in a regression. R^2 is always between 0 and 1. An R^2 of 0 means that x provides no information about y. An R^2 of 1 means that use of x information allows perfect prediction of y with every point of the scatterplot exactly on the regression line. Anything in between represents different levels of closeness of the scattered points around the regression line.

So for the corn problem we can say that 93.4% of the total variation in plant weight can be explained by regressing on the amount of nitrogen added. Unfortunately, there is no clear general interpretation of the values of R^2. While $R^2 = 0.6$ might indicate a great finding in social sciences, it might indicate a very poor finding in a chemistry experiment.

R^2 is a measure of the fraction of the total variation in the outcome that can be explained by the explanatory variable. It runs from 0 to 1, with 1 indicating perfect prediction of y from x.

9.9 Using transformations

If you find a problem with the equal variance or Normality assumptions, you will probably want to see if the problem goes away if you use $\log(y)$ or y^2 or $\sqrt{y}$ or $1/y$ instead of y for the outcome. (It never matters whether you choose natural vs. common log.) For non-linearity problems, you can try transformation of x, y, or both. If regression on the transformed scale appears to meet the assumptions of linear regression, then go with the transformations. In most cases, when reporting your results, you will want to back transform point estimates and the ends of confidence intervals for better interpretability. By "back transform" I mean do the inverse of the transformation to return to the original scale. The inverse of common log of y is 10^y; the inverse of natural log of y is e^y; the inverse of y^2 is $\sqrt{y}$; the inverse of $\sqrt{y}$ is y^2; and the inverse of $1/y$ is $1/y$ again. *Do not transform a p-value – the p-value remains unchanged.*

Here are a couple of examples of transformation and how the interpretations of the coefficients are modified. If the explanatory variable is dose of a drug and the outcome is log of time to complete a task, and $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = 1.5$, then we can say the best estimate of the log of the task time when no drug is given is 2 or that the best estimate of the time is $10^2 = 100$ or $e^2 = 7.39$ depending on which log was used. We also say that for each 1 unit increase in drug, the log of task time increases by 1.5 (additively). On the original scale this is a *multiplicative* increase of $10^{1.5} = 31.6$ or $e^{1.5} = 4.48$. Assuming natural log, this says every time the dose goes up by another 1 unit, the mean task time gets multiplied by 4.48.

If the explanatory variable is common log of dose and the outcome is blood sugar level, and $\hat{\beta}_0 = 85$ and $\hat{\beta}_1 = 18$ then we can say that when log(dose)=0, blood sugar is 85. Using $10^0 = 1$, this tells us that blood sugar is 85 when dose equals 1. For every 1 unit increase in log dose, the glucose goes up by 18. But a one unit increase in log dose is a ten fold increase in dose (e.g., dose from 10 to 100 is log dose from 1 to 2). So we can say that every time the dose increases 10-fold the glucose goes up by 18.

As an example of how to use transformations in linear regression in `R`, here's the code to run a linear regression between `weight` and squared `nitrogen` for the corn example:

```
> quadReg = lm(weight~I(nitrogen^2), data = corn)
> summary(quadReg)

Call:
```

```
lm(formula = weight ~ I(nitrogen^2), data = corn)

Residuals:
     Min       1Q   Median       3Q      Max
-133.331  -70.561    4.101   67.108  141.101

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.743e+02  2.409e+01   7.236 3.00e-07 ***
I(nitrogen^2) 4.743e-02  4.715e-03  10.059 1.08e-09 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 82.2 on 22 degrees of freedom
Multiple R-squared:  0.8214,  Adjusted R-squared:  0.8133
F-statistic: 101.2 on 1 and 22 DF,  p-value: 1.084e-09
```

Note that you have to use the `I()` function notation to insert transformations in linear regressions in `R`; just writing `weight~nitrogen^2` wouldn't have worked. From the output, we may conclude that there is a quadratic relationship between nitrogen and weight (the p-value for the slope coefficient is quite low, and $R^2 = 0.82$); however, we can note that the R^2 was higher when we just included nitrogen instead of squared nitrogen, and in fact the quadratic term becomes non-significant after we add in the linear term for nitrogen (not shown here). We will talk about these points in much more depth in the next chapter.

Transformations of x or y to a different scale are very useful for fixing broken assumptions. However, your interpretations of the results have to change according to the change of scale of the variables being studied.

9.10 How to perform simple linear regression in R

We've already demonstrated how to use `R` to run linear regression throughout this chapter. However, for reference, we'll include a short description of the functions

we used throughout this chapter:

- `lm()`: Run a linear regression ("linear model"). For example, write `lm(y~x, data = dataset)` to run a linear regression between the variables `y` (as the outcome) and `x` (as the explanatory variable) in a dataset called `dataset`.
- `summary()`: Get output from a linear regression model. You put something from `lm()` as the argument for this function. For example, you'll write something of the form `summary(lm(y~x, data = dataset))`.
- `confint()`: Get confidence intervals for a linear regression model. You put something from `lm()` as the argument for this function. For example, you'll write something of the form `confint(lm(y~x, data = dataset))`.
- `residuals()`: Get the residuals of linear regression model. These values represent the deviations between the real data and the fitted line, i.e., $y_i - \hat{y}_i$. Use this in conjunction with `qqnorm()` and `qqline()` to make the appropriate quantile-normal plot as a residual diagnostic check for a linear model.
- `predict()`: Get the fitted values of a linear regression model. These values represent our estimate of the outcome based on the linear model, i.e., $\hat{y}_i$. Use this in conjunction with `residuals()` to make a residuals-versus-fit plot as a residual diagnostic check for a linear regression.
- `I()`: Define a transformation to be used in a linear regression model. Use this function within the `lm()` function. For example, write `lm(y~I(x^2), data = dataset)` to run a linear regression between the variables `y` (as the outcome) and $\texttt{x}^2$ (as the explanatory variable) in a dataset called `dataset`.

In a nutshell: Simple linear regression is used to explore the relationship between a quantitative outcome and a quantitative explanatory variable. The p-value for the slope, $\hat{\beta}_1$, is a test of whether or not changes in the explanatory variable really are associated with changes in the outcome. The interpretation of the confidence interval for β_1 is usually the best way to convey what has been learned from a study. Occasionally there is also interest in the intercept. No interpretations should be given if the assumptions are violated, as determined by thinking about the fixed-x and independent errors assumptions, and checking the residual-versus-fit and residual QN plots for the other three assumptions.

Chapter 10

Analysis of Covariance

An analysis procedure for looking at group effects on a continuous outcome when other continuous explanatory variables may also have an effect on the outcome.

This chapter introduces several new important concepts, including multiple regression, interaction, and use of indicator variables, and then uses them to present a model appropriate for experiments where there is a quantitative outcome, a categorical treatment variable, and one or more quantitative explanatory variables. Generally, the main interest is in the effects of the categorical variable, and the quantitative explanatory variable is considered to be a "control" variable, such that power is improved if its value is controlled for. Using the principles explained here, it is relatively easy to extend the ideas to additional categorical and quantitative explanatory variables.

The term ANCOVA, analysis of covariance, is commonly used in this setting, although there is some variation in how the term is used. In some sense ANCOVA is a blending of ANOVA and regression.

10.1 Multiple regression

Before you can understand ANCOVA, you need to understand multiple regression. Multiple regression is a straightforward extension of simple regression from one to several quantitative explanatory variables (and also categorical variables as we will see in Section 10.4). For example, if we vary water, sunlight, and fertilizer to see

their effects on plant growth, we have three quantitative explanatory variables. In this case we write the structural model as

$$E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Remember that $E(Y|x_1, x_2, x_3)$ is read as expected (i.e., average) value of Y (the outcome) given the values of the explanatory variables x_1 through x_3. Here, x_1 is the amount of water, x_2 is the amount of sunlight, x_3 is the amount of fertilizer, β_0 is the intercept, and the other βs are all slopes. Of course, we can have any number of explanatory variables as long as we have a different β parameter for each explanatory variable.

Although the use of numeric subscripts for the different explanatory variables (x's) and parameters (β's) is quite common, I think that it is usually nicer to use meaningful mnemonic letters for the explanatory variables and corresponding text subscripts for the parameters to remove the necessity of remembering which number goes with which explanatory variable. Unless referring to variables in a completely generic way, I will avoid using numeric subscripts here (except for using β_0 to refer to the intercept). So the above structural equation is better written as

$$E(Y|W, S, F) = \beta_0 + \beta_W W + \beta_S S + \beta_F F.$$

In multiple regression, we still make the same five assumptions as we did for simple linear regression in Chapter 9:

1. Normality
2. Equal variance (σ^2)
3. Independent errors
4. Linearity
5. Fixed x

Let's examine what the above multiple regression structural model is claiming, i.e., in what situations it might be plausible. By examining the equation for the multiple regression structural model you can see that the meaning of each slope coefficient is that it is the change in the mean outcome associated with (or caused by) a one-unit rise in the corresponding explanatory variable *when all of the other explanatory variables are held constant.*

We can see this by taking the approach of writing down the structural model equation and then making it reflect specific cases. Here is how we find what happens to the mean outcome when x_1 is fixed at, say 5, and x_2 at, say 10, and x_3 is allowed to vary.

$$\begin{aligned} E(Y|x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ E(Y|x_1 = 5, x_2 = 10, x_3) &= \beta_0 + 5\beta_1 + 10\beta_2 + \beta_3 x_3 \\ E(Y|x_1 = 5, x_2 = 10, x_3) &= (\beta_0 + 5\beta_1 + 10\beta_2) + \beta_3 x_3 \end{aligned}$$

Because the βs are fixed (but unknown) constants, this equation tells us that when x_1 and x_2 are fixed at the specified values, the relationship between $E(Y)$ and x_3 can be represented on a plot with the outcome on the y-axis and x_3 on the x-axis as a straight line with slope β_3 and intercept equal to the number $\beta_0 + 5\beta_1 + 10\beta_2$. Similarly, we get the same slope with respect to x_3 for any combination of x_1 and x_2, and this idea extends to changing any one explanatory variable when the others are held fixed.

By simplifying the structural model to specific cases, we learn that the above multiple regression model claims that not only is there a linear relationship between $E(Y)$ and any x when the other x's are held constant, but also the effect of a given change in an x value does not depend on what the values of the other x variables are set to, as long as they are held constant. In other words, there are no *interactions* among the x variables. We will discuss models that include interactions shortly.

When we discussed t-tests and ANOVA for analyzing experiments in Chapters 5 and 6, respectively, our primary interest was in estimating treatment effects. The statistical models behind t-tests and ANOVA assume that there are only two variables: outcome variables and treatment variables. However, including additional explanatory variables (or *control variables*) in statistical models can give us useful information to better estimate treatment effects in experiments. Primary interpretation is still focused on the experimental treatment variable in these statistical models, but including control variables in the models often increases power in detecting treatment effects if they exist. Control variables function in the same way as blocking variables (see Section 7.5) in that they affect the outcome but are not of primary interest. Ideally, control variables are highly related to the outcome—when this is the case, including control variables can greatly increase the precision of estimating treatment effects. Examples of control variables for many psychological studies include things like ability (as determined by some aux-

iliary information) and age. Examples in education include pre-test scores and socioeconomic status. Examples in medicine include age, gender, pre-existing conditions, and other health-related variables.

As an example of multiple regression with two manipulated quantitative variables, consider an analysis of the data of MRdistract.dat which is from a (fake) experiment testing the effects of both visual and auditory distractions on reading comprehension. The outcome is a reading comprehension test score administered after each subject reads an article in a room with various distractions. The test is scored from 0 to 100 with 100 being best. The subjects are exposed to auditory distractions that consist of recorded construction noise with the volume randomly set to vary between 10 and 90 decibels from subject to subject. The visual distraction is a flashing light at a fixed intensity but with frequency randomly set to between 1 and 20 times per minute.

Exploratory data analysis is difficult in the multiple regression setting because we need more than a two dimensional graph. In this example we have three quantitative variables: The outcome is the test score, and the two explanatory variables are the decibel level and number of distractions. How can we visualize all three quantitative variables in a single graph?

You can make three-dimensional scatterplots in R, but they are often hard to read. One alternative is to plot the outcome separately against each explanatory variable (i.e., make a scatterplot for score-versus-decibel and score-versus-distractions). However, this provides a limited amount of information; in particular, this won't tell us anything about the joint distribution of the two explanatory variables decibel and number of distractions. Another option commonly done is to categorize one of the explanatory variables, and then make a scatterplot colored by those categories. For example, note that the number of distractions variable (called `freq` in the dataset) ranges between 1 and 20; so, we can categorize this variable into the ranges [1-5], [6-10], [11-15], [16-20]. The following code (which uses the `ceiling()` function) creates this categorical variable (either 1, 2, 3, or 4):

```
#the original freq variable (ranging 1-20)
> distract$freq
7  6 15  2  9 18 15  7  5  2 16  5 19 19  5  4 14  3 12  4 14 16
   5 12 18 16  6 15 11 10 10 17  5 10  5 17 10  6 14  4
#a categorized version (ranging 1-4)
> ceiling(distract$freq/5)
2 2 3 1 2 4 3 2 1 1 4 1 4 4 1 1 3 1 3 1 3 4 1 3 4 4 2 3 3 2 2 4 1
  2 1 4 2 2 3 1
```

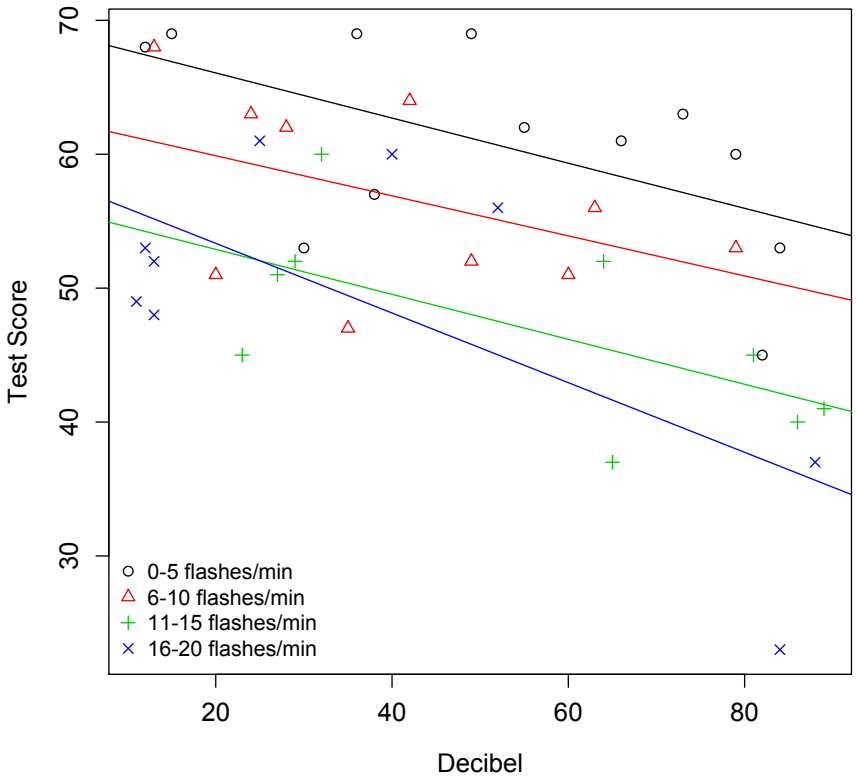


Figure 10.1: EDA for the distraction example.

In other words, we have now created a variable that can be defined as "low distractions" (1), "medium-low distractions" (2), "medium-high distractions" (3), and "high distractions" (4). Then, we can make a scatterplot of score-versus-decibel colored by this categorical variable (Figure 10.1). For those curious, here is the code to make this figure:

```
> plot(distract$db,distract$score, xlab="Decibel", ylab="Test
   Score",
+       cex.axis=1.2, cex.lab=1.2,
+       pch=ceiling(distract$freq/5),col=ceiling(distract$freq/5)
   )
#add regression lines for each level of freq
> for (i in 1:4) {
+   #grab subset of data that corresponds
+   #to the particular level of freq
+   currentSubset = subset(distract, ceiling(freq/5) == i)
+   abline(lm(score~db, data = currentSubset), col=i)
+ }
#add a legend to the plot
> legend("bottomleft",
+   legend = c("0-5 flashes/min", "6-10 flashes/min",
+     "11-15 flashes/min", "16-20 flashes/min"),
+   pch = 1:4, col = 1:4,
+   bty = "n")
```

Note that we added regression lines for each level of the categorical variable to make it easier to see general trends, and we also changed the shape *and* color of each dot so that datapoints belonging to different categories stand out even more (which makes the graph black-and-white friendly).

Here we can see that increasing the value of either explanatory variable tends to reduce the mean outcome. Although the fit lines are not parallel, with a little practice you will be able to see that given the uncertainty in setting their slopes from the data, they are actually consistent with parallel lines, which is an indication that no interaction is needed. We'll discuss how to formally test for interaction effects (i.e., non-parallel lines) in the next section.

But first, let's look at some results when we regress the outcome variable (`score`) on the two explanatory variables (`db` and `freq`):

```
#regression with main effects
> distractReg = lm(score~db + freq, data = distract)
> #regression output
> summary(distractReg)

```

```
Call:
lm(formula = score ~ db + freq, data = distract)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8741  -5.2087   0.6954   5.3594  11.8417

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.68812    3.26002  22.910  < 2e-16 ***
db          -0.20007    0.04261  -4.695 3.60e-05 ***
freq        -1.11812    0.20796  -5.377 4.39e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 6.939 on 37 degrees of freedom
Multiple R-squared:  0.5528,  Adjusted R-squared:  0.5286
F-statistic: 22.87 on 2 and 37 DF,  p-value: 3.422e-07

#confidence intervals
> confint(distractReg)
                 2.5 %     97.5 %
(Intercept) 68.0826923 81.2935561
db          -0.2864126 -0.1137295
freq        -1.5394883 -0.6967490
```

Note that we're still using the `y~x` formula notation within the `lm()`, but now we have two variables on the right side (`db` and `freq`), and we use the `+` symbol to "add" more variables to the model. In other words, the structural model we are fitting here is:

$$E[\texttt{score}|\texttt{db}, \texttt{freq}] = \beta_0 + \beta_{db} \cdot \texttt{db} + \beta_f \cdot \texttt{freq}$$

which motivates the `+` notation within `lm()`. Importantly, note that our estimates in the coefficient table are indeed our estimates of the parameters β_0, β_{db}, and β_f in this structural model.

Now here is an interpretation of the analysis of this experiment. A multiple regression analysis (additive model, i.e., with no interaction) was performed using sound distraction volume in decibels and visual distraction frequency in flashes per minute as explanatory variables, and test score as the outcome. Changes in both distraction types cause a statistically significant reduction in test scores. For each 10 db increase in noise level, the test score drops by 2.00 points (95% CI=[1.14,

2.86]) at any fixed visual distraction level. For each per minute increase in the visual distraction blink rate, the test score drops by 1.12 points (95%CI=[0.70,1.54]) at any fixed auditory distraction value. From the adjusted R^2, we see that about 53% of the variability in test scores is accounted for by taking the values of the two distractions into account.

The validity of these conclusions can be confirmed with the following assumption checks:

- Checking Normality using a quantile-normal plot of the residuals
- Checking equal variance and linearity using a residual vs. fit plot. It is also a good idea to further confirm linearity for each explanatory variable with plots of each explanatory variable vs. the residuals.
- Qualitatively confirming that the fixed-x assumption is met by realizing that the values of the distractions are precisely set by the experimenter.
- Qualitatively confirming that the independent errors assumption is met by realizing that separate subjects are used for each test, and the subjects were not allowed to collaborate.

An additional assumption implicit in the regression model discussed here is what is called the "additivity" assumption. The "additivity" assumption says that the effect (on the outcome) of a one-unit rise of one explanatory variable is the same at *every* fixed value of the other variable (and vice versa). When this assumption is violated, there is said to be an *interaction* among the explanatory variables. We will discuss interactions and how to test for them in the next section.

There is one piece of the `lm()` regression output that we haven't discussed yet. Look at the last line of the output: We see an F-statistic of 22.87 and a small p-value. This p-value is for the null hypothesis that *all* of the slope parameters, but not the intercept parameter, are equal to zero. So for this experiment, we reject $H_0 : \beta_{db} = \beta_f = 0$.

Multiple regression is a direct extension of simple regression to multiple explanatory variables. Each new explanatory variable adds one term to the structural model.

Setting	x_{db}	x_f	E(Y)	difference from baseline
1	2	4	100-5(2)-3(4)=78	
2	3	4	100-5(3)-3(4)=73	-5
3	2	6	100-5(2)-3(6)=72	-6
4	3	6	100-5(3)-3(6)=67	-11

Table 10.1: Demonstration of the additivity of $E(Y) = 100 - 5x_{db} - 3x_f$.

10.2 Interaction

Interaction is a major concept in statistics that applies whenever there are two or more explanatory variables. Interactions exist between two or more explanatory variables in their effect on an outcome; *interaction is* **never** *between an explanatory variable and an outcome, or between levels of a single explanatory variable.* The term interaction applies to both quantitative and categorical explanatory variables. The definition of interaction is that the effect of a change in the level or value of one explanatory variable on the mean outcome *depends* on the level or value of another explanatory variable. Therefore interaction relates to the structural part of a statistical model.

In the absence of interaction, the effect *on the outcome* of any specific *change* in one explanatory variable, e.g., a one unit rise in a quantitative variable or a change from, e.g., level 3 to level 1 of a categorical variable, does not depend on the level or value of the other explanatory variable(s), as long as they are held constant. This also tells us that, e.g., the effect on the outcome of changing from level 1 of explanatory variable 1 and level 3 of explanatory variable 2 to level 4 of explanatory variable 1 and level 2 of explanatory variable 2 is equal to the sum of the effects on the outcome of only changing variable 1 from level 1 to 4 plus the effect of only changing variable 2 from level 3 to 1. For this reason the lack of an interaction is called **additivity**. The distraction example of the previous section is an example of a multiple regression model for which additivity holds (and therefore there is no interaction of the two explanatory variables in their effects on the outcome).

A mathematic example may make this more clear. Consider a model with quantitative explanatory variables "decibels of distracting sound" and "frequency

of light flashing", represented by x_{db} and x_f respectively. Imagine that the parameters are actually known, so that we can use numbers instead of symbols for this example. The structural model demonstrated here is $E(Y) = 100 - 5x_{db} - 3x_f$. Sample calculations are shown in Table 10.1. Line 1 shows the arbitrary starting values $x_{db} = 2,\ x_f = 4$. The mean outcome is 78, which we can call the "baseline" for these calculations. If we leave the light level the same and change the sound to 3 (setting 2), the mean outcome drops by 5. If we return to $x_{db} = 2$, but change x_f to 6 (setting 3), then the mean outcome drops by 6. Because this is a non-interactive, i.e., additive, model we expect that the effect of simultaneously changing x_{db} from 2 to 3 and x_f from 4 to 6 will be a drop of 5+6=11. As shown for setting 4, this is indeed so. This would not be true in a model with interaction.

Note that the component explanatory variables of an interaction and the lines containing these individual explanatory variables in the coefficient table of the multiple regression output, are referred to as **main effects**. In the presence of an interaction, when the signs of the coefficient estimates of the main effects are the same, we use the term **synergy** if the interaction coefficient has the same sign. This indicates a "super-additive" effect, where the whole is more than the sum of the parts. If the interaction coefficient has opposite sign to the main effects, we use the term **antagonism** to indicate a "sub-additive" effect, where simultaneous changes in both explanatory variables has less effect than the sum of the individual effects.

The key to understanding the concept of interaction, how to put it into a structural model, and how to interpret it, is to understand the construction of one or more new interaction variables from the existing explanatory variables. An interaction variable is created as the product of two (or more) explanatory variables. That is why some programs and textbooks use the notation "A*B" to refer to the interaction of explanatory variables A and B.

The creation, use, and interpretation of interaction variables for two quantitative explanatory variables is discussed next. The extension to more than two variables is analogous but more complex. Interactions that include a categorical variable are discussed in the next section.

Consider an example of an experiment testing the effects of the dose of a drug (in mg) on the induction of lethargy in rats as measured by number of minutes that the rat spends resting or sleeping in a 4 hour period. Rats of different ages are used and age (in months) is used as a control variable. Data for this (fake) experiment are found in lethargy.dat.

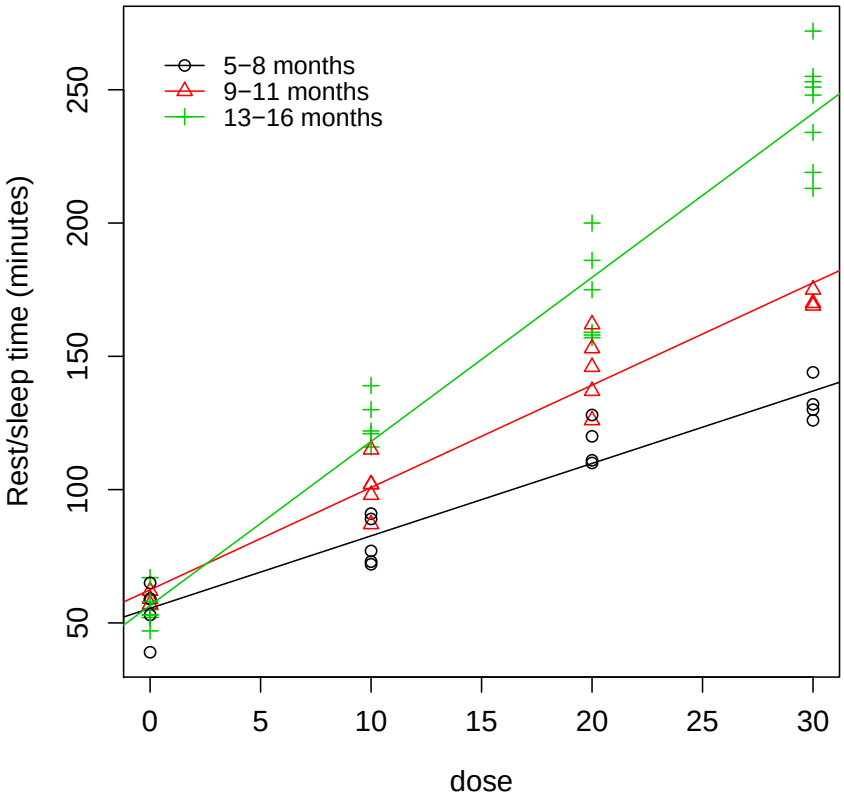


Figure 10.2: EDA for the lethargy example.

Figure 10.2 shows some EDA. Here the control variable, age, is again categorized, and regression fit lines are added to the plot for each level of the age categories. (Further analysis uses the complete, quantitative version of the age variable.) What you should see here is that the slope appears to change as the control variable changes. It looks like more drug causes more lethargy, and older rats are more lethargic at any dose. But what suggests interaction here is that the three fit lines are *not* parallel, so we get the (correct) impression that the effect of any dose increase on lethargy is *stronger* in old rats than in young rats.

In multiple regression with interaction we add the new (product) interaction variable(s) as additional explanatory variables. For the case with two explanatory variables, this becomes

$$E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}(x_1 \cdot x_2)$$

where β_{12} is the single parameter that represents the interaction effect and $(x_1 \cdot x_2)$ can either be thought of as the single new interaction variable (data column) or as the product of the two individual explanatory variables.

Let's examine what the multiple regression with interaction model is claiming, i.e., in what situations it might be plausible. By examining the equation for the structural model you can see that the effect of a one unit change in either explanatory variable *depends* on the value of the other explanatory variable.

We can understand the details by taking the approach of writing down the model equation and then making it reflect specific cases. Here, we use more meaningful variable names and parameter subscripts. Specifically, $\beta_{\mathrm{d*a}}$ is the symbol for the single interaction parameter. Let's consider a case where we fix age to be equal to some value a:

$$\begin{aligned} E(Y|\text{dose}, \text{age}) &= \beta_0 + \beta_{\text{dose}}\text{dose} + \beta_{\text{age}}\text{age} + \beta_{\mathrm{d*a}}\text{dose} \cdot \text{age} \\ E(Y|\text{dose}, \text{age} = a) &= \beta_0 + \beta_{\text{dose}}\text{dose} + a\beta_{\text{age}} + a\beta_{\mathrm{d*a}} \cdot \text{dose} \\ E(Y|\text{dose}, \text{age} = a) &= (\beta_0 + a\beta_{\text{age}}) + (\beta_{\text{dose}} + a\beta_{\mathrm{d*a}})\text{dose} \end{aligned}$$

The first case is the linear regression model with an interaction term, and the second and third lines are two ways to write the model when we set age $= a$. Because the βs are fixed (unknown) constants, this equation tells us that when age is fixed at some particular number, a, the relationship between $E(Y)$ and dose is a straight line with intercept equal to the number $\beta_0 + a\beta_{\text{age}}$ and slope

equal to the number $\beta_{\text{dose}} + a\beta_{\text{d*a}}$. The key feature of the interaction is the fact that the slope with respect to dose *is different* for each value of a, i.e., for each age. A similar equation can be written for fixed dose and varying age. When we considered simple linear regression in Chapter 9, there was just a single intercept parameter and a single slope parameter. As we can see, for linear regression models with interactions, the intercept and slope for a particular subject depends on that subject's explanatory variables.

Explaining the meaning of the interaction parameter in a multiple regression with continuous explanatory variables is difficult. Luckily, as we will see below, it is much easier in the simplest version of ANCOVA, where there is one categorical and one continuous explanatory variable.

Here is the code to run a linear regression of the outcome (`resttime`) on the two explanatory variables (`dose` and `age`) as well as their interaction:

```
> lethargy.linReg = lm(resttime ~ dose*age, data = lethargy)
> summary(lethargy.linReg)

Call:
lm(formula = resttime ~ dose * age, data = lethargy)

Residuals:
     Min       1Q   Median       3Q      Max
-17.8286  -5.2131   0.4527   5.4958  15.7428

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.99513    5.49304   8.919 2.43e-12 ***
dose         0.39759    0.28196   1.410    0.164
age          0.75875    0.50010   1.517    0.135
dose:age     0.39556    0.02493  15.865  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 7.883 on 56 degrees of freedom
Multiple R-squared:  0.9846,  Adjusted R-squared:  0.9838
F-statistic:  1192 on 3 and 56 DF,  p-value: < 2.2e-16
```

Note that there are four numbers in the "Estimate" column: These correspond to our estimates for β_0, β_{dose}, β_{age}, and $\beta_{\text{d*a}}$. Furthermore, note that we just had to include `dose*age` on the right-hand side of the formula within `lm()`, which automatically adds the main effects *and* interaction for these variables. Alternatively,

we could have written `dose + age + dose:age` or `dose + age + dose*age`.

Here is an interpretation of the analysis of this experiment. A multiple regression analysis including interaction was performed using drug dose in mg and rat age in months as explanatory variables, and minutes resting or sleeping during a 4 hour test period as the outcome. There is a significant interaction (t=15.86, $p< 2 * 10^{-16}$) between dose and age in their effect on lethargy. (Therefore changes in either or both explanatory variables cause changes in the lethargy outcome.) Because the coefficient estimate for the interaction is of the same sign as the signs of the individual coefficients, it is easy to give a general idea about the effects of the explanatory variables on the outcome: Increases in both dose and age are associated with (cause, for dose) an increase in lethargy, and the effects are "super-additive" or "synergistic" in the sense that the effect of simultaneous fixed increases in both variables is more than the sum of the effects of the same increases made separately for each explanatory variable. We can also see that about 98% of the variability in resting/sleeping time is accounted for by taking the values of dose and age into account.

As usual, the validity of these conclusions needs to be confirmed with checks for the five linear model assumptions discussed earlier.

Note that the p-value for the interaction line of the regression results tells us that the interaction is an important part of the model. Also note that the component explanatory variables of the interaction (main effects) are almost always included in a model if the interaction is included. In the presence of a significant interaction, both explanatory variables affect each other in terms of their own effects on the outcome. Therefore, if an interaction effect has a significant p-value, it is often misleading to interpret the main effects by themselves (because, by definition of interactions, the effect of one variable depends on another variable). On the other hand, if the interaction is not significant, generally the appropriate next step is to perform a new multiple regression analysis excluding the interaction term, i.e., run an additive model.

To write prediction equations with numbers (based on estimates of the parameters) instead of symbols, we use $\hat{Y}$ on the left side to indicate a "best estimate" of $E(Y)$ (not it's true value!), which depends on the β values. For this example, the prediction equation for resting/sleeping minutes for rats of age 12 months at any dose is

$$\hat{Y} = 49.00 + 0.398(\text{dose}) + 0.76(12) + 0.396(\text{dose} \cdot 12) = 58.12 + 5.15(\text{dose})$$

Interaction between two explanatory variables is present when the effect of one on the outcome depends on the value of the other. Interaction is implemented in multiple regression by including a new explanatory variable that is the product of two existing explanatory variables. The model can be better understood by writing equations for the relationship between one explanatory variable and the outcome for some fixed values of the other explanatory variable.

10.3 Categorical variables in multiple regression

To use a categorical variable with k levels in multiple regression we must re-code the data column as $k-1$ new columns, each with only two different codes (most commonly we use 0 and 1). Variables that only take on the values 0 or 1 are called **indicator** or **dummy** variables. They should be considered as quantitative variables and should be named corresponding to their "1" level (e.g., if old = 1 and young = 0, the variable should be named "old", not "young").

An indicator variable is coded 0 for any case that does not match the variable name and 1 for any case that does match the variable name.

One level of the original categorical variable is designated the "baseline". If there is a control or placebo, the baseline is usually set to that level. The baseline level does not have a corresponding variable in the new coding; instead subjects with that level of the categorical variable have 0's in all of the new variables. Each new variable is coded to have a "1" for the level of the categorical variable that matches its name and a zero otherwise.

It is very important to realize that when new variables like these are constructed, they *replace* the original categorical variable when entering variables into a multiple regression analysis, so the original variables are no longer used at all. (The originals should not be erased, because they are useful for EDA, and because you want to be able to verify correct coding of the indicator variables.) Furthermore, note that the choice of the baseline variable only affects the convenience of

presentation of results and does not affect the interpretation of the model or the prediction of future values.

As an example consider a data set with a categorical variable for favorite condiment. The categories are ketchup, mustard, hot sauce, and other. If we arbitrarily choose ketchup as the baseline category we get a coding like this:

	Indicator Variable		
Level	mustard	hot sauce	other
ketchup	0	0	0
mustard	1	0	0
hot sauce	0	1	0
other	0	0	1

Note that this indicates, e.g., that every subject that likes mustard best has a 1 for their "mustard" variable, and zeros for their "hot sauce" and "other" variables.

As shown in the next section, this coding flexibly allows a model to have no restrictions on the relationships of population means when comparing levels of the categorical variable. It is important to understand that if we "accidentally" use a categorical variable, usually with values 1 through k, in a multiple regression, then we are inappropriately forcing the mean outcome to be ordered according to the levels of a nominal variable, *and* we are forcing these means to be equally spaced. Both of these problems are fixed by using indicator variable recoding.

To code the interaction between a categorical variable and a quantitative variable, we need to create another $k-1$ new variables. These variables are the products of the $k-1$ indicator variable(s) and the quantitative variable. Each of the resulting new data columns has zeros for all rows corresponding to all levels of the categorical variable except one (the one included in the name of the interaction variable), and has the value of the quantitative variable for the rows corresponding to the named level.

Generally a model includes all or none of a set of indicator variables that correspond with a single categorical variable. The same goes for the $k-1$ interaction variables corresponding to a given categorical variable and quantitative explanatory variable.

Categorical explanatory variables can be incorporated into multiple regression models by substituting $k-1$ indicator variables for any k-level categorical variable. For an interaction between a categorical and a quantitative variable $k-1$ product variables should be created.

10.4 ANCOVA

The term ANCOVA (analysis of covariance) is used somewhat differently by different analysts and computer programs, but the most common meaning, and the one we will use here, is for a multiple regression analysis in which there is at least one quantitative and one categorical explanatory variable. Usually the categorical variable is a treatment of primary interest, and the quantitative variable is a "control variable" of secondary interest, which is included to improve power (without sacrificing generalizability).

Consider a particular quantitative outcome and two or more treatments that we are comparing for their effects on the outcome. If we think one or more explanatory variables are associated with the outcome, then we can include it in the regression to increase power and improve precision. The intuition behind this is that, by "regressing out" the variation that's due to the explanatory variable(s), we can better identify the variation that's due to treatment. Ignoring the other explanatory variables and performing a simple ANOVA increases σ^2 and makes it harder to detect any real differences in treatment effects.

ANCOVA extends the idea of blocking to continuous explanatory variables, as long as a simple mathematical relationship (usually linear) holds between the control variable and the outcome.

10.4.1 ANCOVA with no interaction

An example will make this more concrete. The data in mathtest.dat come from a (fake) experiment testing the effects of two computer aided instruction (CAI) programs on performance on a math test. The programs are labeled A and B, where A is the control, older program, and B is suspected to be an improved

version. We know that performance depends on general mathematical ability so the students math SAT is used as a control variable.

First let's look at t-test results, ignoring the SAT score. EDA shows a slightly higher mean math test score, but lower median for program B. A t-test shows no significant difference with t=0.786, p=0.435. It is worth noting that the CI for the mean difference between programs is [-5.36, 12.30], so we are 95% confident that the effect of program B relative to the old program A is somewhere between lowering the mean score by 5 points and raising it by 12 points. The estimate of σ (square root of MS_{within} from an ANOVA) is 17.1 test points.

EDA showing the relationship between math SAT (MSAT) and test score separately for each program is shown in Figure 10.3. The steepness of the lines and the fact that the variation in y at any x is smaller than the overall variation in y for either program demonstrates the value of using MSAT as a control variable. The lines are roughly parallel, suggesting that an additive, no-interaction model is appropriate. The line for program B is higher than for program A, suggesting its superiority.

First it is a good idea to run an ANCOVA model with interaction to verify that the fit lines are parallel (the slopes are not statistically significantly different). This is done by running a multiple regression model that includes the explanatory variables ProgB, MSAT, and the interaction between them (i.e, the product variable). Note that we do not need to create a new set of indicator variables because there are only two levels of program, and the existing variable is already an indicator variable for program B. The interaction p-value is 0.375 (not shown), so there is no evidence of a significant interaction (different slopes).

However, it is important to note that the `progB` variable is coded as a 0 or 1, and thus by default `R` will assume this is a quantitative variable. However, this is conceptually incorrect: It is a categorical variable! In terms of the linear regression results, this won't make a difference for a binary categorical variable, but this could make a huge difference for variables with more than two categories. To make sure that the `progB` variable is treated as a categorical variable, we will use the `factor()` function:

```
#read in the data
> mathtest = read.table("mathtest.txt", header = TRUE)
#the progB variable is coded as a number:
> mathtest$progB
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

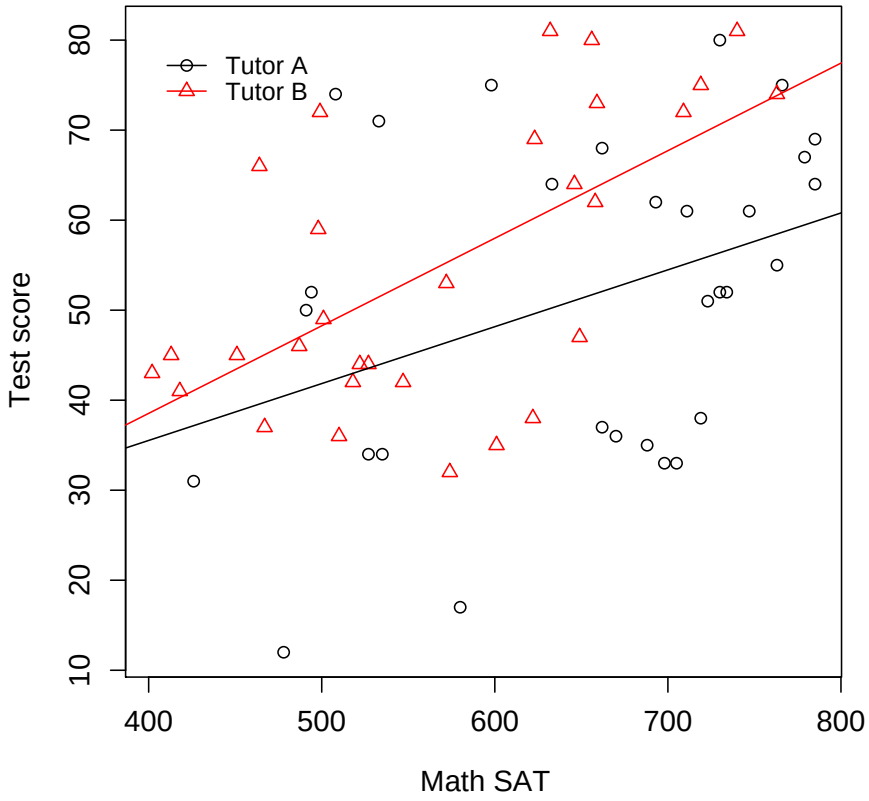


Figure 10.3: EDA for the math test / CAI example.

```
> class(mathtest$progB)
"integer"
#instead code it as a factor:
> mathtest$progB = factor(mathtest$progB)
#now we see it is categorical:
> mathtest$progB
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Levels: 0 1
> class(mathtest$progB)
[1] "factor"
```

It is often the case that categorical variables are coded with numbers (e.g., 1, 2, 3 may represent "low," "medium," and "high"). By default, R will treat such variables as quantitative, which is conceptually incorrect. To make sure that these variables are encoded as categorical in R, it is very important that you use the `factor()` function to redefine these variables.

Now let's run the additive linear regression model and look at the output:

```
> mathtest.linReg = lm(score~progB + MSAT, data = mathtest)
> summary(mathtest.linReg)

Call:
lm(formula = score ~ progB + MSAT, data = mathtest)

Residuals:
    Min      1Q  Median      3Q     Max
-28.740  -9.154  -0.294  10.042  33.971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26955   12.69771  -0.021 0.983138
progB1      10.09314    4.20585   2.400 0.019695 *
MSAT         0.07933    0.01902   4.171 0.000104 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 15.08 on 57 degrees of freedom
Multiple R-squared:  0.2419,  Adjusted R-squared:  0.2153
```

```
F-statistic: 9.095 on 2 and 57 DF,  p-value: 0.0003731

> confint(mathtest.linReg)
                    2.5 %     97.5 %
(Intercept) -25.69628269 25.1571882
progB1        1.67106164 18.5152169
MSAT          0.04124367  0.1174109
```

Of primary interest is the estimate of the benefit of using program B over program A, which is 10.09 points (t=2.40, p=0.020) with a 95% confidence interval of 1.67 to 18.52 points. Somewhat surprisingly, the estimate of σ, which now refers to the standard deviation of test score for any combination of program *and* MSAT is only slightly reduced from 17.1 to 15.1 points. The ANCOVA model explains 22% of the variabilty in test scores (adjusted $R^2 = 0.215$), so there are probably some other important variables "out there" to be discovered.

Of minor interest is the fact that the "control" variable, math SAT score, is highly statistically significant (t=4.17, p = 0.0001). We estimate that every 10 additional math SAT points is associated with a 0.4 to 1.2 point rise in test score. Note that this is just an association and *not* a causal conclusion. We can only make a causal conclusion about Program A vs. Program B, because we randomized the programs. We did not randomize SAT test scores.

In conclusion, program B improves test scores by a few points on average for students of all ability levels (as determined by MSAT scores).

This is a typical ANOVA story where the power to detect the effects of a treatment is improved by including one or more control and/or blocking variables, which are chosen by subject matter experts based on prior knowledge. In this case the effect of program B compared to control program A was detectable using MSAT in an ANCOVA, but not when ignoring it in the t-test.

For ANCOVA, it is extremely helpful to write out the full structural model equation, and then simplify it based on different levels of a categorical explanatory variable. In this case, the full structural model equation and simplified structural model equations are:

$$\begin{aligned} E(Y|\text{ProgB}, MSAT) &= \beta_0 + \beta_{\text{ProgB}}\text{ProgB} + \beta_{\text{MSAT}}\text{MSAT} \\ \text{Program A: } E(Y|\text{ProgB} = 0, MSAT) &= \beta_0 + \beta_{\text{MSAT}}\text{MSAT} \\ \text{Program B: } E(Y|\text{ProgB} = 1, MSAT) &= (\beta_0 + \beta_{\text{ProgB}}) + \beta_{\text{MSAT}}\text{MSAT} \end{aligned}$$

To be explicit, β_{MSAT} is the slope parameter for MSAT and β_{ProgB} is the parameter for the indicator variable ProgB. By comparing the second and third equations, we can see that this parameter determines a difference in intercept for Program A vs. Program B.

For the analysis of the data shown here, the predictions are:

$$\begin{aligned} \hat{Y}(\text{ProgB}, MSAT) &= -0.27 + 10.09\text{ProgB} + 0.08\text{MSAT} \\ \text{Program A: } \hat{Y}(\text{ProgB} = 0, MSAT) &= -0.27 + 0.08\text{MSAT} \\ \text{Program B: } \hat{Y}(\text{ProgB} = 1, MSAT) &= 9.82 + 0.08\text{MSAT} \end{aligned}$$

We can also see how the point estimate for the treatment effect is obtained: It is the third equation minus the second equation. Note that although (for this example) the intercept is a meaningless extrapolation to an impossible MSAT score of 0, we still need to use it in the prediction equation. Also note that in this no-interaction model, the simplified equations for the different treatment levels have different intercepts, but the same slope.

ANCOVA is used in the case of a quantitative outcome with both a categorical and a quantitative explanatory variable. The main use is for testing a treatment effect while using a quantitative control variable to gain power.

10.4.2 ANCOVA with interaction

It is also possible that a significant interaction between a control variable and treatment will occur, or that the quantitative explanatory variable is a variable of primary interest that interacts with the categorical explanatory variable. A significant interaction between an explanatory variable and the treatment variable implies that the treatment somehow alters the relationship between the explanatory variable and the outcome.

Let's consider a dataset (available at Performance.dat) that involves three different treatments (A, B, and C), a skill variable S (which can range from 0 to 100), and a quantitative outcome, performance, which can range from 0 to 200. EDA showing the relationship between skill and performance separately for each

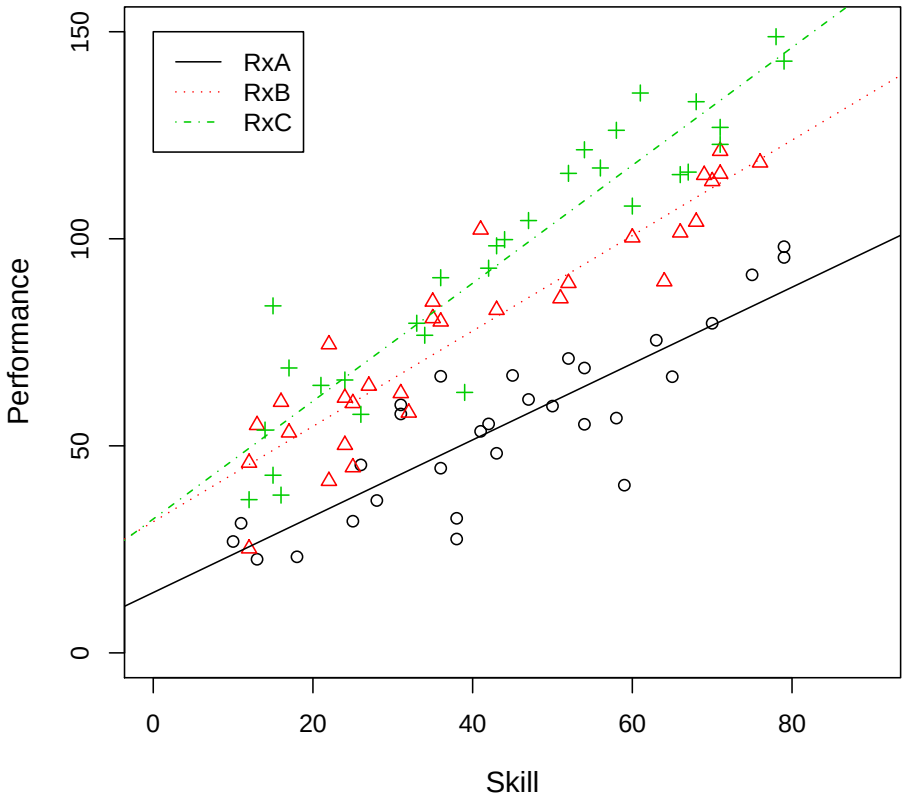


Figure 10.4: EDA for the performance ANCOVA example.

treatment is shown in Figure 10.4. The treatment variable, called Rx, was recoded to $k - 1 = 2$ indicator variables, which we will call RxB and RxC, with level A as the baseline. In other words, if a subject was assigned to treatment A, then RxB = RxC = 0; if a subject was assigned to treatment B, then RxB = 1 and RxC = 0; and if a subject was assigned to treatment C, then RxB = 0 and RxC = 1.

Meanwhile, here are the results from the corresponding multiple regression:

```
#performance example
> performance = read.table("Performance.txt", header = TRUE)
#recode as factors
> performance$RxB = factor(performance$RxB)
> performance$RxC = factor(performance$RxC)
#linear regression with interactions
```

```
> performance.linReg.int = lm(y ~ RxB*S + RxC*S, data =
    performance)
> summary(performance.linReg.int)

Call:
lm(formula = y ~ RxB * S + RxC * S, data = performance)

Residuals:
    Min      1Q  Median      3Q     Max
-28.444  -6.421   1.343   6.871  30.116

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5646     5.0023   2.912 0.004604 **
RxB1         17.0960     6.6305   2.578 0.011670 *
S             0.9217     0.1044   8.824 1.34e-13 ***
RxC1         17.7669     6.8281   2.602 0.010952 *
RxB1:S        0.2303     0.1418   1.624 0.108019
S:RxC1        0.5018     0.1415   3.547 0.000641 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 10.95 on 84 degrees of freedom
Multiple R-squared:  0.8849,  Adjusted R-squared:  0.878
F-statistic: 129.1 on 5 and 84 DF,  p-value: < 2.2e-16
```

Before we interpret the results, we should first write down the structural model we are positing with this model (which you should be able to determine from the above output and description of the data). Using mnemonic labels for the parameters, the structural model that goes with this analysis is:

$$E(Y|\text{Rx},\text{S}) = \beta_0 + \beta_{\text{RxB}}\text{RxB} + \beta_{\text{RxC}}\text{RxC} + \beta_{\text{S}}\text{S} + \beta_{\text{RxB*S}}\text{RxB}\cdot\text{S} + \beta_{\text{RxC*S}}\text{RxC}\cdot\text{S}$$

Using the above output, the parameter estimates are $\hat{\beta}_0 = 14.56$, $\hat{\beta}_{\text{RxB}} = 17.10$, $\hat{\beta}_{\text{RxC}} = 17.77$, $\hat{\beta}_{\text{S}} = 0.92$, $\hat{\beta}_{\text{RxB*S}} = 0.23$, and $\hat{\beta}_{\text{RxC*S}} = 0.50$.

To understand this complicated model, let us use the **very important** technique of simplying the structural model based on different values of the categorical explanatory variable (in this case, Rx). Remember that Rx takes on three values (A, B, or C), which in turn correspond to different values (0 or 1) for RxB and

RxC. Here's what the structural model looks like for each value of Rx:

$$\begin{aligned}
\text{RxA: } E(Y|\text{Rx=A, S}) &= \beta_0 + \beta_{\text{S}}\text{S} \\
\text{RxB: } E(Y|\text{Rx=B, S}) &= (\beta_0 + \beta_{\text{RxB}}) + (\beta_{\text{S}} + \beta_{\text{RxB*S}})\text{S} \\
\text{RxC: } E(Y|\text{Rx=C, S}) &= (\beta_0 + \beta_{\text{RxC}}) + (\beta_{\text{S}} + \beta_{\text{RxC*S}})\text{S}
\end{aligned}$$

These simplified equations are created by substituting in 0's and 1's for RxB and RxC (but not into parameter subscripts), and then fully simplifying the equations.

By examining these three equations we can fully understand the model. From the first equation, we see that β_0 is the mean outcome for subjects given treatment A and who have S=0. (Sometimes it is worthwhile to "center" a variable like S by subtracting its mean from every value; then the intercept will refer to the mean of S, which is never an extrapolation.)

Again using the first equation, we see that β_{S} is the slope of Y vs. S for subjects given treatment A.

From the second equation, the intercept for treatment B can be seen to be $(\beta_0 + \beta_{\text{RxB}})$, and this is the mean outcome when S=0 for subjects given treatment B. Therefore the interpretation of β_{RxB} is the *difference* in mean outcome when S=0 when comparing treatment B to treatment A (a positive parameter value would indicate a higher outcome for B than A, and a negative parameter value would indicate a lower outcome). Similarly, the interpretation of $\beta_{\text{RxB*S}}$ is the *change* in slope from treatment A to treatment B, where a positive $\beta_{\text{RxB*S}}$ means that the B slope is steeper than the A slope and a negative $\beta_{\text{RxB*S}}$ means that the B slope is less steep than the A slope.

The null hypotheses then have these specific meanings. $\beta_{\text{RxB}} = 0$ is a test of whether the intercepts differ for treatments A and B. $\beta_{\text{RxC}} = 0$ is a test of whether the intercepts differ for treatments A and C. $\beta_{\text{RxB*S}} = 0$ is a test of whether the slopes differ for treatments A and B. And $\beta_{\text{RxC*S}} = 0$ is a test of whether the slopes differ for treatments A and C.

Here is a full interpretation of the performance ANCOVA example. Analysis of the data from the performance dataset shows that treatment and skill interact in their effects on performance. Because skill levels of zero are a gross extrapolation, we should not interpret the intercepts.

Just for educational purposes, if skill=0 were a meaningful, observed state, then we would say all of the things in this paragraph. The estimated mean performance

for subjects with zero skill given treatment A is 14.56 points (however, a 95% CI would be more informative). If it were scientifically interesting, we could also say that this value of 14.56 is statistically different from zero (t=2.91, df=84, p=0.005). The intercepts for treatments B and C (mean performances when skill level is zero) are both statistically significantly different from the intercept for treatment A (t=2.58,2.60, respectively, df=84, and p=0.012, 0.011, respectively). The estimates are 17.10 and 17.77 points higher for B and C, respectively, compared to A.

We can also say that there is a statistically significant effect of skill on performance for subjects given treatment A (t=8.82, p= $1.34 * 10^{-13}$). We estimate that the mean performance increases by 9.2 points for each 10-point increase in skill. The slope of performance vs. skill for treatment B is not statistically significantly different from that of treatment A (t=1.15, p=0.108). The slope of performance vs. skill for treatment C is statistically significantly different for that of treatment A (t=3.55, p=0.001). The best estimate is that the slope for subjects given treatment C is 0.50 higher than for treatment A (i.e., the mean change in performance for a 1 unit increase in skill is 0.50 points *more for treatment C than for treatment A*). We can also say that the best estimate for the slope of the effect of skill on performance for treatment C is 0.92+0.50=1.42.

In summary, increasing skill has a positive effect on performance for treatment A (of about 9 points per 10 point rise in skill level). Treatment B has a higher projected intercept than treatment A, and the effect of skill on subjects given treatment B is not statistically different from the effect on those given treatment A. Treatment C has a higher projected intercept than treatment A, and the effect of skill on subjects given treatment C is statistically different from the effect on those given treatment A (by about 5 additional points per 10 unit rise in skill).

One important thing to note is that for this model, we can only make comparisons for A vs. B and A vs. C but not B vs C. (Look back at the previous paragraph; we only made conclusions about A vs. B and A vs. C comparisons.) If we wanted to study B vs. C specifically, we would have to change the baseline for the Rx variable. For example, we could change the baseline to B, and thus the coefficient for RxC would correspond to a B vs. C comparison, rather than an A vs. C comparison.

Meanwhile, additional testing using methods we have not learned (but will in Chapter 12!) can be performed to show that performance is better for treatments B and C than treatment A at all observed levels of skill.

If an ANCOVA has a significant interaction between the categorical and quantitative explanatory variables, then the slope of the equation relating the quantitative variable to the outcome differs for different levels of the categorical variable. The p-values for indicator variables test intercept differences from the baseline treatment, while the interaction p-values test slope differences from the baseline treatment.

Finally, we should note that in the above model, the interaction between S and RxC is significant, but the interaction between S and RxB is not. In general, one may ask: Is there *any* interaction effect? In other words, maybe we would like to test the null hypothesis $H_0 : \beta_{\text{RxB*S}} = \beta_{\text{RxC*S}} = 0$; this is slightly different from the above p-values for these coefficients, which test the null hypotheses $H_0 : \beta_{\text{RxB*S}} = 0$ and $H_0 : \beta_{\text{RxC*S}} = 0$ simultaneously. Testing the null hypothesis $H_0 : \beta_{\text{RxB*S}} = \beta_{\text{RxC*S}} = 0$ is equivalent to testing: Do we need a model with interactions, or is an additive model adequate? Thus, in `R`, we will make an additive model, and then use the `anova()` function:

```
#linear regression without interactions
> performance.linReg.noInt = lm(y ~ RxB + RxC + S, data =
    performance)

#compare the additive and interactive models:
> anova(performance.linReg.noInt, performance.linReg.int)
Analysis of Variance Table

Model 1: y ~ RxB + RxC + S
Model 2: y ~ RxB * S + RxC * S
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     86 11600
2     84 10076  2    1524.7 6.3559 0.002689 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
```

We haven't seen the `anova()` function before (remember that we use `aov()` to run one-way ANOVA, not `anova()`!) The `anova()` function is used to compare two models where one is *nested* in another. In this example, the additive model ("Model 1" in the above output) is nested in the interactive model ("Model 2") because it is equivalent to the interactive model where $\beta_{\text{RxB*S}} = \beta_{\text{RxC*S}} = 0$. In this sense,

the above p-value is indeed for the null hypothesis $H_0 : \beta_{\text{RxB*S}} = \beta_{\text{RxC*S}} = 0$; we see that it is equal to 0.003, and we therefore reject this null hypothesis. Also note that the degrees of freedom for this test is equal to two, because there are two parameters that we are testing for in our null hypothesis ($\beta_{\text{RxB*S}}$ and $\beta_{\text{RxC*S}}$).

10.5 Do it in R

To perform multiple linear regression (as well as ANCOVA) in `R`, you use many of the same functions we used to do simple linear regression in Chapter 9. You still use the `lm()` function to run a regression model; you still use the `summary()` function to get output from a regression model; and you use `confint()` to get confidence interals from a regression model. See Section 9.10 for more functions used for regression in `R`.

In terms of coding syntax, the main novelty in this chapter that we didn't have to worry about in Chapter 9 is how to include multiple explanatory variables into a single regression model. Here is the notation used within `lm()` for using multiple explanatory variables in a regression model:

- `+`: Adds a new explanatory variable to the regression model (main effects *or* interactions). For example, `lm(y~a)` is a regression of just `y` on `a`, while `lm(y~a + b)` is a regression of `y` on `a` and `b` (only their main effects, not their interaction!)

- `*`: This is the most common notation used to add interactions to a regression; it also adds main effects. For example, `lm(y~a*b)` is a regression of `y` on `a` and `b` *as well as their interaction*. When including an interaction effect, you should almost always also include main effects.

- `:`: Adds an interaction *only* to a regression (i.e., it excludes main effects). For example, `lm(y~a:b)` is a regression of `y` on the interaction of `a` and `b` *but not their main effects*. Because it's very uncommon (and not recommended) to include interaction effects without main effects, you rarely need to use the `:` notation.

One last function we've discussed in this chapter is `anova()`; this function is used to compare two regression models, where one model is nested within another.

Specifically, you write `anova(model1, model2)`, where `model1` and `model2` are objects defined with `lm()`. This function is used when you want to test the null hypothesis that *multiple* coefficients in a regression model *simultaneously* equal zero. For example, let's say we run the following:

```
model1 = lm(y ~ a)
model2 = lm(y ~ a + b + c)
```

where `a`, `b`, and `c` are explanatory variables, and `y` is an outcome variable. The structural model equation for `model1` is

$$E[Y|a] = \beta_0 + \beta_A \cdot a$$

Meanwhile, the structural model equation for `model2` is

$$E[Y|a] = \beta_0 + \beta_A \cdot a + \beta_B \cdot b + \beta_C \cdot c$$

Thus, if we looked at `summary(model2)`, we would be able to test the null hypotheses $H_0 : \beta_B = 0$ and $H_0 : \beta_C = 0$ separately; however, we would need to write `anova(model1, model2)` if we wanted to test $H_0 : \beta_B = \beta_C = 0$.

Chapter 11

Statistical Power

As we have discussed throughout this book, we conduct experiments to investigate if there is a significant difference in response among some set of treatments. Often, statistical tests are developed to control the Type 1 error rate, such that we protect ourselves from falsely concluding there is a treatment effect when really there isn't one. For example, the t-test is guaranteed to falsely reject the null hypothesis only 5% of the time when the null hypothesis is true and modeling assumptions hold. However, if there *is* a true treatment effect, we want to be pretty confident that our experiment will actually detect it. If an experiment doesn't have the *power* to detect a true treatment effect, then the experiment may not be worth running. Thus, it is very important to assess how *powerful* an experiment is before conducting it.

11.1 The concept

The power of an experiment that you are about to carry out quantifies the chance that you will correctly reject the null hypothesis if some alternative hypothesis is really true.

Consider analyzing a k-level one-factor experiment using ANOVA. We arbitrarily choose $\alpha = 0.05$ (or some other value) as our significance level. We reject the null hypothesis, $\mu_1 = \cdots = \mu_k$, if the F statistic is so large as to occur less than 5% of the time when the null hypothesis is true (and the assumptions are met).

This approach requires computation of the distribution of F values that we

would get if the model assumptions were true, the null hypothesis were true, and we would repeat the experiment many times, calculating a new F-value each time. This is called the null sampling distribution of the F-statistic (see Section 5.2.5).

For any sample size (n per group) and significance level (α) we can use the null sampling distribution to find a critical F-value "cutoff" *before* running the experiment, and know that we will reject H_0 if $F_{\text{observed}} \geq F_{\text{critical}}$, where F_{observed} denotes the value of the statistic we will observe after running the experiment, and F_{critical} denotes the critical value for that statistic. (Throughout this chapter, we will use n to denote the sample size *per treatment group*, not the overall sample size of the experiment.) If the assumptions are met (I won't keep repeating this), then—because of the way that F_{critical} is defined—if we were to run the experiment many times, then we should expect to falsely reject H_0 5% of the time, simply due to "bad luck" that a particular experiment's F-value happens to fall above F_{critical}. This is the so-called Type 1 error (see Section 7.4). We could lower α to reduce the chance that we will make such an error, but this will adversely affect the power of the experiment, as explained next.

Under each combination of n, underlying variance (σ^2) and some particular non-zero difference in population means (non-zero effect size) there is an *alternative* sampling distribution of F. Remember that the null sampling distribution is the distribution of F values we would expect from an experiment when there is no effect (i.e., when the null hypothesis is true). Thus, the alternative sampling distribution is the distribution of F values we would expect from an experiment when there *is* an effect (i.e., when the alternative hypothesis is true).

As an example, Figure 11.1 shows the null sampling distribution of the F-statistic for $k = 3$ treatments and $n = 50$ subjects per treatment (black, solid curve) plus the alternative sampling distribution of the F-statistic for two specific "alternative hypothesis scenarios" (red and green curves) labeled "n.c.p.=4" and "n.c.p.=9". For the moment, just recognize that n.c.p. stands for something called the "non-centrality parameter", that the n.c.p. for the null hypothesis is 0, and that larger n.c.p. values correspond to less "null-like" alternatives.

Regarding this specific example, we note that the numerator of the F-statistic ($\text{MS}_{\text{between}}$) will have $k-1 = 2$ df, and the denominator($\text{MS}_{\text{within}}$) will have $k(n-1) = 147$ df. Therefore the null sampling distribution

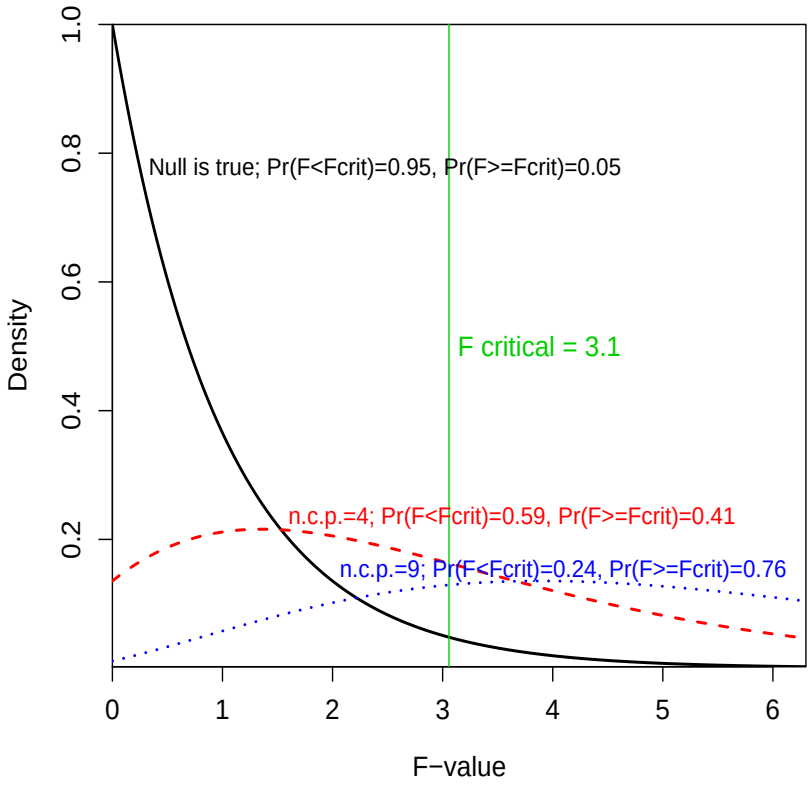


Figure 11.1: Null and alternative F sampling distributions.

for the F-statistic that the computer has drawn for us is the (central) F-distribution (see Section 3.9.7) with 2 and 147 df. This is equivalent to the F-distribution with 2 and 147 df and with n.c.p.=0. The two alternative null sampling distributions (curves) that the computer has drawn correspond to two specific alternative scenarios. The two alternative distributions are called non-central F-distributions. They also have 2 and 147 df, but in addition have "non-centrality parameter" values equal to 4 and 9 respectively.

The whole concept of power is explained in this figure. First focus on the black curve labeled "null is true". This curve is the null sampling distribution of F for *any* experiment with 1) three (categorical) levels of treatment; 2) a quantitative outcome for which the assumptions of Normality (at each level of treatment), equal variance and independent errors apply; 3) no difference in the three population means; and 4) a total of 150 subjects. The curve shows the values of the F-statistic that we are likely (high regions) or unlikely (low regions) to see if we repeat the experiment many times. The value of $F_{critical}$ of 3.1 separates (for k=3, n=50) the area under the null sampling distribution corresponding to the highest 5% of F-statistic values from the lowest 95% of F-statistic values. *Regardless of whether or not the null hypothesis is in fact true*, we will reject $H_0 : \mu_1 = \mu_2 = \mu_3$, i.e., we will *claim* that the null hypothesis is false, if our single observed F-statistic is greater than 3.1. Therefore it is built into our approach to statistical inference that among those experiments in which we study treatments that all have the same effect on the outcome, we will falsely reject the null hypothesis for about 5% of those experiments. It is very important to remember that when we analyze any experiment of this type, we will always reject H_0 if $F_{\text{observed}} > F_{\text{critical}}$, *regardless of whether or not the null hypothesis is in fact true* (because we don't know if it's true—that's why we're conducting an experiment in the first place!) F_{critical} was chosen such that we will reject the null 5% of the time when the null is true, but we will (hopefully) reject it more frequently when the null is false.

Now consider what happens if the null hypothesis is not true (but the model assumptions still hold). There are many ways that the null hypothesis can be false; for example, the true difference in population mean outcomes among treatments $(\mu_1, \ldots, \mu_k)$ could be quite large, or it could be quite small but still non-zero.

Thus, for any experiment, although there is only one null sampling distribution of F, there are (infinitely) many alternative sampling distributions of F. Two are shown in the figure. The information that needs to be specified to characterize a specific alternative sampling distribution is the spacing of the population means, the underlying variance at each fixed combination of explanatory variables (σ^2), and the number of subjects given each treatment (n). I call all of this information an "alternative scenario". The alternative scenario information can be reduced through a simple formula to a single number called the non-centrality parameter (n.c.p.), and this additional parameter value is all that the computer needs to draw the alternative sampling distribution for an ANOVA F-statistic. Note that n.c.p.=0 represents the null scenario.

The figure shows alternative sampling distributions for two alternative scenarios in red (dashed) and blue (dotted). The red curve represents the scenario where $\sigma = 10$ and the true means are 10.0, 12.0, and 14.0, which can be shown to correspond to n.c.p.=4. The blue curve represents the scenario where $\sigma = 10$ and the true means are 10.0, 13.0, and 16.0, which can be shown to correspond to n.c.p.=9. Obviously when the mean parameters are spaced 3 apart (blue) the scenario is more un-null-like than when they are spaced 2 apart (red). Note that the different alternative sampling distributions also depend on σ, but the null sampling distribution does not (instead, it only depends on the number of groups and the sample size of each group).

The alternative sampling distributions of F show how likely different F-statistic values are if the given alternative scenario is true. Looking at the red curve, we see that if you run many experiments when $\sigma = 10$ and $\mu_1 = 10.0, \mu_2 = 12.0$, and $\mu_3 = 14.0$, then about 59% of the time you will get $F < 3.1$ and $p > 0.05$, while the remaining 41% of the time you will get $F \geq 3.1$ and $p \leq 0.05$. This indicates that for the *one* experiment that you can really afford to do, you have a 59% chance of arriving at the incorrect conclusion that the population means are equal, and a 41% chance of arriving at the correct conclusion that the population means are not all the same. This is not a very good situation to be in, because there is a large chance of missing the interesting finding that the treatments have a real effect on the outcome.

We call the chance of incorrectly retaining the null hypothesis the Type 2 error rate, and we call the chance of correctly rejecting the null hypothesis for any given alternative the power. Power is always equal to 1 (or 100%) minus the Type 2 error rate. High power is good, and typically power greater than 80% is arbitrarily

considered "good enough".

In the figure, the alternative scenario with population mean spacing of 3.0 has fairly good power, 76%. If the true mean outcomes are 3.0 apart, and $\sigma = 10$ and there are 50 subjects in each of the three treatment groups, and the Normality, equal variance, and independent error assumptions are met, then any given experiment has a 76% chance of producing a p-value less than or equal to 0.05, which will result in the experimenter correctly concluding that the population means differ. But even if the experimenter does a terrific job of running this experiment, there is still a 24% chance of getting $p > 0.05$ and falsely concluding that the population means do *not* differ, thus making a Type 2 error.

Of course, describing power in terms of the F-statistic in ANOVA is only one example of a general concept. The same concept applies with minor modifications for the t-statistic that we learned about for both the independent samples t-test and the t-tests of the coefficients in regression and ANCOVA, as well as other statistics we haven't yet discussed. In the cases of the t-statistic, the modification relates to the fact that "un-null-like" corresponds to t-statistic values far from zero on either side (i.e., very positive or very negative treatment effects), rather than just larger values as for the F-statistic. Although the F-statistic will be used for the remainder of the power discussion, remember that the concepts apply to hypothesis testing in general. Any statistical test will use some kind of distribution (e.g., a t distribution, an F distribution) as a rejection rule, and this same kind of distribution is used in power calculations for that statistical test.

Finally, note that, for any given experiment and inference method (i.e., statistical test), the power to correctly reject a given alternative hypothesis lies somewhere between 5% and (almost) 100%. The reason power is bounded by 5% is because, even when the null hypothesis is true, we are guaranteed to reject it 5% of the time, and we should reject it more frequently when the null is false. The next section discusses ways to improve power.

For one-way ANOVA, the null sampling distribution of the F-statistic shows that when the null hypothesis is true, an experimenter has a 95% chance of obtaining a p-value greater than 0.05, in which case she will make the correct conclusion, but 5% of the time she will obtain $p \leq 0.05$ and make a Type 1 error. The various alternative sampling distributions of the F-statistic show that the chance of making a Type 2 error can range from 95% down to near zero, because the Type 2 error rate is equal to 1 - power, and power will always be between 5% and 100%.

11.2 Improving power

When designing an experiment (and before conducting it!), it is important to think about how different aspects of the experimental design can be changed to improve the power of the experiment. For this section we will focus on the two-group continuous outcome case because it is easier to demonstrate the effects of various factors on power in this simple setup. To make things concrete, assume that the experimental units are a random selection of news websites, the outcome is number of clicks (C) between 7 PM and 8 PM Eastern Standard Time for an associated online ad, and the two treatments are two fonts for the ads, say Palatino (P) vs. Verdana (V).

One way to think about this problem is in terms of the two confidence intervals for the population means. Anything that reduces the overlap of these confidence intervals will increase the power. The overlap is reduced by reducing the common variance (σ^2), increasing the number of subjects in each group (n), or by increasing the distance between the population means, $|\mu_V - \mu_P|$. The first two things decrease the length of the intervals, and the last thing shifts the intervals further away from each other. Relatedly, it is worth knowing that—roughly speaking—wide confidence intervals correspond to experiments with low power, and narrow confidence intervals correspond to experiments with good power.

This is demonstrated in Figure 11.2. This figure shows an intuitive (rather than mathematically rigorous) view of the process of testing the equivalence of the population means of ad clicks for treatment P vs. treatment V. The top row represents population distributions of clicks for the two treatments. Each curve

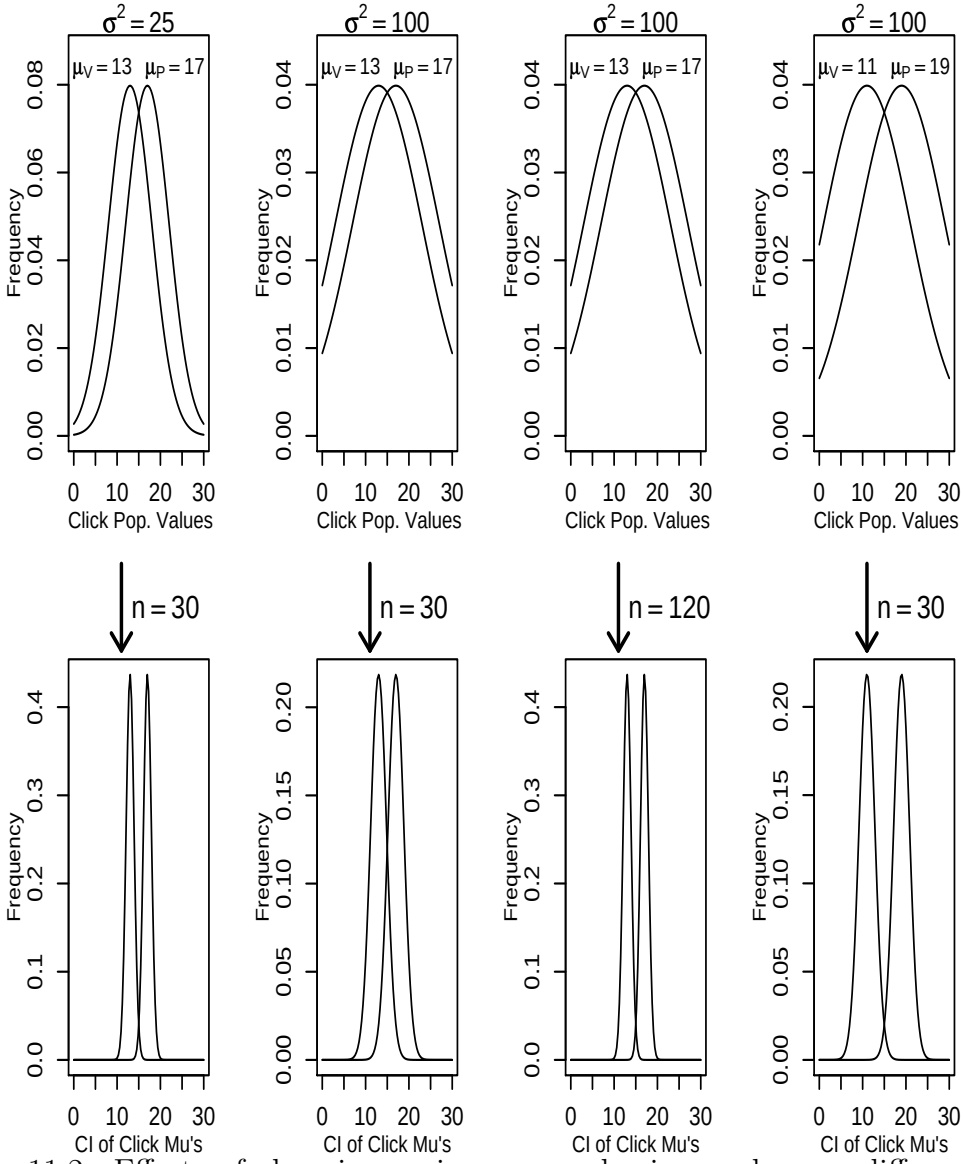


Figure 11.2: Effects of changing variance, sample size, and mean difference on power. Top row: population distributions of the outcome. Bottom row: sampling distributions of the sample mean for the given sample size.

can be thought of as the histogram of the actual click outcomes for one font for all news websites on the Internet. There is a lot of overlap between the two curves, so obviously it would not be very accurate to use, say, one website per font to try to determine if the population means differ.

The bottom row represents the sampling distributions of the sample means for the two treatments based on the given sample size (n) for each treatment. The key idea here is that, although the two curves always overlap, a smaller overlap corresponds to a greater chance that we will get a significant p-value for our one experiment.

Start with the second column of the figure. The upper panel shows that the truth is that σ^2 is 100, and $\mu_V = 13$, while $\mu_P = 17$. The arrow indicates that our sample has $n = 30$ websites with each font. The bottom panel of the second column shows the sampling distributions of sample means for the two treatments. The moderate degree of overlap, best seen by looking at the lower middle portion of the panel, is suggestive of less than ideal power.

The leftmost column shows the situation where the true common variance is now 25 instead of 100 (i.e., the s.d. is now 5 clicks instead of 10 clicks). This markedly reduces the overlap, so the power is improved. When we are designing an experiment, what are some ways that we can reduce the common variance? We could use a within-subjects design (as we will discuss in Chapter 13), or we could use a blocking variable (as we discussed in Chapters 7 and 8) or quantitative control variable (as we discussed in Chapter 10). Specific examples for reducing variation in this experiment include using only television-related websites, controlling the position of the ad on the website, and using only one font size for the ad. A within-subjects design would, e.g., randomly present one font from 7:00 to 7:30 and the other font from 7:30 to 8:00 for each website (which is considered the "subject" here), but would need a different analysis than the independent-samples t-test. Blocking would involve, e.g., using some important (categorical) aspect of the news websites, such as television-related vs. non-television related as a second factor whose p-value is not of primary interest (as in two-way ANOVA, as we discussed in Chapter 8). We would hope that for each level of this second variable the variance of the outcome for either treatment would be smaller than if we had ignored this variable. Finally, using a quantitative variable like site volume as an additional explanatory variable in an ANCOVA setting would similarly reduce variability (i.e., σ^2).

The third column shows what happens if the sample size is increased. Increasing

the sample size four-fold turns out to have the same effect on the confidence curves, and therefore the power, as reducing the variance four-fold. Of course, increasing sample size increases cost and duration of the study.

The fourth column shows what happens if the population mean difference, sometimes called (unadjusted) effect size, is increased. Although the sampling distributions are not narrowed, they are more distantly separated, thus reducing overlap and increasing the power. Although it is hard to see how the difference between the two fonts can be made larger in this example, sometimes we do have some control over the true population mean difference among treatment groups. For example, in a treatment-versus-control study, we could make the treatment "stronger" (e.g., increasing the dose), thereby increasing the population mean difference if there is one. However, this may also make the experiment less scientifically interesting—perhaps we are scientifically interested in a low-dose treatment versus a control. In that case, we would look towards changing other factors to increase the power.

Here is a description of another experiment with examples of how to improve the power. We want to test the effect of three kinds of fertilizer on plant growth (in grams). First we consider reducing the common variability of final plant weight for each fertilizer type. We can reduce measurement error by using a high quality laboratory balance instead of a cheap hardware store scale. And we can have a detailed, careful procedure for washing off the dirt from the roots and removing excess water before weighing. Subject-to-subject variation can be reduced by using only one variety of plant and doing whatever is possible to ensure that the plants are of similar size at the start of the experiment. Environmental variation can be reduced by assuring equal sunlight and water during the experiment. And treatment application variation can be reduced by carefully measuring and applying the fertilizer to the plants. As mentioned in Section 7.5, reduction in these different sources of variation (except measurement variability) also tends to reduce generalizability.

As usual, having more plants per fertilizer improves power, but at the expense of extra cost. We can also increase population mean differences by using a larger amount of fertilizer and/or running the experiment for a longer period of time. (Both of the latter ideas are based on the assumption that the plants grow at a constant rate proportional to the amount of fertilizer, but with different rates per unit time for the same amount of different fertilizers.)

A within-subjects design is not possible here, because a single plant cannot be

tested on more than one fertilizer type. In other words, once we have observed how much a plant has grown, we cannot reverse time in order to see how much it would have grown using a different fertilizer.

Blocking could be done based on different fields if the plants are grown outside in several different fields, or based on a subjective measure of initial "healthiness" of the plants (determined before randomizing plants to the different fertilizers). If the fertilizer is a source of, say, magnesium in different chemical forms, and if the plants are grown outside in natural soil, a possible control variable is the amount of nitrogen in the soil near each plant. Each of these blocking/control variables are expected to affect the outcome, but are not of primary interest. By including them in the means model, we are creating finer, more homogeneous divisions of "the set of experimental units with all explanatory variables set to the same values". The inherent variability of each of these sets of units, which we call σ^2 for any model, is smaller than for the larger, less homogeneous sets that we get when we don't include these variables in our model.

Reducing σ^2, increasing n, and increasing the spacing between population means will all reduce the overlap of the sampling distributions of the means, thus increasing power.

11.3 Case Studies on Type 1 and 2 error rates, power, and positive and negative error rates

People often confuse the probability of a Type 1 error and/or the probability of a Type 2 error with the probability that a given research result is false. This section attempts to clarify the situation by looking at several specific (fake) researchers' experiences over the course of their careers.

Remember that a given null hypothesis, H_0, is either true or false, but we can never know this truth for sure. Also, for a given experiment, the standard decision rule tells us that when $p \leq \alpha$ we should reject the null hypothesis, and when $p > \alpha$ we should retain it. But again, we can never know for sure whether our inference is actually correct or incorrect.

Next we need to clarify the definitions of some common terms. A "positive"

result for an experiment means finding $p \leq \alpha$, which is the situation for which we reject H_0. A "negative" result means finding $p > \alpha$, which is the situation for which we fail to reject (or retain) H_0. "True" means correct (i.e. reject H_0 when H_0 is false or retain H_0 when H_0 is true), and "false" means incorrect. These terms are commonly put together, e.g., a false positive refers to the case where $p \leq \alpha$, but the null hypothesis is actually true.

Here are some examples in which we pretend that we have omniscience, although the researcher in question does not. Let $\alpha = 0.05$ unless otherwise specified.

1. Neetika Null studies the effects of various chants on blood sugar level. Every week she studies 15 controls and 15 people who chant a particular word from the dictionary for 5 minutes. After 1000 weeks (and 1000 words) what is her Type 1 error rate (positives among null experiments), Type 2 error rate (negatives among non-null experiments) and power (positives among non-null experiments)? What percent of her positives are true? What percent of her negatives are true?

   This description suggests that the null hypothesis is always true, i.e. I assume that chants don't change blood sugar level, and certainly not within five minutes. Her Type 1 error rate is $\alpha = 0.05$. Her Type 2 error rate (sometimes called β) and power are not applicable because no alternative hypothesis is ever true. Out of 1000 experiments, 1000 are null in the sense that the null hypothesis is true. Because the probability of getting $p \leq 0.05$ in an experiment where the null hypothesis is true is 5%, she will see about 50 positive and 950 negative experiments. For Neetika, although she does not know it, every time she sees $p \leq 0.05$ she will mistakenly reject the null hypothesis, for a 100% error rate. But every time she sees $p > 0.05$ she will correctly retain the null hypothesis for an error rate of 0%.

2. Stacy Safety studies the effects on glucose levels of injecting cats with subcutaneous insulin at different body locations. She divides the surface of a cat into 1000 zones and each week studies injection of 10 cats with water and 10 cats with insulin in a different zone.

   This description suggests that the null hypothesis is always false. Because Stacy is studying a powerful treatment and will have a small measurement error, her power will be large; let's use 80%=0.80 as an example. Her Type 2 error rate will be β=1-power=0.2, or 20%. Out of 1000 experiments, all

1000 are non-null, so Type 1 error is not applicable. With a power of 80% we know that each experiment has an 80% chance of giving $p \leq 0.05$ and a 20% chance of given $p > 0.05$. So we expect around 800 positives and 200 negatives. Although Stacy doesn't know it, every time she sees $p \leq 0.05$ she will correctly reject the null hypothesis, for a 0% error rate. But every time she sees $p > 0.05$ she will mistakenly retain the null hypothesis for an error rate of 100%.

3. Rima Regular works for a large pharmaceutical firm performing initial screening of potential new oral hypoglycemic drugs. Each week for 1000 weeks she gives 100 rats a placebo and 100 rats a new drug, then tests blood sugar. To increase power (at the expense of more false positives) she chooses $\alpha = 0.10$.

   For concreteness let's assume that the null hypothesis is true 90% of the time. Let's consider the situation where among the 10% of candidate drugs that work, half have a strength that corresponds to power equal to 50% (for the given n and σ^2) and the other half correspond to power equal to 70%.

   Out of 1000 experiments, 900 are null with around 0.10*900=90 positive and 810 negative experiments. Of the 50 non-null experiments with 50% power, we expect around 0.50*50=25 positive and 25 negative experiments. Of the 50 non-null experiments with 70% power, we expect around 0.70*50=35 positive and 15 negative experiments. So among the 100 non-null experiments (i.e., when Rima is studying drugs that really work) 25+35=60 out of 100 will correctly give $p \leq 0.05$. Therefore Rima's average power is 60/100 or 60%.

   Although Rima doesn't know it, when she sees $p \leq 0.05$ and rejects the null hypothesis, around 60/(90+60)=0.40=40% of the time she is correctly rejecting the null hypothesis, and therefore 60% of the time when she rejects the null hypothesis she is making a mistake. Of the 810+40=850 experiments for which she finds $p > 0.05$ and retains the null hypothesis, she is correct 810/(810+40)=0.953=95.3% of time and she makes an error 4.7% of the time. (Note that this value of approximately 95% is only a coincidence, and not related to $\alpha = 0.05$; in fact $\alpha = 0.10$ for this problem.)

   These error rates are not too bad given Rima's goals, but they are not very intuitively related to $\alpha = 0.10$ and power equal to 50 or 70%. The 60% error rate among drugs that are flagged for further study (i.e., have $p \leq 0.05$) just indicates that some time and money will be spent to find out which of these drugs are not really useful. This is better than not investigating a drug that

really works. (The chance of this happening is 40%, i.e., the Type 2 error rate. Note that this is equal to 1 - power, and power is 60%.) In fact, Rima might make even more money for her company if she raises α to 0.20, causing more money to be wasted investigating truly useless drugs, but preventing some possible money-making drugs from slipping through as useless. By the way, the overall error rate is (90+40)/1000=13%.

Conclusion: For *your* career, you cannot know the chance that a negative result is an error or the chance that a positive result is an error. And these are what you would really like to know! But you do know that when you study "ineffective" treatments (and perform an appropriate statistical analysis) you have only a 5% chance of incorrectly claiming they are "effective". And you know that the more you increase the power of an experiment, the better your chances are of detecting a truly effective treatment.

The error rates that experimenters are really interested in—i.e., the probability that they are making an error for their current experiment—are not knowable. These error rates differ from both α and β=1-power.

11.4 Expected Mean Square

In Section 11.5, we will discuss how to calculate the power of a particular experiment for a given alternative hypothesis. Power calculations—as well as many calculations in statistics—revolve around "expected mean squares" (EMS), and so first we will delve shallowly into EMS. Learning about EMS will not only give you a better understanding of power calculations in the next section, but it will also give you a deeper understanding of the calculations involved in two-way ANOVA, which we saw in the previous chapter.

Although a full treatment of expected mean squares is quite technical, a superficial understanding is not difficult and greatly aids understanding of several other topics. EMS tells us what values we will get for any given mean square (MS) statistic under either the null or an alternative distribution, on average over

Source of Variation	MS	EMS
Factor A	MS_A	$\sigma_e^2 + n\sigma_A^2$
Error (residual)	MS_{error}	σ_e^2

Table 11.1: Expected mean squares for a one-way ANOVA.

repeated experiments. Remember that MS statistics are essential to one- and two-way ANOVA, because F statistics are computed as the ratio of two MS values.

If we have k population treatment means, we can define $\bar{\mu} = \frac{\sum_{i=1}^{k} \mu_i}{k}$ as the mean of the population treatment means, and $\lambda_i = \mu_i - \bar{\mu}$ (where λ is read "lambda"), and $\sigma_A^2 = \frac{\sum_{i=1}^{k} \lambda_i^2}{k-1}$. The quantity σ_A^2 is simply the empirical variance of the μ_i. Notice that we can express our usual null hypothesis as $H_0 : \sigma_A^2 = 0$ because if all of the μ's are equal, then all of the λ's equal zero. We can similarly define σ_B^2 and σ_{A*B}^2 for a 2 way design.

Let σ_e^2 be the true error variance. (We haven't been using the subscript "e" up to this point, but here we will use it to be sure we can distinguish various symbols that all include σ^2.) As usual, n is the number of subjects per group. For 2-way ANOVA, a (instead of k) is the number of levels of factor A and b is the number of levels of factor B.

The EMS tables for one-way and two-way designs are shown in table 11.1 and 11.2.

Remember that all of the between-subjects ANOVA F-statistics are ratios of mean squares with various means squares in the numerator and with the error mean square in the denominator. From the EMS tables, you can see why, for either design, under the null hypothesis, the F ratios that we have been using are appropriate and have "central F" sampling distributions (mean near 1). You can also see why, under any alternative, these F ratios tend to get bigger. You can also see that power can be increased by increasing the spacing between population means ("treatment strength") via increased values of $|\lambda|$, by increasing n, or by decreasing σ_e^2. This formula also demonstrates that the value of σ_e^2 is irrelevant to the sampling distributing of the F-statistic (cancels out) when the null hypothesis is true, i.e., $\sigma_A^2 = 0$.

Source of Variation	MS	EMS
Factor A	MS_A	$\sigma_e^2 + bn\sigma_A^2$
Factor B	MS_B	$\sigma_e^2 + an\sigma_B^2$
A*B interaction	MS_{A*B}	$\sigma_e^2 + n\sigma_{AB}^2$
Error (residual)	MS_{error}	σ_e^2

Table 11.2: Expected mean squares for a two-way ANOVA.

For the mathematically inclined, the EMS formulas give a good idea of what aspects of an experiment affect the F ratio.

11.5 Power Calculations

In case it is not yet obvious, I want to reiterate why it is imperative to calculate power for your experiment *before* running it. It is possible and common for experiments to have low power, e.g., in the range of 20 to 70%. If you are studying a treatment which is effective in changing the population mean of your outcome, and your experiment has, e.g., 40% power for detecting the true mean difference, and you conduct the experiment perfectly and analyze it appropriately, you have a 60% chance of getting a p-value greater than 0.05, in which case you will erroneously conclude that the treatment is ineffective. To prevent wasted experiments, you should calculate power and only perform the experiment if there is reasonably high power.

It is worth noting that you will not be able to calculate the "true" power of your experiment. This is because power depends on how much the population means truly differ, which you can never know. Instead, you will use a combination of mathematics and judgment to make a useful estimation of the power.

There are an infinite number of alternative hypotheses. For any of them we can increase power by 1) increasing n (sample size) or 2) decreasing experimental error (σ_e^2). Also, among the alternatives, those with larger effect sizes (population mean differences) will have more power. These statements derive directly from the EMS interpretive form of the F equation (shown here for 1-way ANOVA):

$$\text{Expected Value of F} = \text{Expected value of } \frac{MS_A}{MS_{\text{error}}} \approx \frac{\sigma_e^2 + n\sigma_A^2}{\sigma_e^2}$$

Note that this is only an approximation because the expectation of a ratio is *not* the ratio of expectations (which we're using as an approximation here). Nonetheless, from the above we can see that increasing n or σ_A^2 increases the average value of F. Regarding the effect of changing σ_e^2, a small example will make this more clear. Consider the case where $n\sigma_A^2 = 10$ and $\sigma_e^2 = 10$. In this case, the average F value is 20/10=2. Now reduce σ_e^2 to 1. In this case, the average F value is 11/1=11, which is much bigger, resulting in more power. In short, the larger the proportion of variance that is due to treatment effects ($n\sigma_A^2$) instead of error (σ_e^2), the larger the power.

In practice, we try to calculate the power of an experiment for one or a few reasonable alternative hypotheses. We try not to get carried away by considering alternatives with huge effects that are unlikely to occur. Instead we try to devise alternatives that are fairly conservative and reflect what might really happen (see the next section).

What do we need to know to calculate power? Beyond k and alpha (α), we need to know sample size (which we may be able to increase if we have enough resources), an estimate of experimental error (variance or σ_e^2, which we may be able to reduce, possibly in a trade-off with generalizability), and reasonable estimates of true effect sizes.

For any set of these three things, which we will call an "alternative hypothesis scenario", we can find the sampling distribution of F under that alternative hypothesis. Then it is easy to find the power.

We often estimate σ_e^2 with residual MS, or error MS (MSE), or within-group MS from previous similar experiments. Or we can use the square of the actual or guessed standard deviation of the outcome measurement for a number of subjects exposed to the same (any) treatment. Or, assuming Normality, we can use expert knowledge to guesstimate the 95% range of a homogenous group of subjects, then estimate σ_e as that range divided by 4. (This works because 95% of a normal distribution is encompassed by mean plus or minus 2 s.d.) A similar trick is to estimate σ_e as 3/4 of the IQR (see Section 4.2.4), then square that quantity. It's also very common to choose σ_e^2 to be proportional to the effect size. In short, there are many rules-of-thumb for specifying the variance σ_e^2 in power calculations.

But be careful! If you use too large (pessimistic) of a value for σ_e^2 your computed power will be smaller than your true power. If you use too small (optimistic) of a value for σ_e^2 your computed power will be larger than your true power.

11.6 Choosing effect sizes

As mentioned above, you want to calculate power for "reasonable" effect sizes that you consider achievable. A similar goal is to choose effects sizes such that smaller effects would not be scientifically interesting. In either case, it is obvious that choosing effect sizes is not a statistical exercise, but rather one requiring subject matter or possibly policy level expertise.

I will give a few simple examples here, choosing subject matter that is known to most people or easily explainable. The first example is for a categorical outcome, even though we haven't yet discussed statistical analyses for such experiments (we will discuss this in Chapter 15). Consider an experiment to see if a certain change in a TV commercial for a political advisor's candidate will make a difference in an election. Here is the kind of thinking that goes into defining the effect sizes for which we will calculate the power. Let's say that, from prior subject matter knowledge, the candidate's PR manager estimates that about one fourth of the voting public will see the commercial. He also estimates that a change of 1% in the total vote will be enough to get him excited that redoing this commercial is a worthwhile expense. So therefore an effect size of 4% difference in a favorable response towards his candidate is the effect size that is reasonable to test for.

Now consider an example of a farmer who wants to know if it's worth it to move her tomato crop in the future to a farther, but more sunny slope. She estimates that the cost of initially preparing the field is \$2000, the yearly extra cost of transportation to the new field is \$200, and she would like any payoff to happen within 4 years. The effect size is the difference in crop yield in pounds of tomatoes per plant. She can put 1000 plants in either field, and a pound of tomatoes sells for \$1 wholesale. So for each 1 pound of effect size, she gains \$1000 per year. Over 4 years she needs to pay off \$2000+4(\$200)=\$2800. She concludes that she needs to have good power, say 80%, to detect an effect size of 2.8/4=0.7 additional pounds of tomatoes per plant (i.e., a gain of \$700 per year).

Finally consider a psychologist who wants to test the effects of a drug on memory. She knows that people typically remember 40 out of 50 items on this test. She

really wouldn't get too excited if the drug raised the score to 41, but she certainly wouldn't want to miss it if the drug raised the score to 45. She decides to "power her study" for $\mu_1 = 40$ vs. $\mu_2 = 42.5$. If she adjusts n to get 80% power for these population test score means, then she has an 80% chance of getting $p \leq 0.05$ when the true effect is a difference of 2.5, and some larger (calculable) power for a difference of 5.0, and some smaller (calculable) non-zero, but less than ideal, power for a difference of 1.0.

In general, you should consider the smallest effect size that you consider interesting and try to achieve reasonable power for that effect size, while also realizing that there is more power for larger effects and less power for smaller effects. Sometimes it is worth calculating power for a range of different effect sizes.

11.7 Using n.c.p. to calculate power

The material in this section is optional.

Here we will focus on the simple case of power in a one-way between-subjects design. The "manual" calculation steps are shown here. Understanding these may aid your understanding of power calculation in general, but ordinarily you will use a computer (perhaps a web applet) to calculate power.

Under any particular alternative distribution the numerator of F is inflated, and F follows the non-central F distribution with $k-1$ and $k(n-1)$ degrees of freedom and with "non-centrality parameter" equal to:

$$\text{n.c.p.} = \frac{n \cdot \sum_{i=1}^{k} \lambda_i^2}{\sigma_e^2}$$

where n is the proposed number of subjects in *each* of the groups we are comparing. The bigger the n.c.p., the more the alternative sampling distribution moves to the right and the more power we have.

Manual calculation example: Let $\alpha = 0.10$ and $n = 11$ per cell. In a similar experiment MSE=36. What is the power for the alternative hypothesis $H_A : \mu_1 = 10$, $\mu_2 = 12$, $\mu_3 = 14$, $\mu_4 = 16$?

1. Under the null hypothesis the F-statistic will follow the central F distribution (i.e., n.c.p.=0) with $k - 1 = 3$ and $k(n-1) = 40$ df. Using a computer or F table we find $F_{\text{critical}} = 2.23$.

2. Since $\bar{\mu}$=(10+12+14+16)/4=13, the λ's are -3,-1,1,3, so the non-centrality parameter is

$$\frac{11(9+1+1+9)}{36} = 6.11.$$

3. The power is the area under the non-central F curve with 3,40 df and n.c.p.=6.11 that is to the right of 2.23. Using a computer or non-central F table, we find that the area is 0.62. This means that we have a 62% chance of rejecting the null hypothesis if the given alternate hypothesis is true.

4. An interesting question is what is the power if we double the sample size to 22 per cell. df_{error} is now 21*4=84 and $F_{critical}$ is now 2.15. The n.c.p.=12.22. From the appropriate non-central F distribution we find that the power increases to 90%.

In practice we will use a Java applet to calculate power.

In R, the commands that give the values in the above example are:

```
qf(1-0.10, 3, 40) # result is 2.226092 for alpha=0.10
1-pf(2.23, 3, 40, 6.11) # result is 0.6168411
qf(1-0.10, 3, 84) # result is 2.150162
1-pf(2.15,3, 84, 12.22) # result is 0.8994447
```

11.8 A power applet

This section is meant to be a reference if ever you find the need to calculate power in your own research. The Russ Lenth power applet is very nice way to calculate power. It is available at http://www.cs.uiowa.edu/~rlenth/Power. You will have to download a Java file and follow some instructions on the website for opening the file. Here I will cover ANOVA and regression.

11.8.1 Overview

To get started with the Lenth Power Applet, select a method such as Linear Regression or Balanced ANOVA, then click the "Run Selection" button. A new window will open with the applet for the statistical method you have chosen. Every time you see sliders for entering numeric values, you may also click the small square at upper right to change to a text box form for entering the value. The Help menu item explains what each input slider or box is for.

11.8.2 One-way ANOVA

This part of the applet works for one-way and two-way balanced ANOVA. Remember that balanced indicates equal numbers of subjects per group. For one-way ANOVA, leave the "Built-in models" drop-down box at the default value of "One-way ANOVA".

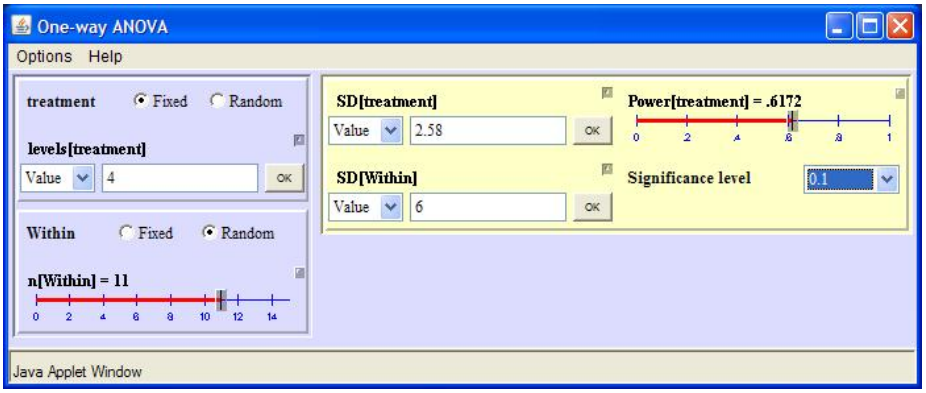


Figure 11.3: One-way ANOVA with Lenth power applet.

Enter "n" under "Observations per factor combination", and click to study the power of "F tests". A window opens that looks like Figure 11.3.

On the left, enter "k" under "levels[treatment] (Fixed)". Under "n[Within] (Random)" you can change n.

On the right enter σ_e (σ) under "SD[Within]" (on the standard deviation, not variance scale) and α under "Significance level". Finally you need to enter the

"effect size" in the form of "SD[treatment]". For this applet the formula is

$$\text{SD[treatment]} = \sqrt{\frac{\sum_{i=1}^{k} \lambda_i^2}{k-1}}$$

where λ_i is $\mu_i - \bar{\mu}$ as in Section 11.4.

For $H_A : \mu_1 = 10$, $\mu_2 = 12$, $\mu_3 = 14$, $\mu_4 = 16$, $\bar{\mu} = 13$ and $\lambda_1 = -3$, $\lambda_2 = -1$, $\lambda_3 = +1$, $\lambda_4 = +3$.

$$\begin{aligned} \text{SD[treatment]} &= \sqrt{\frac{\sum_{i=1}^{k} \lambda_i^2}{k-1}} \\ &= \sqrt{\frac{(-3)^2 + (-1)^2 + (+1)^2 + (+3)^2}{3}} \\ &= \sqrt{20/3} \\ &= 2.58 \end{aligned}$$

You can also use the menu item "SD Helper" under Options to graphically set the means and have the applet calculate SD[treatment].

Following the example of Section 11.7 we can plug in SD[treatment]=2.58, $n = 11$, and $\sigma_e = 6$ to get power=0.6172, which matches the manual calculation of section 11.7

At this point it is often useful to make a power plot. Choose Graph under the Options menu item. The most useful graph has "Power[treatment]" on the y-axis and "n[Within]" on the x-axis. Continuing with the above example I would choose to plot power "from" 5 "to" 40 "by" 1. When I click "Draw", I see the power for this experiment for different possible sample sizes. An interesting addition can be obtained by clicking "Persistent", then changing "SD[treatment]" in the main window to another reasonable value, e.g., 2 (for $H_A : \mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = 10$, $\mu_4 = 14$), and clicking OK. Now the plot shows power as a function of n for two (or more) effect sizes. In Windows you can use the Alt-PrintScreen key combination to copy the plot to the clipboard, then paste it into another application. The result is shown in Figure 11.4. The lower curve is for the smaller value of SD[treatment].

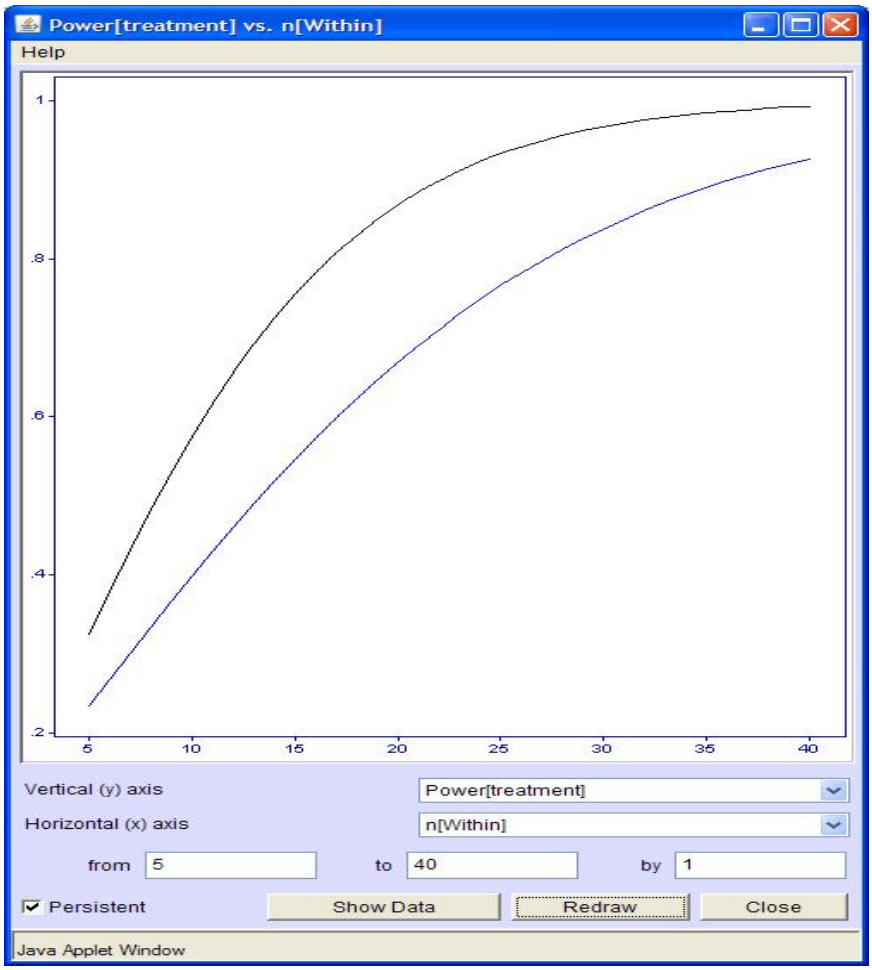


Figure 11.4: One-way ANOVA power plot from Lenth power applet.

11.8.3 Two-way ANOVA without interaction

Select "Two-way ANOVA (additive model)". Click "F tests". In the new window, on the left enter the number of levels for each of the two factors under "levels[row] (Fixed)" and "levels[col] (Fixed)". Enter the number of subjects for each cell under "Replications (Random)".

Enter the estimate of σ under "SD[Residual]" and the enter the "Significance level".

Calculate "SD[row]" and "SD[col]" as in the one-way ANOVA calculation for "SD[treatment]", but the means for either factor are now averaged over all levels of the other factor.

Here is an example. The table shows cell population means for each combination of levels of the two treatment factors for which additivity holds (e.g., a profile plot would show parallel lines).

Row factor / Column Factor	Level 1	Level 2	Level 3	Row Mean
Level 1	10	20	15	15
Level 2	13	23	18	18
Col. Mean	11.5	21.5	16.5	16.5

Averaging over the other factor we see that for the column means, using some fairly obvious invented notation we get $H_{ColAlt} : \mu_{C1} = 11.5, \mu_{C2} = 21.5, \mu_{C3} = 16.5$. The row means are $H_{RowAlt} : \mu_{R1} = 15, \mu_{R2} = 18$.

Therefore SD[row] is the square root of $((-1.5)^2 + (+1.5)^2)/1$ which is 2.12. The value of SD[col] is the square root of $((-5)^2 + (+5)^2 + (0)^2)/2$ which equals 5. If we choose $\alpha = 0.05$, $n = 8$ per cell, and estimate σ at 8, then the power is a not-so-good 24.6% for H_{RowAlt}, but a very good 87.4% for H_{ColAlt}.

11.8.4 Two-way ANOVA with interaction

You may someday find it useful to calculate the power for a two-way ANOVA interaction. It's fairly complicated!

Select "Two-way ANOVA". Click "F tests". In the new window, on the left enter the number of levels for each of the two factors under "levels[row] (Fixed)" and "levels[col] (Fixed)". Enter the number of subjects for each cell under "Replications (Random)".

Enter the estimate of σ under "SD[Residual]" and the enter the "Significance level".

The treatment effects are a bit more complicated here. Consider a table of cell means in which additivity does not hold.

Row factor / Column Factor	Level 1	Level 2	Level 3	Row Mean
Level 1	10	20	15	15
Level 2	13	20	18	17
Col. Mean	11.5	20.0	16.5	16

For the row effects, which come from the row means of 15 and 17, we subtract 16 from each to get the λ values of -1 and 1, then find SD[row]=$\sqrt{\frac{(-1)^2+(1)^2}{1}} = 1.41$.

For the column effects, which come from the column means of 11.5, 20.0, and 16.5, we subtract their common mean of 16 to get λ values of -4.5, 4.0, and 0.5, and then find that SD[col]=$\sqrt{\frac{(-4.5)^2+(4.0)^2+(0.5)^2}{2}} = 4.27$.

To calculate "SD[row*col]" we need to calculate for each of the 6 cells, the value of $\mu_{ij} - (\bar{\mu} + \lambda_{i.} + \lambda_{.j})$ where μ_{ij} indicates the i^{th} row and j^{th} column, and $\lambda_{i.}$ is the λ value for the i^{th} row mean, and $\lambda_{.j}$ is the λ value for the j^{th} column mean. For example, for the top left cell we get 10-(16-4.5-1.0)=-0.5. The complete table is

Row factor / Column Factor	Level 1	Level 2	Level 3	Row Mean
Level 1	-0.5	1.0	-0.5	0.0
Level 2	+0.5	-1.0	0.5	0.0
Col. Mean	0.0	0.0	0.0	0.0

You will know you constructed the table correctly if all of the margins are zero. To find SD[row*col], sum the squares of all of the (non-marginal) cells, then divide by $(r-1)$ and $(c-1)$ where r and c are the number of levels in the row and column factors, then take the square root. Here we get SD[row*col]=$\sqrt{\frac{0.25+1.0+0.25+0.25+1.0+0.25}{1 \cdot 2}} = 1.22$.

If we choose $\alpha = 0.05$, $n = 7$ per cell, and estimate σ at 3, then the power is a not-so-good 23.8% for detecting the interaction (gettin an interaction p-value less than 0.05). This is shown in Figure 11.5.

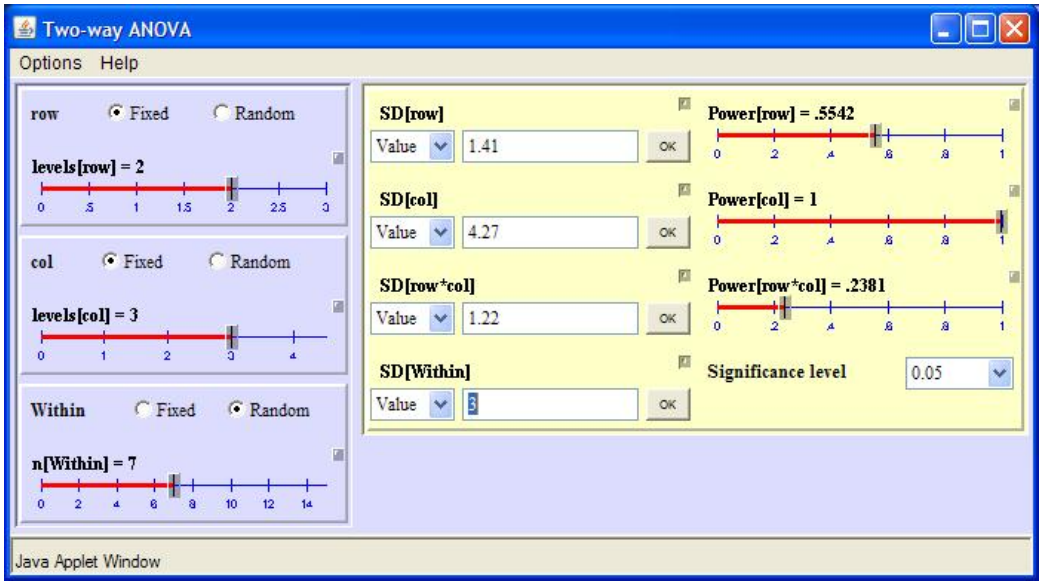


Figure 11.5: Two-way ANOVA with Lenth power applet.

11.8.5 Linear Regression

We will just look at simple linear regression (one explanatory variable). In addition to the α, n, and σ, and the effect size for the slope, we need to characterize the spacing of the *explanatory* variable.

Choose "Linear regression" in the applet and the Linear Regression dialog shown in Figure 11.6 appears. Leave "No. of predictors" (number of explanatory variables) at 1, and set "Alpha", "Error SD" (estimate of σ), and "(Total) Sample size".

Under "SD of x[j]" enter the standard deviation of the x values you will use. Here we use the fact that the spread of any number of repetitions of a set of values is the same as just one set of those values. Also, because the x values are fixed, we use n instead of $n-1$ in the denominator of the standard deviation formula. E.g., if we plan to use 5 subjects each at doses, 0, 25, 50, and 100 (which have a mean of 43.75), then SD of x[j] $= \sqrt{\frac{(0-43.75)^2+(25-43.75)^2+(50-43.75)^2+(100-43.75)^2}{4}} = 36.98$.

Plugging in this value and $\sigma = 30$, and a sample size of 3*4=12, and an effect size of beta[j] (slope) equal to 0.5, we get power = 48.8%, which is not good enough.

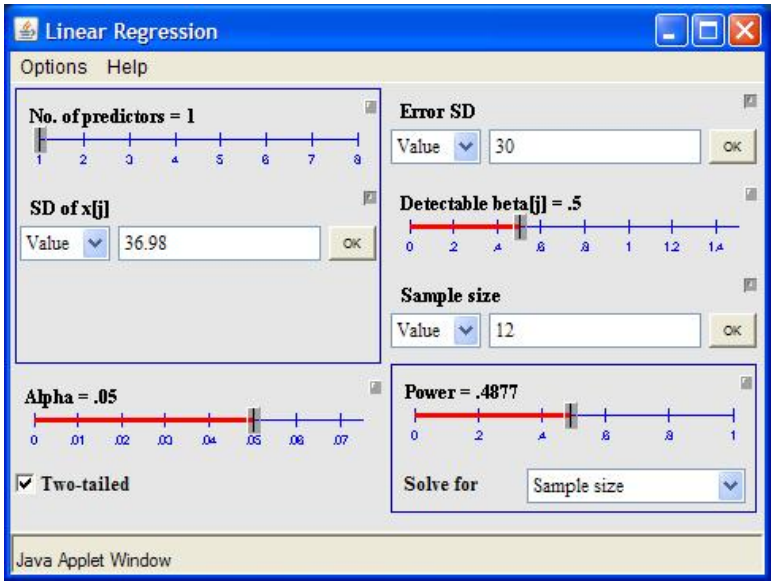


Figure 11.6: Linear regression with Lenth power applet.

In a nutshell: Just like the most commonly used value for alpha is 0.05, you will find that (arbitrarily) the most common approach people take is to find the value of n that achieves a power of 80% for some specific, carefully chosen alternative hypothesis. Although there is a bit of educated guesswork in calculating (estimating) power, it is strongly advised to make some power calculations before running an experiment to find out if you have enough power to make running the experiment worthwhile.

Chapter 12

Contrasts and Custom Hypotheses

Contrasts ask specific questions as opposed to the general ANOVA null vs. alternative hypotheses.

In a one-way ANOVA with a k level factor, the null hypothesis is $\mu_1 = \cdots = \mu_k$, and the alternative is that at least one group (treatment) population mean of the outcome differs from the others. If $k = 2$, and the null hypothesis is rejected we need only look at the sample means to see which treatment is "better". But if $k > 2$, rejection of the null hypothesis does not give the full information of interest. All we know is that there is some significant difference among the population means, but we don't know which particular population means significantly differ. For example, in a test of the effects of control and two active treatments to increase vocabulary, we might find that based on a high value for the F-statistic we are justified in rejecting the null hypothesis $\mu_1 = \mu_2 = \mu_3$. If the sample means of the outcome are 50, 75, and 80 respectively, we need additional testing to answer specific questions like "Is the control population mean lower than the average of the two active treatment population means?" and "Are the two active treatment population means different?" To answer questions like these, we frame "custom" hypotheses, which are formally expressed as **contrast hypotheses**.

12.1 Contrasts in general

A contrast null hypothesis compares two population means or combinations of population means. A **simple contrast hypothesis** compares two population means, e.g. $H_0 : \mu_1 = \mu_5$. The corresponding inequality is the alternative hypothesis: $H_1 : \mu_1 \neq \mu_5$. These are the same kind of hypotheses we saw when conducting t-tests in Chapter 5.

A contrast null hypothesis that has multiple population means on either or both sides of the equal sign is called a **complex contrast hypothesis**. In the vast majority of practical cases, the multiple population means are combined as their mean, e.g., the custom null hypothesis $H_0 : \frac{\mu_1+\mu_2}{2} = \frac{\mu_3+\mu_4+\mu_5}{3}$ represents a test of the equality of the average of the first two treatment population means to the average of the next three. An example where this would be useful and interesting is when we are studying five ways to improve vocabulary, the first two of which are different written methods and the last three of which are different verbal methods.

It is customary to rewrite the null hypothesis with all of the population means on one side of the equal sign and a zero on the other side. E.g., $H_0 : \mu_1 - \mu_5 = 0$ or $H_0 : \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4+\mu_5}{3} = 0$. This mathematical form—whose left side is checked for equality to zero—is the standard form for a contrast. In addition to hypothesis testing, it is also often of interest to place a confidence interval around a contrast of population means, e.g., we might calculate that the 95% CI for $\mu_3 - \mu_4$ is [-5.0, +3.5]. As we've noted previously, this is different from computing two 95% CIs (one for μ_3 and one for μ_4).

As in the rest of classical statistics, we proceed by finding the null sampling distribution of the contrast statistic. A little bit of formalism is needed so that we can enter the correct custom information into a computer program, which will then calculate the contrast statistic (estimate of the population contrast), the standard error of the statistic, a corresponding t-statistic, and the appropriate p-value. As shown later, this process only works under the special circumstances called "planned comparisons"; otherwise it requires some modifications.

Let γ (gamma) represent the population contrast. In this section, we will use an example from a single six level one-way ANOVA, and use subscripts 1 and 2 to distinguish two specific contrasts. As an example of a simple (population) contrast, define γ_1 to be $\mu_3 - \mu_4$, a contrast of the population means of the outcomes for the third vs. the fourth treatments. As an example of a complex contrast let γ_2 be $\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4+\mu_5}{3}$, a contrast of the population mean of the outcome for the first

two treatments to the population mean of the outcome for the third through fifth treatments. We can write the corresponding hypotheses as $H_{01} : \gamma_1 = 0,\ H_{A1} : \gamma_1 \neq 0$ and $H_{02} : \gamma_2 = 0,\ H_{A2} : \gamma_2 \neq 0$.

If we call the corresponding estimates, g_1 and g_2 then the appropriate estimates are $g_1 = \bar{y}_3 - \bar{y}_4$ and $g_2 = \frac{\bar{y}_1+\bar{y}_2}{2} - \frac{\bar{y}_3+\bar{y}_4+\bar{y}_5}{3}$. In the hypothesis testing situation, we are testing whether or not these estimates are consistent with the corresponding null hypothesis. For a confidence interval on a particular population contrast (γ), these estimates will be at the center of the confidence interval.

In the chapter on probability theory, we saw that the sampling distribution of any of the sample means from a (one treatment) sample of size n using the assumptions of Normality, equal variance, and independent errors is $\bar{y}_i \sim N(\mu_i, \sigma^2/n)$, i.e., across repeated experiments, a sample mean is Normally distributed with the "correct" mean and the variance equal to the common group variance reduced by a factor of n. Now we need to find the sampling distribution for some particular combination of sample means.

To do this, we need to write the contrast in "standard form". The standard form involves writing a sum with one term for *each* population mean (μ), *whether or not it is in the particular contrast*, and with a single number, called a **contrast coefficient** in front of each population mean. For our examples we get:

$$\gamma_1 = (0)\mu_1 + (0)\mu_2 + (0)\mu_3 + (1)\mu_4 + (-1)\mu_5 + (0)\mu_6$$

and

$$\gamma_2 = (1/2)\mu_1 + (1/2)\mu_2 + (-1/3)\mu_3 + (-1/3)\mu_4 + (-1/3)\mu_5 + (0)\mu_6.$$

In a more general framing of the contrast we would write

$$\gamma = C_1\mu_1 + \cdots + C_k\mu_k.$$

In other words, each contrast can be summarized by specifying its k coefficients (C values). And it turns out that the k coefficients are what most computer programs want as input when you specify the contrast of a custom null hypothesis.

In our examples, the coefficients (and computer input) for null hypothesis H_{01} are [0, 0, 1, -1, 0, 0], and for H_{02} they are [1/2, 1/2, -1/3, -1/3, -1/3, 0]. Note that the zeros *are* necessary. For example, if you just entered [1, -1], the computer would not understand which pair of treatment population means you want it to compare. Also, note that any valid set of contrast coefficients must add to zero.

It is okay to multiply the set of coefficients by any (non-zero) number. For example, we could also specify H_{02} as [3, 3, -2, -2, -2, 0] and [-3, -3, 2, 2, 2, 0]. These alternate contrast coefficients give the same p-value, but they do give different estimates of γ, and that must be taken into account when you interpret confidence intervals. However, usually we are most interested in hypotheses that test the difference between averages, which must be written using fractions.

A positive estimate for γ indicates higher means for the groups with positive coefficients compared to those with negative coefficients, while a negative estimate for γ indicates higher means for the groups with negative coefficients compared to those with positive coefficients. This is exactly the same kind of inference that we did when conducting t-tests.

To get a computer program to test a custom hypothesis, you must enter the k coefficients that specify that hypothesis.

If you can handle a bit more math, read the theory behind contrast estimates provided here.

The simplest case is for two independent random variables Y_1 and Y_2 for which the population means are μ_1 and μ_2 and the variances are σ_1^2 and σ_2^2. (We allow unequal variance, because even under the equal variance assumption, the sampling distribution of two means, depends on their sample sizes, which might not be equal.) In this case it is true that $E(C_1Y_1 + C_2Y_2) = C_1\mu_1 + C_2\mu_2$ and $\text{Var}(C_1Y_1 + C_2Y_2) = C_1^2\sigma_1^2 + C_2^2\sigma_2^2$. If in addition, the distributions of the random variables are Normal, we can conclude that the distribution of the linear combination of the random variables is also Normal. Therefore $Y_1 \sim N(\mu_1, \sigma_1^2),\ Y_2 \sim N(\mu_2, \sigma_2^2),\ \Rightarrow C_1Y_1 + C_2Y_2 \sim N(C_1\mu_1 + C_2\mu_2,\ C_1^2\sigma_1^2 + C_2^2\sigma_2^2)$.

We will also use the fact that if each of several independent random variables has variance σ^2, then the variance of a sample mean of n of these has variance σ^2/n.

From these ideas (and some algebra) we find that in a one-way ANOVA with k treatments, where the group sample means are independent, if we let σ^2 be the common population variance, and n_i be the number of subjects sampled for treatment i, then $\text{Var}(g) = \text{Var}(C_1\bar{Y_1} + \cdots + C_k\bar{Y_k}) = \sigma^2[\sum_{i=1}^{k}(C_i^2/n_i)]$.

In a real data analysis, we don't know σ^2 so we substitute its estimate, the within-group mean square. Then the square root of the estimated variance is the standard error of the contrast estimate, SE(g).

For any normally distributed quantity, g, which is an estimate of a parameter, γ, we can construct a t-statistic, $(g - \gamma)/\text{SE}(g)$. Then the sampling distribution of that t-statistic will be that of the t-distribution with df equal to the number of degrees of freedom in the standard error $(\text{df}_{\text{within}})$.

From this we can make a hypothesis test using $H_0 : \gamma = 0$, or we can construct a confidence interval for γ, centered around g.

For two-way (or higher) ANOVA without interaction, main effects contrasts are constructed separately for each factor, where the population means represent setting a specific level for one factor and ignoring (averaging over) all levels of the other factor.

For two-way ANOVA with interaction, contrasts are a bit more complicated. E.g., if one factor is job classification (with k levels) and the other factor is incentive applied (with m levels), and the outcome is productivity, we might be interested in comparing any particular combination of factor levels to any other combination. In this case, a one-way ANOVA with $k \cdot m$ levels is probably the best way to go.

On the other hand, if we are only interested in comparing the size of the mean differences for two particular levels of one factor across two levels of the other factor, then we are more clearly in an "interaction framework", and contrasts written for the two-way ANOVA make the most sense. E.g., if the subscripts on mu represent the levels of the two factors, we might be interested in a confidence

interval on the contrast $(\mu_{1,3} - \mu_{1,5}) - (\mu_{2,3} - \mu_{2,5})$.

The contrast idea extends easily to two-way ANOVA with no interaction, but can be more complicated if there is an interaction.

12.2 The issue of multiple comparisons

So far in this book, we have only considered scenarios where we are testing a single hypothesis, e.g., $H_0 : \mu_1 = \cdots = \mu_k$. In this scenario, the chance that we make a Type 1 error—i.e., reject H_0 when it is true—is α, which is often set at 5%. There are many complications that arise when we want to test multiple hypotheses; in particular, it becomes much more difficult to control the Type 1 error rate. If you don't conduct multiple hypothesis tests correctly, your Type 1 error rate will likely be much higher than 5%—in other words, there will be a large chance that you will erroneously report a significant result that isn't really there. This is the statistical equivalent of convincting an innocent person of a crime; we want to avoid this error if we can.

To understand this issue with multiple comparisons, let's consider the case where we reject $H_0 : \mu_1 = \cdots = \mu_k$, and now we want to assess which group means are actually different. One option is to simply test every pairwise comparison, i.e., $H_0 : \mu_1 = \mu_2$, $H_0 : \mu_1 = \mu_3$, and so on. Since each comparison has a 95% chance of correctly retaining the null hypothesis when it is true, after m independent tests we have a 0.95^m chance of correctly concluding that there are no significant differences when the null hypothesis is true. As examples, for m=3, 5, and 10, the chance of correctly retaining all of the null hypotheses are 86%, 77% and 60% respectively. For k groups, there are $\binom{k}{2} = \frac{k(k-1)}{2}$ possible pairwise comparisons, so m will get huge pretty quickly, meaning that the chance of not making erroneous conclusions dwindles pretty quickly.

Maybe you rightfully say, "Look, I'm not that dumb—I'm not going to test every single pairwise comparison. I'm only going to look at the ones that are interesting and stand out from the data." For example, maybe you will decide to only compare the biggest and smallest sample means to see if they are truly different. However, when you do this, you are implicitly comparing all of the

sample means to find this interesting pair, thereby falling back into the checking-all-pairwise-comparisons scenario.

Correctly conducting multiple hypothesis tests is an old but still oft-debated area of statistics. Indeed, as we move further into the age of "big data" where fields like genomics can introduce thousands or possibly millions of possible comparisons, this area becomes increasingly important but also increasingly challenging. In what follows, we will introduce two ways that are frequently used to address the multiple hypothesis testing issue: Planned comparisons and corrections for unplanned comparisons. As we will discuss, ideally, these two approaches are combined to best address the issue of multiple comparisons.

Note that for some situations, such as genomics and proteomics, where k is very large, a better goal than trying to keep the chance of making any false claim at only 5% is to reduce the total fraction of positive claims that are false positive. This is called control of the false discovery rate (FDR). The most popular procedure for controlling the false discovery rate when conducting multiple hypothesis testing is the Benjamini-Hochberg procedure. This procedure is widely used in the genomics and medical literatures, and as a result, the original paper ("Controlling the false discovery rate: A practical and powerful approach to multiple testing") is one of the most cited papers in statistics.

12.3 Planned comparisons

Planned comparisons are a useful way to limit the number of hypothesis tests that are conducted in an analysis, thereby limiting (but not eliminating) the multiple comparison issue discussed in the previous section. Planned comparisons are usually conducted under stringent conditions. Here we list those conditions by order of importance:

1. The contrasts are selected *before* looking at the results, i.e., they are planned, not post-hoc (after-the-fact).

2. The tests are ignored if the overall null hypothesis (e.g., $\mu_1 = \cdots = \mu_k$ in

one-way ANOVA) is not rejected.

3. The contrasts are orthogonal (see below). This requirement is often ignored, with relatively minor consequences.

4. The number of planned contrasts is no more than the corresponding degrees of freedom ($k-1$ for one-way ANOVA).

The orthogonality idea is that each contrast should be based on independent information from the other contrasts. To test for orthogonality of two contrasts for which the contrast coefficients are $C_1 \cdots C_k$ and $D_1 \cdots D_k$, compute $\sum_{i=1}^{k}(C_i D_i)$. If the sum is zero, then the contrasts are orthogonal. E.g., if k=3, then $\mu_1 - 0.5\mu_2 - 0.5\mu_3$ is orthogonal to $\mu_2 - \mu_3$, but not to $\mu_1 - \mu_2$ because (1)(0)+(-0.5)(1)+(-0.5)(-1)=0, but (1)(1)+(-0.5)(-1)+(-0.5)(0)=1.5.

To understand the reasoning behind each of the above conditions for planned comparisons, let's consider the consequences of breaking each requirement. First of all, selecting the contrasts before looking at the results makes the analysis more transparent and convincing from a scientific point-of-view. There is also a statistical reason for this first condition: If you construct your contrasts after looking at your experimental results, you will naturally choose the most "interesting" comparisons, such as comparing the biggest and the smallest sample means. As we talked about in the previous section, this means that you are implicitly comparing all of the sample means to find something "interesting", and thus you are not avoiding the multiple comparison issue, even if you only conduct a single follow-up test. The same kind of argument applies to looking at your planned comparisons without first "screening" with the overall p-value.

Using orthogonal contrasts is also required to maintain your Type 1 experimentwise error rate. Orthogonal contrasts prevent the multiple hypotheses from being correlated. This is helpful because, when null hypotheses are correlated, there is a higher chance of rejecting several hypotheses simultaneously, thereby making it more difficult to control the Type 1 error rate.

Finally, the requirement that the number of planned contrasts is no more than the degrees of freedom is very related to the above requirement that the contrasts

are orthogonal. Given k group means, there are only $(k-1)$ many unique orthogonal contrasts (we won't focus on the technical reason why here). Thus, if one uses more than $k-1$ planned contrasts among k groups, then the contrasts will not be orthogonal. In general, the degrees of freedom denotes the unique number of orthogonal contrasts that one can construct when making comparisons among groups.

Many computer packages, including SPSS, assume that for any set of custom hypotheses that you enter you have already checked that these four conditions apply. Therefore, any p-value it gives you is wrong if you have not met these conditions.

It is up to you to make sure that your contrasts meet the conditions of "planned contrasts"; otherwise the computer package will give wrong p-values.

As an example, consider a trial of control vs. two active treatments ($k = 3$). Before running the experiment, we might decide to test if the average population means for the active treatments differs from the control, and if the two active treatments differ from each other. The contrast coefficients are [1, -0.5, -0.5] and [0, 1, -1]. These are planned before running the experiment. We need to realize that we should only examine the contrast p-values if the overall (between-groups, 2 df) F test gives a p-value less than 0.05. The contrasts are orthogonal because (1)(0)+(-0.5)(1)+(-0.5)(-1)=0. Finally, there are only k-1=2 contrasts, so we have not selected too many.

However, even if the contrasts are planned, that does not necessarily fix the inflated Type 1 error issue. It is true that we have fixed the issue for the above simple example: Because all four of the above conditions were met—especially the "screening" with the overall p-value—the Type 1 error rate is actually controlled at the α level (i.e., the Type 1 error rate is indeed α). However, if any of the four conditions are not met, we often have to make additional corrections to bring the Type 1 error rate back to α. It is an extremely common misconception that planned comparisons don't need to be corrected for multiple comparisons—the paper "Planned Hypothesis Tests Are Not Necessarily Exempt From Multiplicity Adjustment" (Frane, 2015) provides a great examination of this misconception—and you should be aware that planned comparisons are not a panacea to the inflated

Type 1 error issue. Ideally, planned comparisons are paired with corrections for multiple comparisons, which we turn to next.

12.4 Corrections for multiple comparisons (planned or unplanned)

What should we do if we want to test more than $k-1$ contrasts, or if we find an interesting difference that was not in our planned contrasts after looking at our results? These are examples of what is variously called unplanned comparisons, multiple comparisons, post-hoc (after-the-fact) comparisons, or data snooping. The answer is that we need to add some sort of penalty to preserve our Type 1 experiment-wise error rate. The penalty can either take the form of requiring a larger difference (g value) before an unplanned test is considered "statistically significant", or using a smaller α value (or equivalently, using a bigger critical F-value or critical t-value).

How big of a penalty to apply is mostly a matter of considering the size of the "family" of comparisons within which you are operating. (Amount of dependence among the contrasts can also have an effect.) For example, if you pick out the biggest and the smallest means to compare, you are implicitly comparing all pairs of means. In the field of probability, the symbol $\binom{a}{b}$ (read a choose b) is used to indicate the number of different groups of size b that can be formed from a set of a objects. The formula is $\binom{a}{b} = \frac{a!}{b!(a-b)!}$ where $a! = a \cdot (a-1) \cdots (1)$ is read "a factorial". The simplification for pairs, $b = 2$, is $\binom{a}{2} = \frac{a!}{2!(a-2)!} = a(a-1)/2$. For example, if we have a factor with 6 levels, there are 6(5)/2=15 different paired comparisons we can make.

Note that these penalized procedures are designed to be applied *without* first looking at the overall p-value.

The simplest, but often overly conservative penalty is the Bonferroni correction. If m is the size of the family of comparisons you are making, the Bonferroni procedure says to reject any post-hoc comparison test(s) if $p \leq \alpha/m$. So for $k = 6$ treatment levels, you can make post-hoc comparisons of all pairs while preserving Type 1 error at 5% if you reject H_0 only when $p \leq \alpha/15 = 0.0033$.

By "conservative," we mean that this procedure is often more stringent than necessary, and using some other valid procedure might show a statistically sig-

nificant result in some cases where the Bonferroni correction shows no statistical significance.

The Bonferroni procedure is completely general and can be use for any set of hypothesis tests. For example, if we want to try all contrasts of the class "compare all pairs and compare the mean of any two groups to any other single group", the size of this class can be computed, and the Bonferroni correction applied. If k=5, there are 10 pairs, and for each of these we can compare the mean of the pair to each of the three other groups, so the family has 10*3+10=40 possible comparisons. Using the Bonferroni correction with m=40 will ensure that you make a false positive claim no more than $100\alpha\%$ of the time.

Another procedure that is valid specifically for comparing pairs is the Tukey procedure. The mathematics will not be discussed here, but the procedure is commonly available, and can be used to compare any and all pairs of group population means after seeing the results. For two-way ANOVA without interaction, the Tukey procedure can be applied to each factor (ignoring or averaging over the other factor). For a $k \times m$ ANOVA with a significant interaction, if the desired contrasts are between arbitrary cells (combinations of levels of the two factors), the Tukey procedure can be applied after reformulating the analysis as a one-way ANOVA with $k \times m$ distinct (arbitrary) levels. The Tukey procedure is more powerful (less conservative) than the corresponding Bonferroni procedure.

It is worth mentioning again that none of these procedures are needed for $k = 2$—there is only one hypothesis that can be conducted for $k = 2$ groups, and these procedures only apply to cases where multiple hypotheses are conducted.

Yet another post-hoc procedure is Dunnett's test. This test makes the appropriate penalty correction for comparing one (control) group to all other groups.

In short, there is a huge number of post-hoc procedures, which implies that there is some lack of consensus on how to best address unplanned comparisons. You should avoid unplanned comparisons if possible, but the whole point of unplanned comparisons is that, well, you did not plan for them. In this class we won't be testing on which post-hoc procedure is most appropriate for a particular dataset, but we have discussed them here for reference for your future career if ever you need them. Ultimately, you should pick the procedure that is most appropriate for your application (for example, if you are focusing on comparing pairs, the Tukey procedure is probably most appropriate). Certainly, it is very bad practice to try as many procedures as needed until you get the answer you want!

The final post-hoc procedure discussed here is the Scheffé procedure. This is a very general but conservative procedure. It is applicable for the family of *all* possible contrasts! One way to express the procedure is to consider the usual uncorrected t-test for a contrast of interest. Square the t-statistic to get an F statistic. Instead of the usual F-critical value for the overall null hypothesis, often written as $F(1-\alpha, k-1, N-k)$, the penalized critical F value for a post-hoc contrast is $(k-1)F(1-\alpha, k-1, N-k)$. Here, N is the total sample size for a one-way ANOVA, and $N-k$ is the degrees of freedom in the estimate of σ^2.

The critical F value for a Scheffé penalized contrast can be obtained as $(k-1) \times \text{qf}(0.95, k-1, N-k)$ in R.

In practice, the Scheffé penalty makes sense when you see an interesting complex post-hoc contrast, and then want to see if you actually have good evidence that it is "real" (statistically significant). Implementing the Scheffé penalty involves computing the contrast estimate (g) and its standard error (SE(g)), and then find $F = (g/\text{SE}(g))^2$ and reject H_0 only if this value exceeds the Scheffé penalized F cutoff value.

When you have both planned and unplanned comparisons (which should be most of the time), it is not worthwhile (re-)examining any planned comparisons that also show up in the list of unplanned comparisons. This is because the unplanned comparisons have a penalty, so if the contrast null hypothesis is rejected as a planned comparison we already know to reject it, whether or not it is rejected on the post-hoc list, and if it is retained as a planned comparison, there is no way it will be rejected when the penalty is added.

Unplanned contrasts should be tested only after applying an appropriate penalty to avoid a high chance of Type 1 error. The most useful post-hoc procedures are Bonferroni, Tukey, and Dunnett.

12.5 Do it in R

12.5.1 Contrasts in one-way ANOVA

Let's return to the emotion experiment example from Chapter 6. To save you some time going back to Chapter 6 and trying to remember what that experiment was: Subjects were assigned to one of three emotion-inducing groups ("Control," "Shame," or "Guilt"), and a 0-10 `cooperation` score was measured. Thus, the three-level explanatory variable is `emotion` and the outcome variable is `cooperation`. In Chapter 6, we performed one-way ANOVA on this dataset. One-way ANOVA tests the following hypothesis: $H_0 : \mu_C = \mu_S = \mu_G$, where μ_C, μ_S, and μ_G denote the population-level `cooperation` score under Control, Shame, and Guilt, respectively. Here is the code to run the one-way ANOVA on this dataset:

```
> onewayModel = aov(cooperation ~ emotion, data = moral)
> summary(onewayModel)
             Df Sum Sq Mean Sq F value Pr(>F)
emotion       2   86.4   43.18   4.495 0.0131 *
Residuals   123 1181.4    9.61
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
```

Because the p-value is relatively small (0.01, which is less than 0.05), we reject the null hypothesis that the three population means are equal. Thus, we conclude that the three population means are *not* equal; however, from this test alone, we do not know *which* population means are not equal to each other. To determine which population means are not equal to each other, we need to perform some follow-up hypotheses.

Intuitively, if we want to know which pairs of population means are not equal to each other, the most natural type of follow-up hypotheses to conduct are all pairwise comparisons. As discussed in this chapter, when we want to conduct all pairwise hypothesis tests, we use the Tukey procedure to correct for multiple hypothesis testing. After a one-way ANOVA analysis is defined using `aov()` (as above), you can use the `TukeyHSD()` function to produce adjusted p-values for all pairwise comparisons:

```
#All pairwise comparisons with Tukey correction
> TukeyHSD(onewayModel)
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
Fit: aov(formula = cooperation ~ emotion, data = moral)

$emotion
                     diff        lwr         upr     p adj
Guilt-Control   1.8937729  0.258733   3.5288128 0.0187923
Shame-Control   0.2905983 -1.317986   1.8991828 0.9037976
Shame-Guilt    -1.6031746 -3.180682 -0.0256668 0.0454727
```

Each row corresponds to a different pairwise null hypothesis. Specifically, the first row corresponds to $H_0 : \mu_G = \mu_C$; the second corresponds to $H_0 : \mu_S = \mu_C$; and the third corresponds to $H_0 : \mu_S = \mu_G$. You can reject any of these null hypotheses at the $\alpha = 0.05$ level when the p-values in the "p adj" column are less than 0.05; these p-values are labeled as "p adj" because they are adjusted via the Tukey correction, and thus issues of multiple testing are no longer a concern for these p-values. From the above, we can conclude that $\mu_G > \mu_C$ and $\mu_G > \mu_S$, but we fail to reject the null hypothesis that $\mu_S = \mu_C$. Note that we don't simply conclude that $\mu_G \neq \mu_C$ and $\mu_G \neq \mu_S$; we used the "diff" column to determine the direction of these population means.

Although the Tukey procedure is the most appropriate multiple-testing correction to use for all pairwise comparisons, other corrections can be used by using the `pairwise.t.test()` function. In this function, you specify the outcome variable as `x`, the categorical (or grouping) variable as `g`, and the type of correction you want as `p.adjust.method`. Specifically, `p.adjust.method` can be set to "none" (which is equivalent to doing many separate independent t-tests), "bonferroni" (which is the most conservative correction), "BH" (which stands for "Benjamini-Hochberg", which we did not discuss in depth in this chapter but is a popular correction to use when there are very many comparisons to be made, as in genetics), and other options (see `help(p.adjust)` for other options, which are outside the scope of this class). As a demonstration, here is the output using no correction and using the Bonferroni correction:

```
#pairwise comparisons with no correction
> pairwise.t.test(x = moral$cooperation, g = moral$emotion, p.
    adjust.method = "none")

  Pairwise comparisons using t tests with pooled SD

data:  moral$cooperation and moral$emotion

       Control Guilt
```

```
Guilt 0.0069  -
Shame 0.6690  0.0174

P value adjustment method: none
#pairwise comparisons with bonferonni correction
> pairwise.t.test(x = moral$cooperation, g = moral$emotion, p.
    adjust.method = "bonferroni")

  Pairwise comparisons using t tests with pooled SD

data:  moral$cooperation and moral$emotion

      Control Guilt
Guilt 0.021   -
Shame 1.000   0.052

P value adjustment method: bonferroni
```

Notice that the p-values with no correction are all strictly smaller than the Tukey p-values, and the Bonferroni p-values are all strictly larger than the Tukey p-values. The p-values with no correction are not valid, in the sense that they do not control for multiple hypothesis testing, and thus we will falsely reject the null hypothesis at a rate greater than 5% if we use these p-values (i.e., our Type 1 error rate will be higher than 5%). Meanwhile, the Bonferroni p-values are valid (i.e., they control the Type 1 error rate to be at or below 5%), but note that, compared to the Tukey p-values, we will *correctly* reject the null hypothesis *less often* (because the p-values are higher), meaning that we have less power. This is why the Bonferroni correction is often considered to be overly conservative - although it guarantees to control the Type 1 error rate, it also sacrifices a lot of statistical power; more statistical power can be gained using more nuanced procedures like the Tukey procedure without sacrificing validity.

There are two other types of follow-up tests that we should discuss:

1. Dunnett's procedure (used when you want to test for all pairwise comparisons between a control group and all other treatment groups).

2. Scheffé's procedure (used when you want to correct for testing multiple contrasts - not just pairwise comparisons).

The most useful function to do these in `R` is available in the package `multcomp`. Up to this point, we haven't had to install any packages in `R`. To install this

package, within the console of R, type `install.packages("multcomp")`. After it is installed, load the library by typing `library(multcomp)` in the console.

The function you use to implement Dunnett's procedure and Scheffé's procedure is the `glht()` function (this function will only be available to you in R after you've loaded the `multcomp` package.) In the `glht()` function, you specify an ANOVA model created by `aov()`, and you also have to (carefully) specify the use of Dunnett's procedure or a set of contrasts you want to test. As an example, let's first turn to Dunnett's procedure, which is particularly relevant for our emotion experiment example (because there is a control group). Here is the code to run Dunnett's procedure:

```
#Be sure to load the multcomp library!
> library(multcomp)
#Dunnett's procedure
> glht.dunnett = glht(model = onewayModel, linfct = mcp(emotion =
   "Dunnett"))
#Obtain the output:
> summary(glht.dunnett)

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = cooperation ~ emotion, data = moral)

Linear Hypotheses:
                      Estimate Std. Error t value Pr(>|t|)
Guilt - Control == 0   1.8938     0.6892   2.748   0.0131 *
Shame - Control == 0   0.2906     0.6780   0.429   0.8734
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
(Adjusted p values reported -- single-step method)
```

Notice that we have to carefully specify an argument called `linfct` using the `mcp()` function in the `multcomp` package. Within the `mcp()` function, you write the name of your explanatory variable (in this case, `emotion`) and set it equal to `"Dunnett"`. Let's look at the above output. Note that the p-values are higher than the p-values we got when no correction was done (which makes sense - we are inflating our p-values in order to account for issues of multiple testing), but they are lower than the p-values we got from Tukey's procedure. This is because Dunnett's

procedure is more powerful than Tukey's procedure for the two hypotheses above ($H_0 : \mu_C = \mu_G$ and $H_0 : \mu_C = \mu_S$); however, as a cost, we give up trying to test the hypothesis $H_0 : \mu_G = \mu_S$. Nonetheless, the interpretation from Dunnett's procedure is the same in this case: We reject the null hypothesis $H_0 : \mu_C = \mu_G$ but fail to reject $H_0 : \mu_C = \mu_S$. Finally, note that Dunnett's procedure will always treat the "control" group as the reference group of your categorical variable (i.e., the first level of that variable). In this example, we were lucky that the Control group was indeed the first level of the `emotion` variable. If the control group is not the first level of your explanatory variable, use the `relevel()` function to set the control group to the first level.

Now let's discuss implementing Scheffé's procedure. So far, we have discussed how to conduct pairwise comparisons in `R`, but sometimes you'll be interested in conducting more complex hypotheses. For example, how can we test if the Shame and Guilt groups (together) are different from the Control group? In other words, we would like to test the hypothesis $H_0 : \mu_C = \frac{\mu_G + \mu_S}{2}$, i.e., $H_0 : \mu_C - \frac{\mu_G}{2} - \frac{\mu_S}{2}$. This corresponds to the contrast $(1, -1/2, -1/2)$. (Importantly, note that the contrast would be different if the levels of the categorical variable, `emotion`, were ordered differently. However, in `R`, by default the levels of a categorical variable are ordered alphabetically - e.g., Control, Guilt, and Shame - which is why we ordered our null hypotheses as such.) In the code below, we input the contrast $(1, -1/2, -1/2)$ in the `glht()` function:

```
#Here we test if the control group is different from the other two
    groups
#This corresponds to the contrast (1, -1/2, -1/2)
#First, run glht():
> glht.fit = glht(model = onewayModel, linfct = mcp(emotion = c(1,
    -1/2, -1/2)))
> summary(glht.fit)

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = cooperation ~ emotion, data = moral)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  -1.0922     0.5973  -1.828   0.0699 .
---
```

```
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
        0.1         1
(Adjusted p values reported -- single-step method)
```

The above p-value is *not* Scheffé's procedure; it does not account for possible multiple testing. Computing the p-value from Scheffe's procedure is somewhat complicated. First, you must define the F-statistic, which is the square of the t-statistic provided by the `glht()` function. Then, you must compute the p-value using the `pf()` function (i.e., the CDF of the F distribution). For the sake of 36-309, we do not expect you to implement Scheffé's procedure from scratch, but we wanted to provide the necessary code for reference:

```
#Define the number of groups (in this case 3)
> k = 3
#Define the number of subjects
> N = nrow(moral)
#To compute the p-value from Scheffe's procedure,
#We first need to compute the standardized F statistic,
#which is defined as t^2/(g-1):
> fStat = (summary(glht.fit)$test$tstat)^2
#Then, Scheffe's p-value is:
> 1 - pf(fStat/(k-1), k-1, N-k)
        1
0.1921998
```

Note that the p-value from Scheffé's procedure is much larger compared to the original p-value (0.19 versus 0.07). A benefit of Scheffé's procedure is that it can be used for *as many* contrasts as you want, so it's especially valuable when you want to test many complicated contrasts; however, this is at the expense of sacrificing a lot of statistical power.

12.5.2 Contrasts for Two-way ANOVA

In this section, it's worth considering two scenarios:

1. Two-way ANOVA without interaction
2. Two-way ANOVA with interaction

Contrasts in two-way (between-subjects) ANOVA *without* interaction work just like in one-way ANOVA, but with separate contrasts for each factor. As a demonstration, let's quickly revisit the math example from Chapter 8. In this example,

there were two categorical explanatory variables `courses` ("Algebra only", "Algebra + Geometry", and "through Calculus") and `activity` ("Yes" or "No" for participation in extracurricular activities), and a quantitative outcome `score` (a test score). The interaction between `activity` and `score` was non-significant, so we fit a two-way ANOVA model without an interaction:

```
> twoWay.math = aov(score ~ courses + activity, data = mathAct)
> summary(twoWay.math)
             Df Sum Sq Mean Sq F value   Pr(>F)
courses       2  15619    7809  320.00  < 2e-16 ***
activity      1    517     517   21.17 4.84e-06 ***
Residuals   857  20914      24
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
```

Because the p-value for courses is very small, we reject the null hypothesis that the three population means for `score` are equal (i.e., we reject $H_0 : \mu_A = \mu_{AG} = \mu_C$, where μ_A, μ_{AG}, and μ_C correspond to the population means for "Algebra only", "Algebra + Geometry", and "through Calculus", respectively). However, from this test alone, we don't know *which* of these population means are not equal to each other, which suggests that we should do some follow-up tests. Meanwhile, because `activity` only has two levels, we already know in which direction the corresponding population means are not equal to each other by examining the appropriate outcome sample means (in Chapter 8 we found that $\mu_Y > \mu_N$, where μ_Y and μ_N correspond to the population means under the "Yes" and "No" conditions). Thus, technically no follow-up tests are necessary for the `activity` variable. Nonetheless, as a demonstration, we'll show how to do follow-up tests for both factors.

It's very straightforward to conduct all pairwise comparisons for each factor using the `TukeyHSD()` function; you simply input the `aov()` model into this function:

```
#all pairwise comparisons via the Tukey procedure
> TukeyHSD(twoWay.math)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = score ~ courses + activity, data = mathAct)

$courses
                  diff       lwr       upr p adj
algGeom-alg   4.595334  3.474940  5.715729     0
```

```
calc-alg      15.120534 13.655444 16.585624     0
calc-algGeom 10.525200  9.370969 11.679431     0

$activity
           diff     lwr      upr   p adj
yes-no 1.572066 0.89538 2.248752 5.9e-06
```

From this, we can conclude that $\mu_C > \mu_{AG} > \mu_A$ (we indeed need all three lines of the `courses` table to make this conclusion). Meanwhile, we also see from the `activity` table that $\mu_Y > \mu_N$. Note that the Tukey p-value for `activity` is slightly higher than the original p-value we saw for the two-way ANOVA analysis above. Because `activity` only had two levels and thus there was technically no need for follow-up tests, we would recommend reporting the original p-value of $4.84 * 10^{-6}$.

You can also use the `glht()` function to test more complex contrasts. For example, the following code simultaneously tests the hypotheses $H_0 : \mu_C = \frac{\mu_A + \mu_{AG}}{2}$ and $H_0 : \mu_Y = \mu_N$:

```
#Test average of alg and algGeom against calc,
#while also testing yes versus no
> glht.fit.math = glht(model = twoWay.math,
+     linfct = mcp(courses = c(-1/2, -1/2, 1), activity = c(-1,1))
   )
> summary(glht.fit.math)

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = score ~ courses + activity, data = mathAct)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  12.5247     0.5127  24.431  < 1e-10 ***
2 == 0   1.6006     0.3479   4.601 9.68e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
(Adjusted p values reported -- single-step method)
```

Thus, from the above output, we can conclude that $\mu_C > \frac{\mu_A + \mu_{AG}}{2}$.

Now let's consider the case where we have a two-way ANOVA *with* interaction.

We'll return to another example from Chapter 8 - the car noise example, where there were two explanatory variables (`SIZE` and `TYPE`) and an outcome variable `NOISE`:

```
#Run two-way ANOVA
> twoWay.car = aov(NOISE ~ SIZE * TYPE, data = carNoise)
#Examine output
> summary(twoWay.car)
             Df Sum Sq Mean Sq F value   Pr(>F)
SIZE          2  26051   13026 199.119  < 2e-16 ***
TYPE          1   1056    1056  16.146 0.000363 ***
SIZE:TYPE     2    804     402   6.146 0.005792 **
Residuals    30   1963      65
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1          1
```

We see that the interaction is significant. Follow-up tests can become complicated very quickly when an interactions are present. In general, there are two main types of follow-up tests that are conducted for two-way ANOVA analyses when there is a significant interaction:

1. Comparing pairwise tests for one factor while holding another factor fixed.
2. Prespecified complex contrasts that are of interest.

To demonstrate the first type of follow-up test, let's look at output from the `TukeyHSD()` function for this interactive model:

```
> TukeyHSD(twoWay.car)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = NOISE ~ SIZE * TYPE, data = carNoise)

$SIZE
                  diff        lwr       upr     p adj
medium-large 61.250000  53.10983 69.390167 0.0000000
small-large  51.666667  43.52650 59.806833 0.0000000
small-medium -9.583333 -17.72350 -1.443167 0.0183324

$TYPE
                   diff       lwr       upr     p adj
standard-octel 10.83333 5.327328 16.33934 0.0003631
```

```
$`SIZE:TYPE`
                                        diff         lwr         upr
      p adj
medium:octel-large:octel         51.6666667  37.463511  65.869823
   0.0000000
small:octel-large:octel          52.5000000  38.296844  66.703156
   0.0000000
large:standard-large:octel        5.0000000  -9.203156  19.203156
   0.8890358
medium:standard-large:octel      75.8333333  61.630177  90.036489
   0.0000000
small:standard-large:octel       55.8333333  41.630177  70.036489
   0.0000000
small:octel-medium:octel          0.8333333 -13.369823  15.036489
   0.9999720
large:standard-medium:octel     -46.6666667 -60.869823 -32.463511
   0.0000000
medium:standard-medium:octel     24.1666667   9.963511  38.369823
   0.0001909
small:standard-medium:octel       4.1666667 -10.036489  18.369823
   0.9454142
large:standard-small:octel      -47.5000000 -61.703156 -33.296844
   0.0000000
medium:standard-small:octel      23.3333333   9.130177  37.536489
   0.0003130
small:standard-small:octel        3.3333333 -10.869823  17.536489
   0.9787622
medium:standard-large:standard   70.8333333  56.630177  85.036489
   0.0000000
small:standard-large:standard    50.8333333  36.630177  65.036489
   0.0000000
small:standard-medium:standard -20.0000000 -34.203156  -5.796844
   0.0022033
```

Because the interaction between `SIZE` and `TYPE` is significant, it's really only appropriate to interpret the `SIZE:TYPE` table (in the same way that we should refrain from interpreting the main effects in the two-way ANOVA model when there is a significant interaction). Unless you're a computer or super-human, looking at the `SIZE:TYPE` table is probably overwhelming at first. Let's look at the first row: This row corresponds to the pairwise hypothesis $H_0 : \mu_{MO} = \mu_{LO}$, i.e., the population mean for `NOISE` under the joint condition (Medium, Octel) compared to that under the joint condition (Large, Octel). The above output can be much more interpretable if you focus on pairwise comparisons where one of the factors

is fixed. For example, in the aforementioned comparison in the first row, the TYPE variable is fixed to Octel. To make the output easier to read, let's look at only the comparisons where the TYPE variable is fixed:

```
                                          diff        lwr        upr
        p adj
medium:octel-large:octel         51.6666667  37.463511  65.869823
    0.0000000
small:octel-large:octel          52.5000000  38.296844  66.703156
    0.0000000
small:octel-medium:octel          0.8333333 -13.369823  15.036489
    0.9999720

medium:standard-large:standard   70.8333333  56.630177  85.036489
    0.0000000
small:standard-large:standard    50.8333333  36.630177  65.036489
    0.0000000
small:standard-medium:standard -20.0000000 -34.203156  -5.796844
    0.0022033
```

Hopefully this output is much more manageable. We've also added a space in between the "octel" comparisons and the "standard" comparisons for even easier readability. The first three rows tell us that, for Octel, $\mu_M > \mu_L$ and $\mu_S > \mu_L$ but $\mu_S = \mu_M$. Meanwhile, the last three rows tells us that, for Standard, $\mu_M > \mu_S > \mu_L$. Another way to interpret this output is that NOISE under the small and medium conditions appear to be statistically significant under Standard but not Octel.

Finally, it may be of interest to test complex hypotheses such as $H_0 : \mu_L = \frac{\mu_S+\mu_M}{2}$ for "standard" and "octel" *separately*. The most straightforward way to do this is to first create two subsets of the data, one containing "standard" and one containing "octel":

```
#standard data
> car.standard = subset(carNoise, TYPE == "standard")
#octel data
> car.octel = subset(carNoise, TYPE == "octel")
```

run one-way ANOVA (treating SIZE as the explanatory variable) for both of these subsets:

```
oneway.standard = aov(NOISE ~ SIZE, data = car.standard)
oneway.octel = aov(NOISE ~ SIZE, data = car.octel)
```

and then test the hypothesis $H_0 : \mu_L = \frac{\mu_S + \mu_M}{2}$ for each of these one-way ANOVAs:

```
> glht.fit.standard = glht(model = oneway.standard, linfct = mcp(
    SIZE = c(1, -1/2, -1/2)))
> glht.fit.octel = glht(model = oneway.octel, linfct = mcp(SIZE =
    c(1, -1/2, -1/2)))
```

Here is the resulting output from these analyses:

```
> summary(glht.fit.standard)

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = NOISE ~ SIZE, data = car.standard)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  -60.833      5.231  -11.63 6.63e-09 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
(Adjusted p values reported -- single-step method)

> summary(glht.fit.octel)

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = NOISE ~ SIZE, data = car.octel)

Linear Hypotheses:
       Estimate Std. Error t value Pr(>|t|)
1 == 0  -52.083      2.312  -22.52  5.6e-13 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1
(Adjusted p values reported -- single-step method)
```

From this output, we would conclude that we reject the null hypothesis $H_0 : \mu_L = \frac{\mu_S + \mu_M}{2}$ for "standard" *and* "octel". Note that it would be inappropriate

to test this hypothesis overall across the entire dataset, in the same way that intepreting main effects is inappropriate in the presence of interactions. Instead, it is more appropriate to interpret the main effects of one factor while holding the level of another factor fixed, as we have done here.

Chapter 13

Within-Subjects Designs

ANOVA must be modified to take correlated errors into account when multiple measurements are made for each subject.

13.1 Overview of within-subjects designs

Any categorical variable for which each subject experiences multiple levels of the variable is called a **within-subjects factor**. These levels could be different "treatments", or they may be different outcome measurements for the same treatment (e.g., height and weight as outcomes for each subject), or they may be repetitions of the same outcome over time (or space) for each subject. In a broad sense, the term **repeated measure** is a synonym for a within-subject factor, although often the term repeated measures analysis is used in a narrower sense to indicate the specific set of analyses discussed in Section 13.5.

In contrast to a within-subjects factor, any factor for which each subject experiences only one of the levels is a **between-subjects factor**. Any experiment that has at least one within-subjects factor is said to use a **within-subjects design**, while an experiment that uses only between-subjects factor(s) is called a **between-subjects design**. Often the term **mixed design** or **mixed within- and between-subjects design** is used when there is at least one within-subjects factor and at least one between-subjects factor in the same experiment. (Be careful to distinguish this from the so-called mixed models of Chapter 14.) All of the experiments discussed in the preceding chapters are between-subjects designs; in

all of the analyses we've discussed so far, subjects only received one treatment level (e.g., for any given subject, we only had a "treatment" measurement or a "control" measurement - never both).

Please do not confuse the terms between-groups and within-groups with the terms between-subjects and within-subjects. The first two terms, which we first encountered in the ANOVA chapter, are names of specific SS and MS components and are named because of how we define the deviations that are summed and squared to compute SS. In contrast, the terms within-subjects and between-subjects refer to experimental designs that either do or do not take multiple measurements on each subject.

When a within-subjects factor is used in an experiment, measurements will not be independent, because multiple measurements come from the same subject. Thus, new methods are needed that allow for dependence among measurements. (See Section 5.2.8 to review the independent errors assumption.)

Why would we want to take multiple measurements on the same subject(s)? There are two basic reasons. First, our primary interest may be to study the change of an outcome over time, e.g., a learning effect; in this scenario, we'll need to take multiple measurements over time. Second, studying multiple outcomes for each subject allows each subject to be his or her own "control", i.e., we can effectively remove subject-to-subject variation when studying different treatments. This reduced variability directly increases power, often dramatically.

These two reasons are very important advantages for within-subjects designs, making them very popular. However, there are also two major reasons why someone may not use a within-subjects design. First, it may be impossible to give multiple treatments to a single subject. For example, in studies comparing surgery vs. drug treatment of a disease, subjects generally receive one or the other treatment, not both. The second reason is that there may be concerns about confounding, which we elaborate on below.

The confounding problem of within-subjects designs is an important concern. Consider the case of three kinds of hints for solving a logic problem. Let's take the time till solution as the outcome measure. If each subject first sees problem 1 with hint 1, then problem 2 with hint 2, then problem 3 with hint 3, then we will probably have two major difficulties. First, the effects of the hints **carry-over** from each trial to the next. The truth is that problem 2 is solved when the subject has been exposed to two hints, and problem 3 when the subject has been exposed to all three hints. The effect of hint type (the main focus of inference) is

confounded with the cumulative effects of prior hints.

The carry-over effect is generally dealt with by allowing sufficient time between trials to "wash out" the effects of previous trials. That is often quite effective, e.g., when the treatments are drugs, and we can wait until the previous drug leaves the system before studying the next drug. But in cases such as the hint study, this approach may not be effective or may take too much time.

The other, partially overlapping, source of confounding is the fact that when testing hint 2, the subject has already had practice with problem 1, and when testing hint three she has already had practice with problems 1 and 2. This is the **learning effect**.

The learning effect can be dealt with effectively by using **counterbalancing**. The carryover effect is also partially corrected by counterbalancing. Counterbalancing in this experiment could take the form of collecting subjects in groups of six, then randomizing the group to all possible orderings of the hints (123, 132, 213, 231, 312, 321). Then, because each hint is evenly tested at all points along the learning curve, any learning effects would "balance out" across the three hint types, removing the confounding. (It would probably also be a good idea to randomize the order of the problem presentation in this study.) This is the same justification that's used for randomized experiments in general: By randomizing subjects, they will (on average) be comparable to each other. In other words, we'll have higher internal validity, as discussed in Chapter 7. In this example, the hints are balanced out, just like explanatory variables of subjects are balanced in randomized experiments.

You need to know how to distinguish within-subjects from between-subjects factors. Within-subjects designs have the advantage of more power and allow observation of change over time. The main disadvantages are possible carryover effects and confounding, which can often be addressed with experimental design strategies like counterbalancing and waiting for carryover effects to 'wash out' over time.

13.2 Multivariate distributions

Some of the analyses in this chapter require you to think about **multivariate distributions**. Up to this point, we have dealt with outcomes that, among all subjects that have the same given combination of explanatory variables, are assumed to follow a univariate Normal distribution. By univariate, we mean that there is a single mean μ and single variance σ^2 for the Normal distribution placed on the outcomes Y. A convenient consequence of assuming a univariate Normal distribution for the outcomes is that we can visualize the outcome distribution using a simple two-dimensional plot with an x- and y-axis. The population-level distribution can be visualized as a standard bell-shaped curve with the value of the outcome on the x-axis and the relative frequency of that value on the y-axis (e.g., Figure 3.2). Meanwhile, the sample distribution can be visualized with a histogram (e.g., Figure 4.1).

Most importantly, a univariate distribution makes sense for a between-subjects design because there is only a single measurement for each subject. Meanwhile, for within-subjects designs, we have *multiple* measurements per subject, and these measurements will be correlated. These issues are accounted for using multivariate distributions.

For example, to represent the outcomes of two treatments for each subject, we need a so-called bivariate distribution. To produce a graphical representation of a bivariate distribution, we use the two axes (say, y1 and y2) on a sheet of paper for the two different outcome values, and therefore each pair of outcomes corresponds to a point on the paper with y1 equal to the first outcome and y2 equal to the second outcome. Then the third dimension (coming up out of the paper) represents how likely each combination of outcome is. For a bivariate Normal distribution, this is like a real bell sitting on the paper (rather than the silhouette of a bell that we have been using so far).

Using an analogy between a bivariate distribution and a mountain peak, we can represent a bivariate distribution in 2-dimensions using a figure corresponding to a topographic map. Figure 13.1 shows the center and the contours of one particular bivariate Normal distribution. This distribution has a negative correlation between the two values for each subject, so the distribution is more like a bell squished along a diagonal line from the upper left to the lower right. If we have no correlation between the two values for each subject, we get a nice round bell. You can see that an outcome like $Y_1 = 2,\ Y_2 = 6$ is fairly likely, while one like $Y_1 = 6,\ Y_2 = 2$

is quite unlikely. (By the way, bivariate distributions can have shapes other than Normal.)

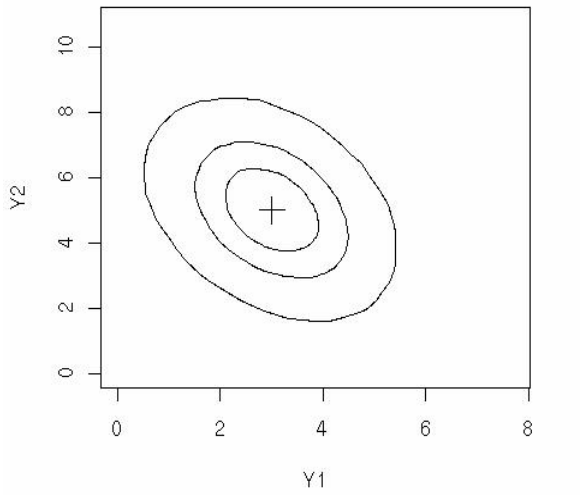


Figure 13.1: Contours enclosing 1/3, 2/3 and 95% of a bivariate Normal distribution with a negative covariance.

The idea of the bivariate distribution can easily be extended to more than two dimensions, but it is of course much harder to visualize. A multivariate distribution with k-dimensions has a k-length vector of means (e.g., a mean for each of the k outcomes). It also has a $k \times k$ dimensional matrix (i.e., a rectangular array of numbers) representing the variances of the individual variables, and all of their paired covariances (see Section 3.6.1).

For example, a 3-dimensional multivariate distribution representing the outcomes of three treatments in a within-subjects experiment would be characterized by a mean vector, e.g.,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \tag{13.1}$$

and a variance-covariance matrix, e.g.,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \gamma_{1,2} & \gamma_{1,3} \\ \gamma_{1,2} & \sigma_2^2 & \gamma_{2,3} \\ \gamma_{1,3} & \gamma_{2,3} & \sigma_3^2 \end{bmatrix} \tag{13.2}$$

Here we are using $\gamma_{i,j}$ to represent the covariance of variable Y_i with Y_j, while σ_i^2 represents the variance of variable Y_i.

Sometimes, as an alternative to a variance-covariance matrix, people use a variance vector, e.g.,

$$\sigma^2 = \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \sigma_3^2 \end{bmatrix},$$

and a correlation matrix, e.g.,

$$\text{Corr} = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} \\ \rho_{1,2} & 1 & \rho_{2,3} \\ \rho_{1,3} & \rho_{2,3} & 1 \end{bmatrix}.$$

Here we are using $\rho_{i,j}$ to represent the correlation of variable Y_i with Y_j.

If the distribution is also Normal, we could write the distribution as $Y \sim \mathcal{N}_k(\mu, \Sigma)$. Here, the notation $N_k(\cdot, \cdot)$ denotes that we have a Normal distribution of k dimensions. Notice that this notation is very similar to what we've seen throughout this book - we have a Normal distribution with a mean and variance - but now the mean is a k-dimensional vector and the variance is an $k \times k$ variance-covariance matrix.

Something nice about Normal distributions is that if $Y \sim \mathcal{N}_k(\mu, \Sigma)$ (where Y is a set of k outcomes, $Y_1, \ldots, Y_k$), each Y_i also has a univariate Normal distribution. For example, let's say that $Y \sim \mathcal{N}_3(\mu, \Sigma)$, where μ and Σ are given by (13.1) and (13.2), respectively. Then, $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_1, \sigma_2^2)$, and $Y_3 \sim N(\mu_1, \sigma_3^2)$, i.e., each outcome has its own familiar-looking Normal distribution. However, the additional information we're given by writing a single multivariate distribution $Y \sim \mathcal{N}_3(\mu, \Sigma)$ instead of three univariate distributions is that Y_1, Y_2, and Y_3 are correlated and not independent.

13.3 Example and alternate approaches

Consider an example related to the disease osteoarthritis. (This comes from the OzDASL web site, OzDASL. For educational purposes, I slightly altered the data, which can be found in both the tall and wide formats on the data web page of this book: osteoTall.sav and osteoWide.sav.) Osteoarthritis is a mechanical degeneration of joint surfaces causing pain, swelling and loss of joint function in one or more joints. Physiotherapists treat the affected joints to increase the range of movement (ROM). In this study 10 subjects were each given a trial of therapy with two treatments, TENS (an electric nerve stimulation) and short wave diathermy (a heat treatment), plus control.

Notice that we have a three-level categorical explanatory variable (the treatment) and a quantitative outocme (ROM). Thus, we may be tempted to use a simple one-way ANOVA for this analysis, as we learned in Chapter 6. However, we cannot perform ordinary (between-subjects) one-way ANOVA for this experiment because each subject was exposed to all three treatments, so the errors (ROM outcomes for a given subject for all three treatments minus the population means of outcome for those treatment) are almost surely correlated, rather than independent. Possible appropriate analyses fall into four categories:

1. Response simplification: This type of analysis takes the multiple measurements of each subject and summarizes them into a single measurement, which allows us to use univariate distributions (and thus standard techniques) again. For example, we may define the difference between the control response and the treatment response as a single outcome. (We've actually already seen special cases of this in the class, where we defined a single outcome as the difference between response before treatment and after treatment.) If the within-subjects factor is the only factor, an appropriate test is a one-sample t-test for the difference outcome, with the null hypothesis being a zero mean difference. In cases where the within-subjects factor is repeated measures over time or space and there is a second, between-subjects factor, the effects of the between-subjects factor on the outcome can be studied by taking the mean of all of the outcomes for each subject and using standard, between-subjects one-way ANOVA. This approach does not fully utilize the available information. Often it cannot answer some interesting questions.

2. Treat the several responses on one subject as a single "multivariate" response and model the correlation between the components of that response.

The main statistics are now matrices rather than individual numbers. This approach corresponds to results labeled "multivariate" under "repeated measures ANOVA" for most statistical packages.

3. Treat each response as a separate (univariate) observation, and treat "subject" as a (random) blocking factor. This corresponds to within-subjects ANOVA with subject included as a random factor and with no interaction in the model. It also corresponds to the "univariate" output under "repeated measures". In this form, there are assumptions about the nature of the within-subject correlation that are often not met. To use the univariate approach when its assumptions are not met, it is common to use some approximate correction (to the degrees of freedom) to compensate for a shifted null sampling distribution.

4. Treat each measurement as univariate, but explicitly model the correlations. This is a more modern univariate approach called "mixed models" that subsumes a variety of models in a single unified approach, is very flexible in modeling correlations, and often has improved interpretability. As opposed to "classical repeated measures analysis" (approaches 2 and 3), mixed models can accommodate missing data instead of dropping all data from every subject who is missing one or more measurements, and it accommodates unequal and/or irregular spacing of repeated measurements.

In what follows, we will discuss the first three approaches in detail. The fourth approach will be our focus for Chapter 14.

13.4 Paired t-test

The paired t-test uses response simplification to handle the correlated errors. It only works with two treatments, so we will ignore the diathermy treatment in our osteoarthritis example for this section. The simplification here is to compute the difference between the two outcomes for each subject. Then there is only one "outcome" for each subject, and there is no longer any concern about correlated errors. (The subtraction is part of the paired t-test, so you don't need to do it yourself.)

In `R`, you can use "wide" or "tall" formats of the data to perform the paired t-test. The tall form has one outcome per row, so it has many rows. The wide

form has one subject per row with two or more outcomes per row (necessitating two or more outcome columns). So far in this book we have only focused on "tall" formats, but we introduce "wide" formats here because they are common in within-subjects designs, and it's slightly easier to perform the paired t-test and other repeated measures analyses with wide-formatted data.

The paired t-test uses a one-sample t-test on the single column of computed differences. Although we have not discussed the one-sample t-test, it is actually a simplified version of the independent-sample t-test from Chapter 5. The point of the independent-sample t-test from Chapter 5 is to test if the mean difference between two sets of measurements (e.g., treatment and control) is significantly non-zero; the point of the one-sample t-test is to test if the mean for a single set of measurements (e.g., the treatment minus control difference for each subject) is significantly non-zero.

For the paired t-test, we have (for each subject) the difference in outcome between two treatments, and the sample mean of these differences serves as our estimate of the true mean for the set of difference measurements. Then, we can compute the standard error for that estimate as $\sqrt{\hat{\sigma}^2/n}$, where $\hat{\sigma}^2$ is the sample variance of the difference measurements and n is the number of subjects. Then, as in Chapter 5, we can construct the t-statistic as the estimate divided by the SE of the estimate. Under the null hypothesis that the population mean difference is zero, this t-statistic will follow a t-distribution with $n - 1$ df.

To implement a paired t-test in `R`, you use the `t.test()` function (just like we did in Chapter 5), but with `paired = TRUE` within this function. If you use wide-format data, within the `t.test()` function, you set `x` equal to one column (i.e., one outcome) and `y` equal to another column. If you use tall-format data, you can use the equation syntax (e.g., `y~x`) within the `t.test()` function, but your dataset needs to be ordered by subject (which is almost always the case for tall-format data, but not always!) As a demonstration, here's how we can implement the paired t-test to compare control to TENS ROM for wide-format data and tall-format data:

```
#wide-format data t-test
> t.test(x = osteoWide$control, y = osteoWide$TENS, paired = TRUE)
```

```
3
4   Paired t-test
5
6 data:  osteoWide$control and osteoWide$TENS
7 t = 2.4395, df = 9, p-value = 0.0374
8 alternative hypothesis: true difference in means is not equal to 0
9 95 percent confidence interval:
10   1.286408 34.113592
11 sample estimates:
12 mean of the differences
13                    17.7
14
15 #tall-format data t-test
16 > t.test(ROM ~ rx, paired = TRUE,
17 +         data = subset(osteoTall, rx != "diathermy"))
18
19   Paired t-test
20
21 data:  ROM by rx
22 t = 2.4395, df = 9, p-value = 0.0374
23 alternative hypothesis: true difference in means is not equal to 0
24 95 percent confidence interval:
25   1.286408 34.113592
26 sample estimates:
27 mean of the differences
28                    17.7
29
```

The t-test gives the same results regardless of whether we input the wide-format or tall-format data, which is reassuring. The last line of the output tells us that our point estimate of the difference in population means for ROM between control and TENS is 17.70 with control being higher (the direction of subtraction is `x - y`, and we set `x` equal to the control group and `y` equal to the TENS group). We are 95% confident that the true reduction in ROM caused by TENS relative to the control is between 1.29 and 34.11, so it may be very small or rather large. The t-statistic of 2.44 will follow the t-distribution with 9 df if the null hypothesis is true and the assumptions are met. This leads to a p-value of 0.037, so we reject the null hypothesis and conclude that TENS reduces range of motion.

For comparison, let's see what happens when we run an independent two-sample t-test (i.e., the method we learned from Chapter 5):

```
1 #INCORRECT two-sample t-test
2 > t.test(x = osteoWide$control, y = osteoWide$TENS, var.equal =
```

```
   TRUE)

  Two Sample t-test

data:  osteoWide$control and osteoWide$TENS
t = 1.6196, df = 18, p-value = 0.1227
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.25951 40.65951
sample estimates:
mean of x mean of y
    101.9      84.2
```

From this analysis, we get a p-value of 0.123, leading to the (probably) incorrect conclusion that the two treatments both have the same population mean of ROM. In terms of implementation in R, this comes down to the simple mistake of forgetting to write `paired = TRUE` within the `t.test()` function!

For educational purposes, it is worth noting that it is possible to get the same correct results in this case (or other one-factor within-subjects experiments) by performing a two-way ANOVA in which "subject" is the other factor (besides treatment). Before looking at the results we need to note several important facts.

A factor is said to be a **fixed factor** if the levels used are the same levels you would use if you repeated the experiment. Treatments are generally fixed factors. A factor is said to be a **random factor** if a different set of levels would be used if you repeated the experiment. Subject is a random factor because if you would repeat the experiment, you would use a different set of subjects. Certain types of blocking factors are also random factors.

The reason that we want to use subject as a factor is that it is reasonable to consider that some subjects will have a high outcome for all treatments and others a low outcome for all treatments. Then it may be true that the errors relative to the overall subject mean are uncorrelated across the k treatments given to a single subject. But if we use both treatment and subject as factors, then each combination of treatment and subject has only one outcome. In this case, we have zero degrees of freedom for the within-subjects (error) SS. The usual solution is to use the interaction MS in place of the error MS in forming the F test for the treatment effect. Based on the formula for expected MS of an interaction (see Section 11.4), we can see that the interaction MS is equal to the error MS if there is no interaction and larger otherwise. Therefore if the assumption of no interaction is correct (i.e., treatment effects are similar for all subjects) then we

get the "correct" p-value, and if there really is an interaction, we get too small of an F value (too large of a p-value), so the test is conservative, which means that it may give excess Type 2 errors, but won't give excess Type 1 errors.

Here are the results we get when we run two-way ANOVA on these data for the control and TENS groups:

```
#two-way ANOVA analysis:
> summary(aov(ROM ~ rx + subject,
+             data = subset(osteoTall, rx != "diathermy")))
            Df Sum Sq Mean Sq F value Pr(>F)
rx           1   1566  1566.5   5.951 0.0374 *
subject      9   8379   931.1   3.537 0.0369 *
Residuals    9   2369   263.2
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1          1
```

The null hypothesis of main interest here is that the two treatment population means are equal, and that is tested and rejected on the "rx" row (note that the p-value is the same as the paired t-test analysis!) The null hypothesis for the random subject effect is that the population variance of the subject-to-subject means (of the two treatments) is zero.

Now let's look at what an incorrect one-way ANOVA analysis would give us (i.e., if we threw out the `subject` factor):

```
#INCORRECT one-way ANOVA analysis:
> summary(aov(ROM ~ rx,
+             data = subset(osteoTall, rx != "diathermy")))
            Df Sum Sq Mean Sq F value Pr(>F)
rx           1   1566  1566.5   2.623  0.123
Residuals   18  10749   597.1
```

Note that the above p-value is the same as the incorrect two-sample t-test p-value.

The key observation from this output is that the treatment (rx) SS and MS are the same in both of these ANOVAs, while the sum of the subject SS and residual SS of the correct two-way ANOVA is equal to the residual SS of the incorrect one-way ANOVA. This is a decomposition of the four sources of error (see Section 7.5) that contribute to σ^2, which is estimated by SS_{within} in the one-way ANOVA. In the two-way ANOVA analysis, the subject-to-subject variability is estimated to be 931.1, and the remaining three sources contribute 263.2 (on the variance scale).

This smaller three-source error MS is the denominator for the numerator (rx) MS for the F-statistic of the treatment effect. Therefore we get a larger F-statistic and more power when we use a within-subjects design.

A two-level one-way within-subjects experiment can equivalently be analyzed by a paired t-test or a two-way ANOVA with a random subject factor. The latter also applies to more than two levels. The extra power comes from mathematically removing the subject-to-subject component of the underlying variance (σ^2).

Although in practice you should use a paired t-test with the `t.test()` function, this connection to ANOVA should (hopefully) solidify the sources of variation that exist in a within-subjects design. Note, however, that the paired t-test is only appropriate when there are two treatment groups. In this example, there are actually three treatment groups, so now we turn to how to conduct a one-way repeated measures analysis that accounts for all three treatment groups.

13.5 One-way Repeated Measures Analysis

Although repeated measures analysis is a very general term for any study in which multiple measurements are made on the same subject, there is a narrow sense of repeated measures analysis which is discussed in this section and the next section. This is a set of specific analysis methods commonly used in social sciences, but less commonly in other fields where alternatives such as mixed models tend to be used (which will be discussed in the next chapter).

This section discusses k-level ($k \geq 2$) one-way within-subjects ANOVA using repeated measures in the narrow sense. This includes the second and third approaches discussed in the introduction to this chapter. The next section discusses mixed within/between subjects two-way ANOVA.

First we need to discuss the assumptions of repeated measures analysis. One-way repeated measures analyses assume a Normal distribution of the outcome for each level of the within-subjects factor. The errors are assumed to be uncorrelated between subjects, while the multiple measurements within a subject are assumed to be correlated. The last assumption is that a technical condition called **sphericity**

is met. Although the technical condition is difficult to understand, there is a simpler condition that is nearly equivalent: compound symmetry. **Compound symmetry** indicates that all of the outcome variances between subjects are equal and all of the covariances (and correlations) within subjects are equal. For example, let us say that each subject has three measurements Y_1, Y_2, and Y_3, whose variances are all σ^2 and whose covariances are ρ. Then, the *covariance matrix* for these three measurements can be written as:

$$\begin{pmatrix} \sigma^2 & \rho & \rho \\ \rho & \sigma^2 & \rho \\ \rho & \rho & \sigma^2 \end{pmatrix} \tag{13.3}$$

The above pattern is a visual of compound symmetry; we saw a similar visual when we briefly discussed covariance matrices in Chapter 3, specifically in Table 3.6. This variance-covariance pattern is seen fairly often when there are several different treatments, but is unlikely when there are multiple measurements over time, in which case adjacent times are usually more highly correlated than distant times (meaning that the correlations within subjects should not be treated as equal). For example, we may believe that Y_1 and Y_2 are more correlated than Y_1 and Y_3 if Y_1 and Y_2 are measured more closely together in time.

When conducting repeated measures ANOVA, it is customary to first assess if the sphericity assumption holds; Mauchly's test of sphericity is the most common way to assess the sphericity assumption. Like other tests of assumptions (e.g., Levene's test of equal variance), the null hypothesis is that there is no assumption violation (here, that the variance-covariance structure is consistent with sphericity), so a large (>0.05) p-value is good, indicating no problem with the assumption. Unfortunately, the sphericity test is not very reliable, often exhibiting low power in small samples and exhibiting over-sensitivity to mild violations of the Normality assumption in large samples. It is worth knowing that the sphericity assumption cannot be violated with $k = 2$ levels of treatment (because there is only a single covariance between the two measures, so there is nothing for it to be possibly unequal to).

To run repeated measures ANOVA in `R`, you use the anova_test() function, which is available in the `rstatix` package. This function requires your data to be in a *tall* format, where there is a single column for the outcome and multiple rows per subject. When using this function, there are four arguments you have to specify:

- `dv`: The name of the outcome variable (the "dependent variable")

- `wid`: The name of the subject ID (i.e., the column denoting which subject each measurement came from).
- `within`: The name of the within-subjects factor.
- `data`: The name of the dataset.

As a demonstration, here is how you would use the anova_test() function for the osteoarthritis dataset:

```
#first, load the rstatix package, so that
#we can use the anova_test() function.
> library(rstatix)
#dv is the outcome
#wid is the subject ID
#within is the within-subject factor
> anova_test(dv = ROM, wid = subject, within = rx, data =
   osteoTall)
ANOVA Table (type III tests)

$ANOVA
  Effect DFn DFd     F     p p<.05   ges
1     rx   2  18 3.967 0.037     * 0.108

$`Mauchly's Test for Sphericity`
  Effect     W    p p<.05
1     rx 0.918 0.71

$`Sphericity Corrections`
  Effect   GGe      DF[GG] p[GG] p[GG]<.05   HFe     DF[HF] p[HF]
   p[HF]<.05
1     rx 0.924 1.85, 16.63 0.042         * 1.152 2.3, 20.74 0.037
          *
```

There are three rows of results displayed by the anova_test() function:

- ANOVA
- Mauchly's Test for Sphericity
- Sphericity Corrections

Unintuitively, the first result we should look at is the second one, Mauchly's Test for Sphericity, which assesses whether the sphericity assumption is plausible. If we

accept the sphericity assumption (i.e., if the test of sphericity is non-significant), then we use the first line of results (the "ANOVA" results). In this case, we see that the sphericity test yields a p-value of 0.71, suggesting that the sphericity assumption is plausible, and thus that we should look at the "ANOVA" row of results. From here, the interpretation is identical to typical one-way ANOVA: The null hypothesis is that the population mean ROM is equal across the three treatment groups (TENS, diathermy, and control), and we reject that null hypothesis due to the small p-value of 0.037. Meanwhile, the "DFn" and "DFd" denote the numerator and denominator degrees of freedom, respectively, which are analogous to the between and within degrees of freedom from one-way ANOVA. For repeated measures ANOVA, the numerator degrees of freedom is $k - 1$ (same as one-way ANOVA), but the denominator degrees of freedom is $(n - 1)(k - 1)$, which differs from one-way ANOVA. Familiarly, we still have an F statistic (computed as 3.967), and this statistic is compared to the $F_{2,18}$ distribution to obtain the p-value. Finally, the "ges" (standing for "generalized eta squared") is a measure of *effect size*, in the same sense we discussed in Chapter 11.

If the sphericity assumption is violated, then the third line of results ("Sphericity corrections") are used instead. In `R`, two types of corrections are used: Greenhouse-Geisser ("GG") and Huynh-Feldt ("HF"). We won't discuss these corrections in depth here, but at a high level, they adjust the degrees of freedom of the F distribution to account for the (possible) lack of sphericity. In this case, the Greenhouse-Geisser correction lowers the degrees of freedom slightly (1.85 and 16.63, vs 2 and 18) while the Huynh-Feldt increases them slightly (2.3, 20.74). There is some controversy about when to use which correction, but generally it is safe to go with the Huynh-Feldt correction. It's unusual for the corrections to yield wildly different results. In any case, the null hypothesis for the p-values produced by these corrections is again the same as one-way ANOVA, that the population mean outcomes are equal across treatment groups.

Thus, the work-flow of repeated measures ANOVA is quite similar to the work-flow of a t-test. As discussed in Chapter 5, when running a t-test, we first assess whether the equal variance assumption is plausible. If it seems plausible, then we set `var.equal = TRUE` in the `t.test()` function; otherwise, we set `var.equal = FALSE`, thereby correcting the degrees of freedom and yielding a slightly different p-value. Conceptually speaking, we're doing the same thing here, except we check for sphericity, which is the multivariate equivalent of the equal variance assumption.

13.6 Mixed between/within-subjects designs

One of the most common designs is a two-factor ANOVA, where one factor is varied between subjects and the other within subjects. In other words, we have two factors and a quantitative outcome, suggesting that we should use two-way ANOVA. However, because one of the factors is a within-subjects factor, we need to use concepts of within-subject analysis from the previous section. Otherwise, the workflow is nearly identical to the workflow for typical two-way ANOVA (Chapter 8), as we discuss below.

Recall that, in the previous section, we outlined how to use the anova_test() function to run a repeated measures ANOVA, where there is only a within-subjects factor. Within that function, you needed to specify a `within` argument, which corresponded to the within-subjects factor. Now we are discussing a two-way ANOVA where there is a within-subjects factor *and* a between-subjects factor. Luckily, we can again use the anova_test() function, and (luckily again) the extension is very intuitive: In addition to specifying a `within` argument, you also have to specify a `between` argument, which corresponds to the between-subjects factor.

To demonstrate this, we will consider a simulated experiment where 12 subjects were asked to complete three exercises (walking, running, and cycling), and the amount of kilocalories they burned (the outcome of interest) was recorded after each exercise. The gender (male or female) of each subject was also recorded. Thus, `gender` is the between-subjects factor while `exercise` is the within-subjects factor. Here's how you would run a mixed two-way ANOVA for this experiment:

```
#note that the data needs to be in tall format
#kilocalories is the outcome (dv)
#id is the subject ID (wid)
#gender is the between-subjects factor (between)
#exercise is the within-subjects factor (within)
#energy is the dataset (data)
> anova_test(dv = kilocalories, wid = id, between = gender, within
    = exercise, data = energy)
ANOVA Table (type II tests)

$ANOVA
           Effect DFn DFd      F        p p<.05   ges
1          gender   1  10  1.947 1.93e-01       0.114
2        exercise   2  20 21.226 1.13e-05     * 0.419
3 gender:exercise   2  20  1.198 3.23e-01       0.039

```

```
$`Mauchly's Test for Sphericity`
           Effect    W     p p<.05
1        exercise 0.73 0.242
2 gender:exercise 0.73 0.242

$`Sphericity Corrections`
           Effect   GGe      DF[GG]    p[GG] p[GG]<.05   HFe
   DF[HF]    p[HF] p[HF]<.05
1        exercise 0.787 1.57, 15.74 7.43e-05         * 0.909 1.82,
    18.18 2.53e-05         *
2 gender:exercise 0.787 1.57, 15.74 3.17e-01           0.909 1.82,
    18.18 3.20e-01
```

Let's first look at the ANOVA results (the first row). We have results for the between-subjects factor, `gender`; this factor does not show up in the Mauchly's Test for Sphericity (second row) or Sphericity Corrections (third row) of results. The reason is that notions of sphericity are only needed for within-subjects factors; thus, we can look at the ANOVA results to assess the effect of the between-subjects factor. Meanwhile, the workflow is a combination of the workflow for repeated measures ANOVA and two-way ANOVA:

1. First, look at the Mauchley's Test for Sphericity results. If all the p-values are insignificant, use the ANOVA results to assess effects involving the within-subjects factor. If any are significant, use the Sphericity Corrections row.

2. Notice that the ANOVA and Sphericity Corrections results both have an interaction term, denoted by `gender:exercise`. Similar to two-way ANOVA, if the interaction is significant, then the effect of one factor depends on the level of the other factor, regardless of the p-values for the main effects. Thus, if the interaction p-value is significant (which it is *not* in this example), we should simply state that there is a significant interaction between the two factors and refrain from interpreting the main effects.

3. If the interaction is insigificant, then we can interpret both of the "main effects" null hypotheses of equal means for all levels of one factor averaging over (or ignoring) the other factor. For example, the null hypothesis for `gender` is that the mean kilocalories is equal across male and female groups; the null hypothesis for `exercise` is that it's equal across walking, running, and cycling groups.

In this example, the sphericity assumption seems plausible (due to the p-value of 0.242 from Mauchly's Test for Sphericity); thus, we can look at the ANOVA section of results for interpretation. We see that the interaction is insignificant (p-value = 0.323); thus, we can continue to interpret the main effects of `gender` and `exercise`. For `gender`, we see that the p-value is 0.193; thus, we fail to reject the null hypothesis that the mean kilocalorie expenditure is equal between the two gender groups (males and females). On the other hand, we see that the p-value for `exercise` is fairly below 0.05; thus, we reject the null hypothesis for `exercise`, thereby concluding that the mean kilocalorie expenditure differs across the three `exercise` groups (walking, running, and cycling).

Note that, if an interaction is not significant in two-way ANOVA, we usually remove it from the model before interpreting the main effects; however, that is not possible in repeated measures ANOVA. Instead, you should simply ignore the interaction term and interpret the main effects results for each factor, as we did in the previous paragraph.

When interpreting main effects for factors with more than two levels, if you reject the null hypothesis, you can do follow-up hypothesis tests, as discussed in Chapter 12. For example, you may select a set of planned contrasts for either factor, or you may use post-hoc corrections for the between-subjects factor (e.g., Tukey or Dunnett's procedure). For the within-subjects factor, Bonferroni corrections are available for Estimated Marginal Means via the `emmeans` R package. For now, follow-up tests for mixed ANOVA models is outside the scope of this textbook, and so we will not go into further detail here.

Repeated measures analysis is appropriate when one (or more) factors is a within-subjects factor. When only a single within-subjects factor is present, repeated measures analysis is analogous to one-way ANOVA, but with tests for sphericity (and possible corrections for non-sphericity) When a within-subjects factor *and* a between subjects factor are present, the experiment forms a "mixed design." This design is analogous to two-way ANOVA, but again notions of sphericity have to be addressed.

Chapter 14

Mixed Models

A flexible approach to correlated data.

14.1 Overview

Correlated data arise frequently in statistical analyses. This may be due to grouping of subjects, e.g., students within classrooms, or to repeated measurements on each subject over time or space, or to multiple related outcome measures at one point in time. Mixed model analysis provides a general, flexible approach in these situations, because it allows a wide variety of correlation patterns (or variance-covariance structures) to be explicitly modeled.

As mentioned in Chapter 13, multiple measurements per subject generally result in the correlated errors that are explicitly forbidden by the assumptions of standard (between-subjects) AN(C)OVA and regression models. While repeated measures analysis of the type found in SPSS, which I will call "classical repeated measures analysis", can model general (multivariate approach) or spherical (univariate approach) variance-covariance structures, they are not suited for other explicit structures. Even more importantly, these repeated measures approaches discard all results on any subject with even a single missing measurement, while mixed models allow other data on such subjects to be used as long as the missing data meets the so-called missing-at-random definition. Another advantage of mixed models is that they naturally handle uneven spacing of repeated measurements. Also important is the fact that mixed model analysis is often more

interpretable than classical repeated measures. Finally, mixed models can also be extended (as generalized mixed models) to non-Normal outcomes.

The term mixed model refers to the use of both fixed and random effects in the same analysis. As explained in Section 13.1, fixed effects have levels that are of primary interest and would be used again if the experiment were repeated. Random effects have levels that are not of primary interest and would not necessarily be used again if the experiment were repeated; rather, they are viewed as a random selection from a much larger set of levels used across different experiments. For example, subject effects are almost always random effects, because from experiment to experiment you would use different subjects. Meanwhile, treatment levels are almost always fixed effects, because from experiment to experiment you would still use the same treatment levels. Other examples of random effects include cities in a multi-site trial, batches in a chemical or industrial experiment, and classrooms in an educational setting.

As explained in more detail below, the use of both fixed and random effects in the same model can be thought of hierarchically, and there is a very close relationship between mixed models and the class of models called hierarchical linear models. The hierarchy arises because we can think of one level for subjects and another level for measurements within subjects. In more complicated situations, there can be more than two levels of the hierarchy. The hierarchy also plays out in the different roles of the fixed and random effects parameters. Again, this will be discussed more fully below, but the basic idea is that the fixed effects parameters tell how population means differ between any set of treatments, while the random effect parameters represent the general variability among subjects or other units.

Mixed models use both fixed and random effects. These correspond to a hierarchy of levels, with the repeated, correlated measurements occurring among all of the lower level units for each particular upper level unit.

14.2 A video game example

Consider a study of the learning effects of repeated plays of a video game where age is expected to have an effect. The data are in MMvideo.txt. The quantitative

outcome is the score on the video game (in thousands of points). The explanatory variables are age group of the subject and "trial" which represents which time the subject played the game (1 to 5). The "id" variable identifies the subjects. Note that the data are in the tall format with one observation per row, and multiple rows per subject.

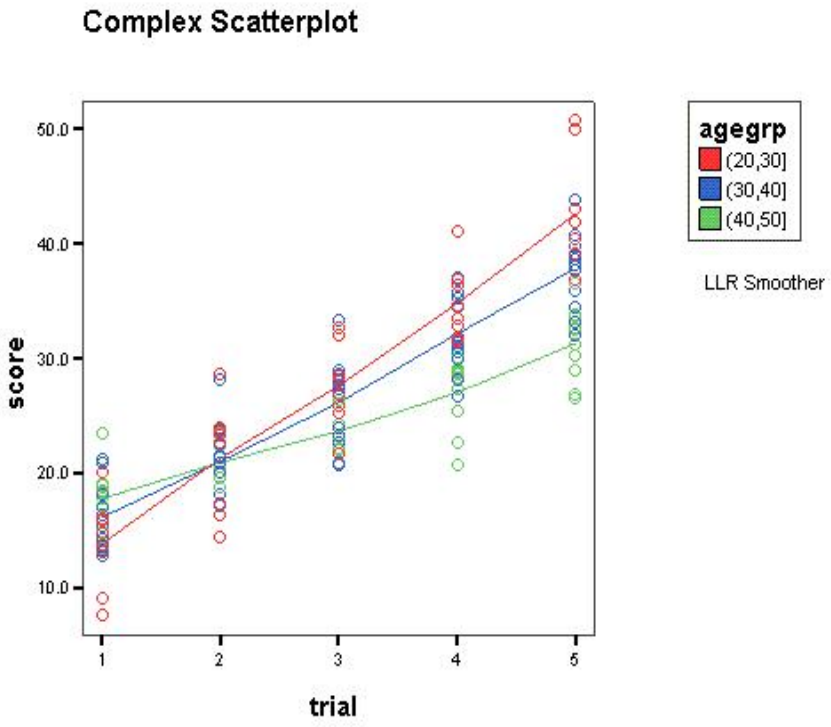


Figure 14.1: EDA for video game example with smoothed lines for each age group.

Some EDA is shown in Figure 14.1. The plot shows all of the data points, with game score plotted against trial number. Smoothed lines are shown for each of the three age groups. The plot shows evidence of learning, with players improving their score for each game over the previous game. The improvement looks fairly linear. The y-intercept (off the graph to the left) appears to be higher for older players. The slope (rate of learning) appears steeper for younger players.

At this point you are most likely thinking that this problem looks like an ANCOVA problem where each age group has a different intercept and slope for the

relationship between the quantitative variables trial and score. But ANCOVA assumes that all of the measurements for a given age group category have uncorrelated errors. In the current problem, each subject has several measurements and the errors for those measurements will almost surely be correlated. This shows up in the data here, because, for each subject, many of their measurements fall on the same side of their group's fitted line. In other words, a particular subject's measurements tend to be "greater than average" or "less than average," meaning that their measurements are correlated.

14.3 Mixed model approach

The solution to the problem of correlated within-subject errors in the video game example is to let each subject have his or her own "personal" intercept (and possibly slope) randomly deviating from the mean intercept for each age group. This results in a group of parallel "personal" regression lines (or non-parallel if the slope is also random). Then, it is reasonable (but not certain) that the errors around the personal regression lines will be uncorrelated. One way to do this is to use subject identification as a categorical variable, but this is treating the inherently random subject-to-subject effects as fixed effects, and "wastes" one parameter for each subject in order to estimate his or her personal intercept. A better approach is to just estimate a single variance parameter which represents how spread out the random intercepts are around the common intercept of each group (usually following a Normal distribution). This is the mixed models approach.

From another point of view, in a mixed model we have a hierarchy of levels. At the top level the units are often subjects or classrooms. At the lower level we could have repeated measurements within subjects or students within classrooms. The lower level measurements that are within the same upper level unit are correlated, when all of their measurements are compared to the mean of all measurements for a given treatment, but often uncorrelated when compared to a personal (or class level) mean or regression line. We also expect that there are various measured and unmeasured aspects of the upper level units that affect all of the lower level measurements similarly for a given unit. For example various subject skills and traits may affect all measurements for each subject, and various classroom traits such as teacher characteristics and classroom environment affect all of the students in a classroom similarly. Treatments are usually applied randomly to whole upper-level units. For example, some subjects receive a drug and some receive a placebo,

or some classrooms get an aide and others do not.

In addition to all of these aspects of hierarchical data analysis, there is a variety of possible variance-covariance structures for the relationships among the lower level units. One common structure is called compound symmetry, which indicates the same correlation between all pairs of measurements, as in the sphericity characteristic of Chapter 13. This is a natural way to represent the relationship between students within a classroom. If the true correlation structure is compound symmetry, then using a random intercept for each upper level unit will remove the correlation among lower level units. Another commonly used structure is autoregressive, in which measurements are ordered, and adjacent measurements are more highly correlated than distant measurements. This is especially common for time series data, where time measurements are clearly ordered (as in the video game example).

To summarize, in each problem the hierarchy is usually fairly obvious, but the user must think about and specify which fixed effects (explanatory variables, including transformations and interactions) affect the average responses for all subjects. Then the user must specify which of the fixed effect coefficients are sufficient without a corresponding random effect as opposed to those fixed coefficients which only represent an average around which individual units vary randomly. In addition, correlations among measurements that are not fully accounted for by the random intercepts and slopes may be specified. And finally, if there are multiple random effects, the correlation of these various effects may need to be specified.

To run a mixed model, the user must make many choices. The main choices are (1) specifying what is a fixed effect, (2) specifying what is a random effect, (3) the assumed correlation structure of the data, and (4) the nature of the hierarchy of the data.

14.4 Analyzing the video game example

Based on Figure 14.1, we should model separate linear relationships between trial number and game score for each age group. Figure 14.2 shows smoothed lines for each subject. From this figure, it looks like we need a separate slope and intercept

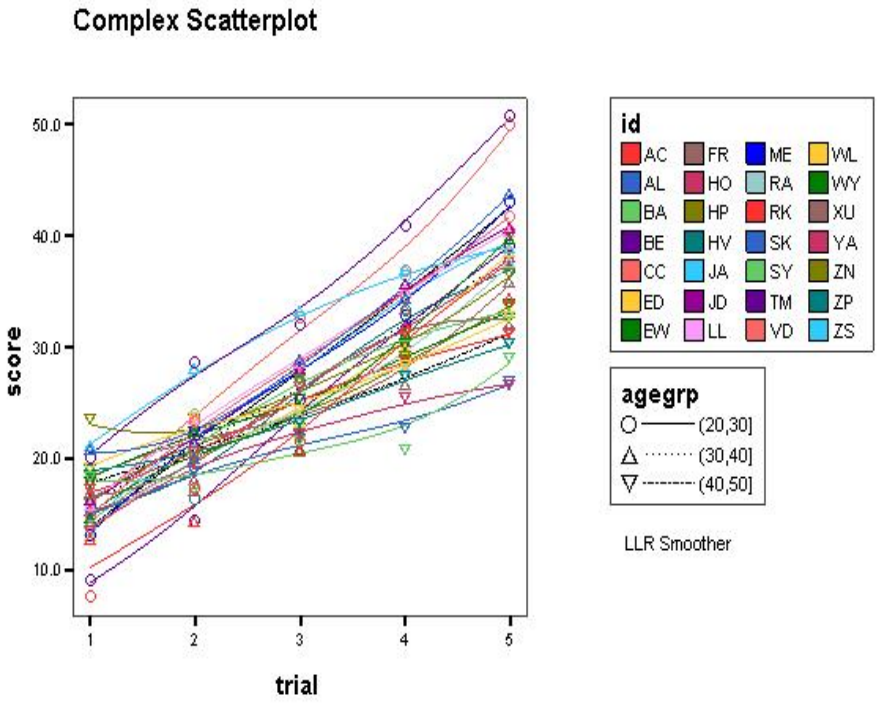


Figure 14.2: EDA for video game example with smoothed lines for each subject.

for each age group. It is also fairly clear that in each group there is random subject-to-subject variation in the intercepts. We should also consider the possibilities that the "learning trajectory" is curved rather than linear, perhaps using the square of the trial number as an additional covariate to create a quadratic curve. We should also check if a random slope is needed. It is also prudent to check if the random intercept is really needed. In addition, we should check if an autoregressive model is needed.

14.5 Setting up a model in SPSS

The mixed models section of SPSS, accessible from the menu item "Analyze / Mixed Models / Linear", has an initial dialog box ("Specify Subjects and Repeated"), a main dialog box, and the usual subsidiary dialog boxes activated by clicking buttons in the main dialog box. In the initial dialog box (Figure 14.3) you will always specify the upper level of the hierarchy by moving the identifier for that level into the "subjects" box. For our video game example this is the subject "id" column. For a classroom example in which we study many students in each classroom, this would be the classroom identifier.

If we want to model the correlation of the repeated measurements for each subject (other than the correlation induced by random intercepts), then we need to specify the order of the measurements within a subject in the bottom ("repeated") box. For the video game example, the trial number could be appropriate.

The main "Linear Mixed Models" dialog box is shown in figure 14.4. (Note that just like in regression analysis use of transformation of the outcome or a quantitative explanatory variable, i.e., a covariate, will allow fitting of curves.) As usual, you must put a quantitative outcome variable in the "Dependent Variable" box. In the "Factor(s)" box you put any categorical explanatory variables (but not the subject variable itself). In the "Covariate(s)" box you put any quantitative explanatory variables. **Important note:** For mixed models, specifying factors and covariates on the main screen does *not* indicate that they will be used in the model, only that they are available for use in a model.

The next step is to specify the fixed effects components of the model, using the Fixed button which brings up the "Fixed Effects" dialog box, as shown in figure 14.5. Here you will specify the structural model for the "typical" subject, which is just like what we did in ANCOVA models. Each explanatory variable or

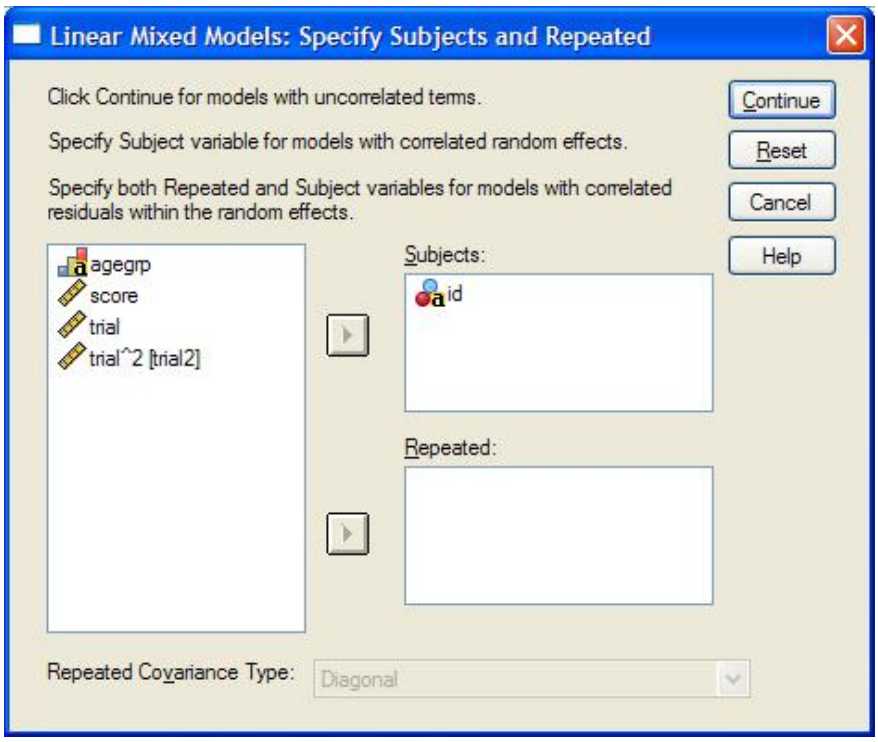


Figure 14.3: Specify Subjects and Repeated Dialog Box.

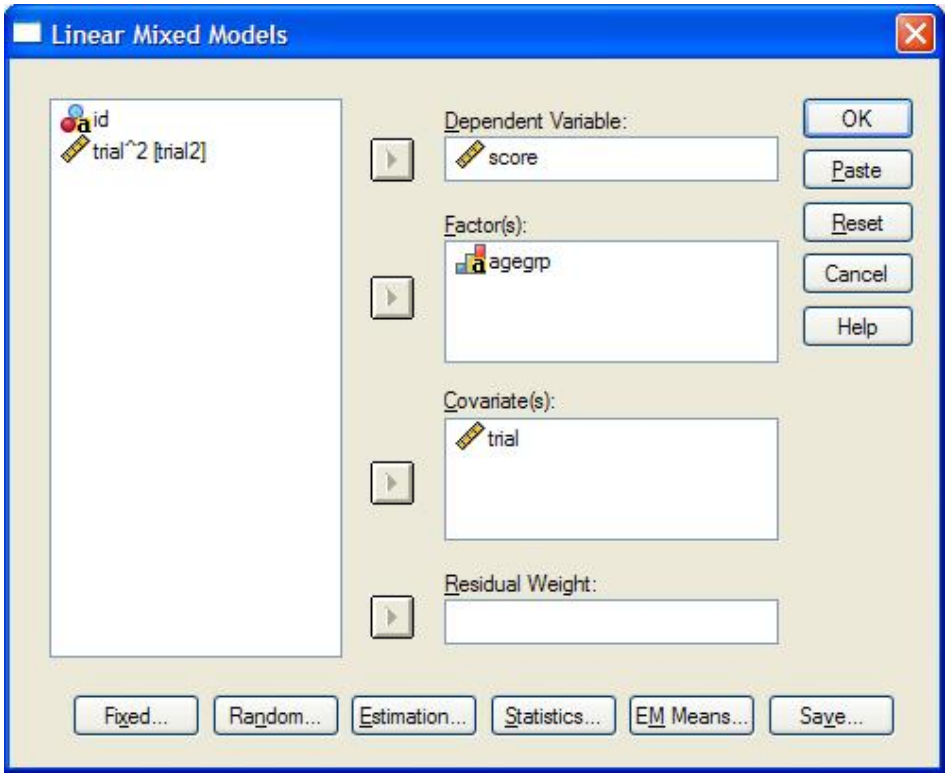


Figure 14.4: Main Linear Mixed Effects Dialog Box.

interaction that you specify will have a corresponding parameter estimated, and that estimate will represent the relationship between that explanatory variable and the outcome if there is no corresponding random effect, and it will represent the mean relationship if there is a corresponding random effect.

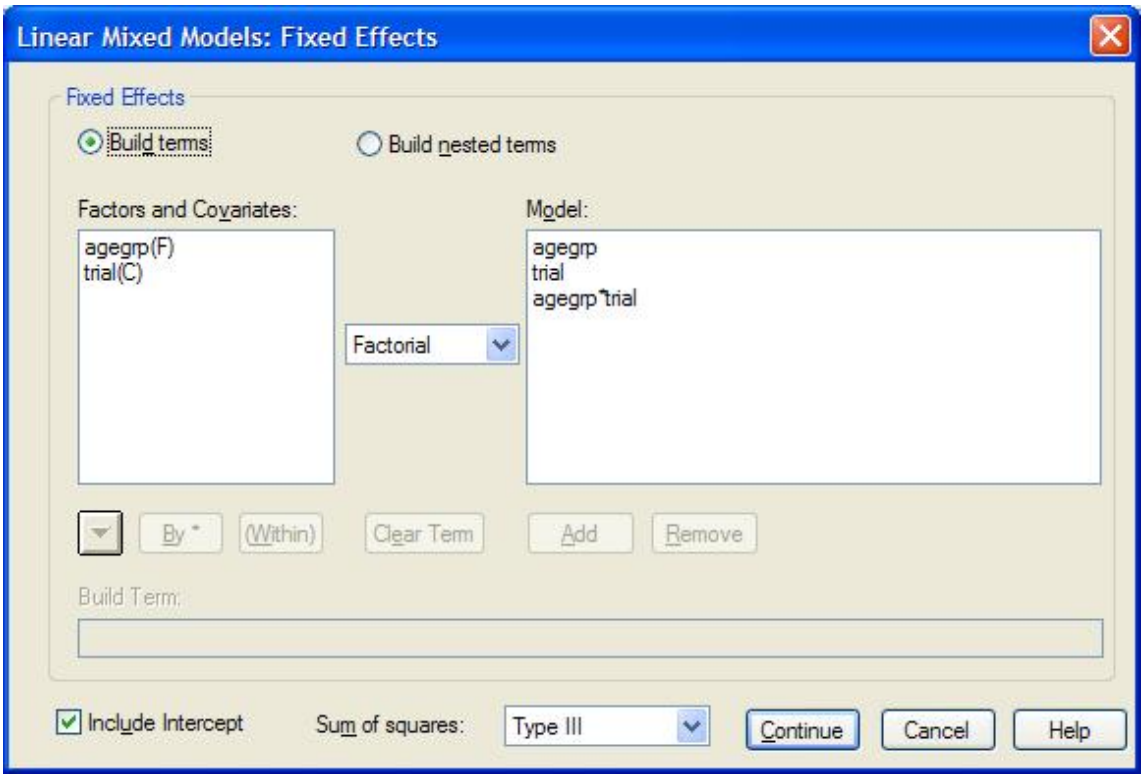


Figure 14.5: Fixed Effects Dialog Box.

For the video game example, I specified main effects for age group and trial plus their interaction. (You will always want to include the main effects for any interaction you specify.) Just like in ANCOVA, this model allows a different intercept and slope for each age group. The fixed intercept (included unless the "Include intercept" check box is unchecked) represents the (mean) intercept for the baseline age group, and the $k-1$ coefficients for the age group factor (with $k=3$ levels) represent differences in (mean) intercept for the other age groups. The trial coefficient represents the (mean) slope for the baseline group, while the interaction coefficients represent the differences in (mean) slope for the other groups relative to the baseline group. (As in other "model" dialog boxes, the actual model depends only on what is in the "Model box", not how you got it there.)

In the "Random Effects" dialog box (Figure 14.6), you will specify which parameters of the fixed effects model are only means around which individual subjects vary randomly, which we think of as having their own personal values. Mathematically these personal values, e.g., a personal intercept for a given subject, are equal to the fixed effect plus a random deviation from that fixed effect, which is zero on average, but which has a magnitude that is controlled by the size of the random effect, which is a variance.

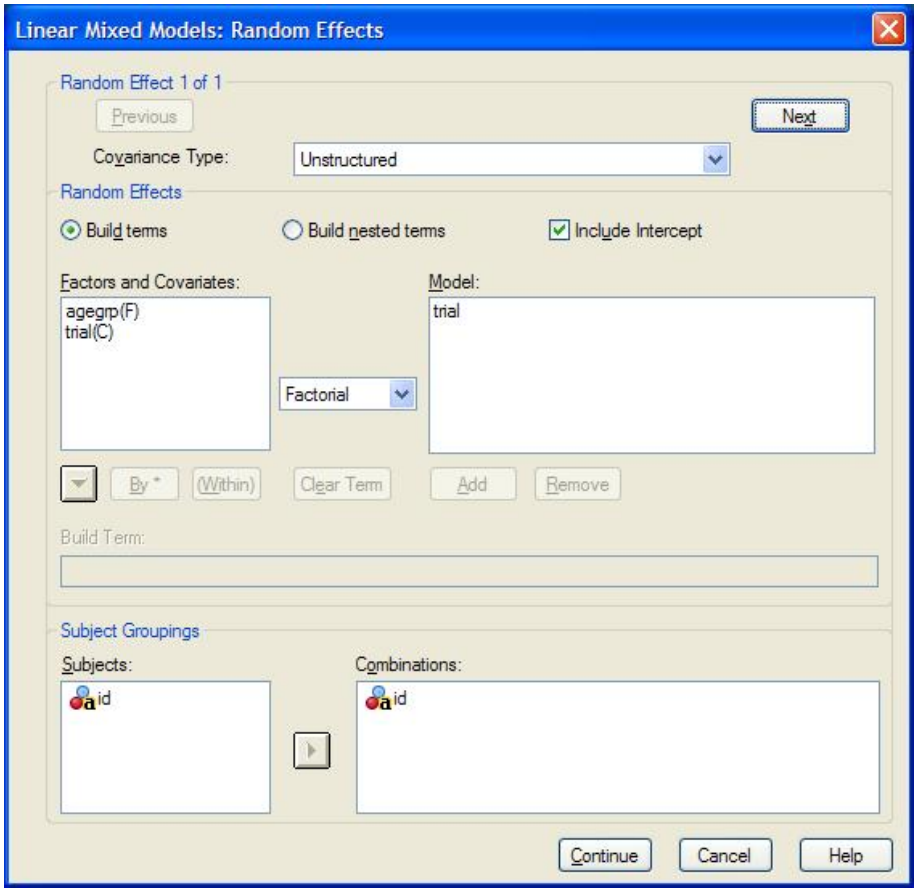


Figure 14.6: Random Effects Dialog Box.

In the random effects dialog box, you will usually want to check "Include Intercept", to allow a separate intercept (or subject mean if no covariate is used) for each subject (or each level of some other upper level variable). If you specify any random effects, then you must indicate that there is a separate "personal" value of, say, the intercept, for each subject by placing the subject identifier in the

"Combinations" box. (This step is very easy to forget, so get in the habit of doing this every time.)

To model a random slope, move the covariate that defines that slope into the "Model" box. In this example, moving trial into the Model box could be used to model a random slope for the score by trial relationship. It does not make sense to include a random effect for any variable unless there is also a fixed effect for that variable, because the fixed effect represents the average value around which the random effect varies. If you have more than one random effect, e.g., a random intercept and a random slope, then you need to specify any correlation between these using the "Covariance Type" drop-down box. For a single random effect, use "identity". Otherwise, "unstructured" is usually most appropriate because it allows correlation among the random effects (see next paragraph). Another choice is "diagonal" which assumes no correlation between the random effects.

What does it mean for two random effects to be correlated? I will illustrate this with the example of a random intercept and a random slope for the trial vs. game score relationship. In this example, there are different intercepts and slopes for each age group, so we need to focus on any one age group for this discussion. The fixed effects define a mean intercept and mean slope for that age group, and of course this defines a mean fitted regression line for the group. The idea of a random intercept and a random slope indicate that any given subject will "wiggle" a bit around this mean regression line both up or down (random intercept) and clockwise or counterclockwise (random slope). The variances (and therefore standard deviations) of the random effects determine the sizes of typical deviations from the mean intercept and slope. But in many situations—like this video game example—subjects with a higher than average intercept tend to have a lower than average slope, so there is a negative correlation between the random intercept effect and the random slope effect. The intuition behind this is that it is often the case that subjects with have high "starting performances" tend to have lower "marginal gains", because there is often a kind of "diminishing marginal returns" behavior in many applications. We can also view this situation with the following statistical formalism: The next subject is represented by a random draw of an intercept deviation and a slope deviation from a distribution with mean zero for both, but with a negative correlation between these two random deviations. Then, the personal intercept and slope are constructed by adding these random deviations to the fixed effect coefficients.

Some other buttons in the main mixed models dialog box are useful. I rec-

ommend that you always click the Statistics button, then check both "Parameter estimates" and "Tests for covariance parameters". The parameter estimates are needed for interpretation of the results, similar to what we did for ANCOVA (see Chapter 10). The tests for covariance parameters aid in determining which random effects are needed in a given situation. The "EM Means" button allows generation of "expected marginal means" which average over all subjects and other treatment variables. In the current video game example, marginal means for the three age groups is not very useful because this averages over the trials, and the score varies dramatically over the trials. Also, in the face of an interaction between age group and trial number, interpreting the averages for each level of age group is probably misleading.

As you can see there are many choices to be made when creating a mixed model. In fact there are many more choices possible than described here. This flexibility makes mixed models an important general purpose tool for statistical analysis, but suggests that it should be used with caution by inexperienced analysts.

Specifying a mixed model requires many steps, each of which requires an informed choice. This is both a weakness and a strength of mixed model analysis.

14.6 Interpreting the results for the video game example

Here is some of the SPSS output for the video game example. We start with the model for a linear relationship between trial and score with separate intercepts and slopes for each age group, and including a random per-subject intercept. Table 14.1 is called "Model Dimension". Focus on the "number of parameters" column. The total is a measure of overall complexity of the model and plays a role in model selection (see next section). For quantitative explanatory variables, there is only one parameter. For categorical variables, this column tells how many parameters are being estimated in the model. The "number of levels" column tells how many lines are devoted to an explanatory variable in the Fixed Effects table (see below). However, when there is a categorical variable with k levels, only $k-1$ parameters

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed	Intercept	1		1	
Effects	agegrp	3		2	
	trial	1		1	
	agegrp * trial	3		2	
Random Effects	Intercept	1	Identity	1	id
Residual				1	
Total		9		8	

Table 14.1: Model dimension for the video game example.

are estimated, because one of the categories is coded as the reference category (this is something we discussed when we learned ANCOVA in Chapter 10, specifically Section 10.3). As a result, some of the estimated parameters in the Fixed Effects table are "blank" (with a period in the rest of the columns). For example, from the table below, we can see that the oldest age group was coded as the reference category. Furthermore, we can see that we have a single random effect, which is an intercept for each level of id (each subject). As a result, there are two additional parameters (which we can see by looking at the "Number of Parameters" column): one for computing the mean intercept across all subjects, and one for computing the variance of these random intercepts. In general, the Model Dimension table is a good quick check that the computer is fitting the model that you intended to fit.

The next table in the output is labeled "Information Criteria" and contains many different measures of how well the model fits the data. (This table will be discussed in the section on model comparison, Section 14.7.) I recommend that you only pay attention to the last one, "Schwartz's Bayesian Criterion (BIC)", also called Bayesian Information Criterion. In this model, the value is 718.4. For information criteria like BIC, the *lower* the number, the better the model.

Next comes the Fixed Effects tables (Tables 14.2 and 14.3). The tests of fixed effects have an ANOVA-style test for each fixed effect in the model. This is nice because it gives a single overall test of the usefulness of a given explanatory variable, without focusing on individual levels. Generally, you will want to remove explanatory variables that do not have a significant fixed effect in this table, and then rerun the mixed effect analysis with the simpler model. In this example, all effects are significant (less than the standard alpha of 0.05). Note that I converted

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	57.8	266.0	<0.0005
agegrp	2	80.1	10.8	<0.0005
trial	1	118.9	1767.0	<0.0005
agegrp * trial	2	118.9	70.8	<0.0005

Table 14.2: Tests of Fixed Effects for the video game example.

						95% Conf. Int.	
		Std.				Lower	Upper
Parameter	Estimate	Error	df	t	Sig.	Bound	Bound
Intercept	14.02	1.11	55.4	12.64	<0.0005	11.80	16.24
agegrp=(20,30)	-7.26	1.57	73.0	-4.62	<0.0005	-10.39	-4.13
agegrp=(30,40)	-3.49	1.45	64.2	-2.40	0.019	-6.39	-0.59
agegrp=(40,50)	0	0	.	.	.	.	.
trial	3.32	0.22	118.9	15.40	<0.0005	2.89	3.74
(20,30)*trial	3.80	0.32	118.9	11.77	<0.0005	3.16	4.44
(30,40)*trial	2.14	0.29	118.9	7.35	<0.0005	1.57	2.72
(40,50)*trial	0	0	.	.	.	.	.

Table 14.3: Estimates of Fixed Effects for the video game example.

the SPSS p-values from 0.000 to the correct form.

The Estimates of Fixed Effects table does not appear by default; it is produced by choosing "parameter estimates" under Statistics. As mentioned before, we can see that age group 40-50 is the "baseline" or reference category (SPSS chooses the last category by default). Therefore the (fixed) intercept value of 14.02 represents the mean game score (in thousands of points) for 40 to 50 year olds for trial zero. Because trials start at one, the intercepts are not meaningful in themselves for this problem, although they are needed for calculating and drawing the best fit lines for each age group.

As in ANCOVA, writing out the full regression model then simplifying tells us that the intercept for 20 to 30 year olds is 14.02-7.26=6.76 and this is significantly lower than for 40 to 50 year olds (t=-4.62, p<0.0005, 95% CI for the difference is

4.13 to 10.39 thousand points lower). Similarly we know that the 30 to 40 years olds have a lower intercept than the 40 to 50 year olds. Again these intercepts themselves are not directly interpretable because they represent trial zero. (It would be worthwhile to recode the trial numbers as zero to four, then rerun the analysis, because then the intercepts would represent game scores the first time someone plays the game.)

The trial coefficient of 3.32 represents that average gain in game score (in thousands of points) for each subsequent trial *for the baseline 40 to 50 year old age group*. The interaction estimates tell the *difference* in slope for other age groups compared to the 40 to 50 year olds. Here both the 20 to 30 year olds and the 30 to 40 year olds learn quicker than the 40 to 50 year olds, as shown by the significant interaction p-values and the positive sign on the estimates. For example, we are 95% confident that the trial to trial "learning" gain is 3.16 to 4.44 thousand points *higher* for the youngest age group compared to the oldest age group.

Interpret the fixed effects for a mixed model in the same way as an ANOVA, regression, or ANCOVA depending on the nature of the explanatory variables(s), but realize that any of the coefficients that have a corresponding random effect represent the mean over all subjects, and each individual subject has their own "personal" value for that coefficient.

The next table is called "Estimates of Covariance Parameters" (table 14.4). It is very important to realize that while the parameter estimates given in the Fixed Effects table are estimates of mean parameters, the parameter estimates in this table are estimates of variance parameters. The intercept variance is estimated as 6.46, so the estimate of the standard deviation is 2.54. This tells us that for any given age group, e.g., the oldest group with mean intercept of 14.02, the individual subjects will have "personal" intercepts that are up to 2.54 higher or lower than the group average about 68% of the time, and up to 4.08 higher or lower about 95% of the time. The null hypothesis for this parameter is a variance of zero, which would indicate that a random effect is not needed. The test statistic is called a Wald Z statistic. Here we reject the null hypothesis (Wald Z=3.15, p=0.002) and conclude that we do need a random intercept. This suggests that there are important unmeasured explanatory variables for each subject that raise or lower their performance in a way that appears random because we do not know the

					95% Conf. Int.	
		Std.	Wald		Lower	Upper
Parameter	Estimate	Error	Z	Sig.	Bound	Bound
Residual	4.63	0.60	7.71	<0.0005	3.59	5.97
Intercept(Subject=id) Variance	6.46	2.05	3.15	0.002	3.47	12.02

Table 14.4: Estimates of Covariance Parameters for the video game example.

value(s) of the missing explanatory variable(s).

The estimate of the residual variance, with standard deviation equal to 2.15 (square root of 4.63), represents the variability of individual trial's game scores around the individual regression lines for each subjects. We are assuming that once a personal best-fit line is drawn for each subject, their actual measurements will randomly vary around this line with about 95% of the values falling within 4.30 of the line. (This is an estimate of the same σ^2 as in a regression or ANCOVA problem.) The p-value for the residual is not very meaningful.

Random effects estimates are variances. Interpret a random effect parameter estimate as the magnitude of the variability of "personal" coefficients from the mean fixed effects coefficient.

All of these interpretations are contingent on choosing the right model. The next section discusses model selection.

14.7 Model selection for the video game example

Because there are many choices among models to fit to a given data set in the mixed model setting, we need an approach to choosing among the models. Even then, we must always remember that all models are wrong (because they are idealized simplifications of Nature), but some are useful. Sometimes a single best model is chosen. Sometimes subject matter knowledge is used to choose the most useful models (for prediction or for interpretation). And sometimes several models, which differ but appear roughly equivalent in terms of fit to the data, are presented as

the final summary for a data analysis problem.

Two of the most commonly used methods for **model selection** are **penalized likelihood** and testing of individual coefficient or variance estimate p-values. Other more sophisticated methods include model averaging and cross-validation, but they will not be covered in this text.

14.7.1 Penalized likelihood methods for model selection

Penalized likelihood methods calculate the likelihood of the observed data using a particular model (see Chapter 3). But because it is a fact that the likelihood always goes up when a model gets more complicated, whether or not the additional complication is "justified", a model complexity penalty is used. Several different penalized likelihoods are available in SPSS, but I recommend using the **BIC (Bayesian information criterion)**. AIC (Akaike information criterion) is another commonly used measure of model adequacy. The BIC number penalizes the likelihood based on both the total number of parameters in a model and the number of subjects studied. The formula varies between different programs based on whether or not a factor of two is used and whether or not the sign is changed. In SPSS, just remember that "smaller is better".

The absolute value of the BIC has no interpretation. Instead the BIC values can be computed for two (or more) models, and the values compared. A smaller BIC indicates a better model. A difference of under 2 is "small" so you might use other considerations to choose between models that differ in their BIC values by less than 2. If one model has a BIC more than 2 lower than another, that is good evidence that the model with the lower BIC is a better balance between complexity and good fit (and hopefully is closer to the true model of Nature).

In our video game problem, several different models were fit and their BIC values are shown in table 14.5. Based on the "smaller is better" interpretation, the (fixed) interaction between trial and age group is clearly needed in the model, as is the random intercept. The additional complexity of a random slope is clearly not justified. The use of quadratic curves (from inclusion of a trial^2 term) is essentially no better than excluding it, so I would not include it on grounds of parsimony. (Parsimony is a concept that comes up a lot in statistics—essentially, it means, "The simpler, the better," i.e., Occam's Razor.)

The BIC approach to model selection is a good one, although there are some

Interaction	random intercept	random slope	quadratic curve	BIC
yes	yes	no	no	718.4
yes	no	no	no	783.8
yes	yes	no	yes	718.3
yes	yes	yes	no	727.1
no	yes	no	no	811.8

Table 14.5: BIC for model selection for the video game example.

technical difficulties. Briefly, there is some controversy about the appropriate penalty for mixed models, and it is probably better to change the estimation method from the default "restricted maximum likelihood" to "maximum likelihood" when comparing models that differ only in fixed effects. Of course you never know if the best model is one you have not checked because you didn't think of it. Ideally the penalized likelihood approach is best done by running all reasonable models and listing them in BIC order. If one model is clearly better than the rest, use that model, otherwise consider whether there are important differing implications among any group of similar low BIC models.

14.7.2 Comparing models with individual p-values

Another approach to model selection is to move incrementally to one-step more or less complex models, and use the corresponding p-values to choose between them. This method has some deficiencies, chief of which is that different "best" models can result just from using different starting places. Nevertheless, this method—called **stepwise model selection**—is commonly used.

Variants of step-wise selection include forward and backward forms. Forward selection starts at a simple model, then considers all of the reasonable one-step-more-complicated models and chooses the one with the smallest p-value for the new parameter. This continues until no addition parameters have a significant p-value. Backward selection starts at a complicated model and removes the term with the largest p-value, as long as that p-value is larger than 0.05. There is no guarantee that any kind of "best model" will be reached by stepwise methods, but in many cases a good model is reached.

14.8 Classroom example

As demonstration, here we will go through another example of how to implement and interpret mixed models (just in case you would benefit from another example, or you just don't like video games).

The (fake) data in schools.txt represent a randomized experiment of two different reading methods which were randomly assigned to third or fifth grade classrooms, one per school, for 20 different schools. The experiment lasted 4 months. The outcome is the after minus before difference for a test of reading given to each student. (In education and psychology, these "after minus before differences" are often called "gains scores".) The average sixth grade reading score for each school on a different statewide standardized test (stdTest) is used as an explanatory variable for each school (classroom).

It seems likely that students within a classroom will be more similar to each other than to students in other classrooms due to whatever school level characteristics are measured by the standardized test. Additional unmeasured characteristics including teacher characteristics, will likely also raise or lower the outcome for a given classroom.

Cross-tabulation shows that each classroom has either grade 3 or 5 and either treatment or control. The classroom sizes are 20 to 30 students. EDA, in the form of a scatterplot of standardized test scores vs. experimental test score difference are shown in figure 14.7. Grade differences are represented in color and treatment differences by symbol type. There is a clear positive correlation of standardized test score and the outcome (reading score difference), indicating that the standardized test score was a good choice of a control variable. The clustering of students within schools is clear once it is realized that each different standardized test score value represents a different school. It appears that fifth graders tend to have a larger rise than third graders. The plot does not show any obvious effect of treatment.

A mixed model was fit with classroom as the upper level ("subjects" in SPSS mixed models) and with students at the lower level. There are main effects for stdTest, grade level, and treatment group. There is a random effect (intercept) to account for school to school differences that induces correlation among scores for students within a school. Model selection included checking for interactions among the fixed effects, and checking the necessity of including the random intercept. The only change suggested is to drop the treatment effect. It was elected to keep the non-significant treatment in the model to allow calculation of a confidence interval

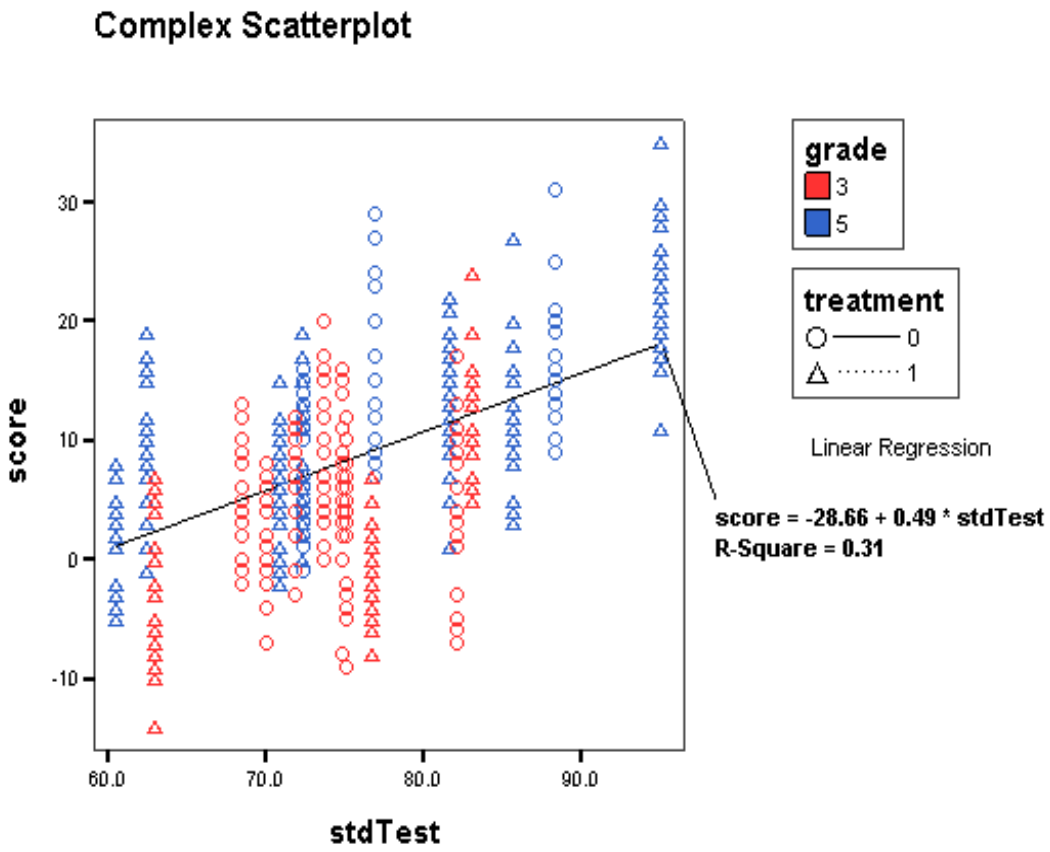


Figure 14.7: EDA for school example

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	15.9	14.3	0.002
grade	1	16.1	12.9	0.002
treatment	1	16.1	1.2	0.289
stdTest	1	15.9	25.6	<0.0005

Table 14.6: Tests of Fixed Effects for the school example.

						95% Conf. Int.	
		Std.				Lower	Upper
Parameter	Estimate	Error	df	t	Sig.	Bound	Bound
Intercept	-23.09	6.80	15.9	-3.40	0.004	-37.52	-8.67
grade=3	-5.94	1.65	16.1	-3.59	0.002	-9.45	-2.43
grade=5	0	0	.	.	.	.	.
treatment=0	1.79	1.63	16.1	1.10	0.289	-1.67	5.26
treatment=1	0	0	.	.	.	.	.
stdTest	0.44	0.09	15.9	5.05	<0.0005	0.26	0.63

Table 14.7: Estimates of Fixed Effects for the school example.

for its effect.

In what follows, we will report results from this mixed model (thereby demonstrating how to report results from these types of models in general).

We note that non-graphical EDA (ignoring the explanatory variables) showed that individual students' test score differences varied between a drop of 14 and a rise of 35 points.

The "Tests of Fixed Effects" table, Table 14.6, shows that grade (F=12.9, p=0.002) and stdTest (F=25.6, p<0.0005) each have a significant effect on a student's reading score difference, but treatment (F=1.2, p=0.289) does not.

The "Estimates of Fixed Effects" table, Table 14.7, gives the same p-values plus estimates of the effect sizes and 95% confidence intervals for those estimates. For example, we are 95% confident that the improvement seen by fifth graders is 2.43 to 9.45 *more* than for third graders. We are particularly interested in the

		Std.	Wald		95% Conf. Int.	
					Lower	Upper
Parameter	Estimate	Error	Z	Sig.	Bound	Bound
Residual	25.87	1.69	15.33	<0.0005	22.76	29.40
Intercept(Subject=sc.) Variance	10.05	3.94	2.55	0.011	4.67	21.65

Table 14.8: Estimates of Covariance Parameters for the school example.

conclusion that we are 95% confident that treatment method 0 (control) has an effect on the outcome that is between 5.26 points more and 1.67 points less than treatment 1 (new, active treatment). In other words, we do not find evidence that the treatment is effective.

We assume that students within a classroom perform similarly due to school and/or classroom characteristics. Some of the effects of the student and school characteristics are represented by the standardized test which has a standard deviation of 8.8 (not shown), and Table 14.7 shows that each one unit rise in standardized test score is associated with a 0.44 unit rise in outcome on average. Consider the comparison of schools at the mean vs. one s.d. above the mean of standardized test score. These values correspond to $\mu_{stdTest}$ and $\mu_{stdTest} + 8.8$. This corresponds to a 0.44*8.8=3.9 point change in average reading scores for a classroom. In addition, other unmeasured characteristics must be in play because Table 14.8 shows that the random classroom-to-classroom variance is 10.05 (s.d.= 3.2 points). Individual student-to-student differences, with a variance 23.1 (s.d. = 4.8 points), have a somewhat larger effect than either school-to-school differences (as measured by the standardized test) or the random classroom-to-classroom differences.

In summary, we find that students typically have a rise in test score over the four month period. Sixth graders improve on average by 5.9 more than third graders. Being in a school with a higher standardized test score tends to raise the reading score gain. Finally there is no evidence that the treatment worked better than the placebo.

In a nutshell: Mixed effects models flexibly give correct estimates of treatment and other fixed effects in the presence of the correlated errors that arise from hierarchical data.

Chapter 15

Analyzing Experiments with Categorical Outcomes

Analyzing data with non-quantitative outcomes

All of the analyses discussed up to this point assume a Normal distribution for the outcome (or for a transformed version of the outcome) at each combination of levels of the explanatory variable(s). This means that we have only been covering statistical methods appropriate for quantitative outcomes. It is important to realize that this restriction only applies to the outcome variable and not to the explanatory variables. In this chapter statistical methods appropriate for categorical outcomes are presented.

15.1 Contingency tables and chi-square analysis

This section discusses analysis of experiments or observational studies with a categorical outcome and a single categorical explanatory variable. We have already discussed methods for analysis of data with a quantitative outcome and categorical explanatory variable(s) (ANOVA and ANCOVA). The methods in this section are also useful for observational data with two categorical "outcomes" and no explanatory variable.

15.1.1 Why ANOVA and regression don't work

There is nothing in most statistical computer programs that would prevent you from analyzing data with, say, a two-level categorical outcome (usually designated generically as "success" and "failure") using ANOVA or regression or ANCOVA. But if you do, your conclusion will be wrong in a number of different ways. The basic reason that these methods don't work is that the assumptions of Normality and equal variance are strongly violated. Remember that these assumptions relate to groups of subjects with the same levels of all of the explanatory variables. The Normality assumption says that in each of these groups the outcomes are Normally distributed. We call ANOVA, ANCOVA, and regression "robust" to this assumption because moderate deviations from Normality alter the null sampling distributions of the statistics from which we calculate p-values only a small amount. But in the case of a categorical outcome with only a few (as few as two) possible outcome values, the outcome is so far from the smooth bell-shaped curve of a Normal distribution, that the null sampling distribution is drastically altered and the p-value completely unreliable.

The equal variance assumption is that, for any two groups of subjects with different levels of the explanatory variables between groups and the same levels within groups, we should find that the variance of the outcome is the same. If we consider the case of a binary outcome with coding 0=failure and 1=success, the variance of the outcome can be shown to be equal to $p_i(1 - p_i)$ where p_i is the probability of getting a success in group i (or, equivalently, the mean outcome for group i). Therefore groups with different means have different variances, violating the equal variance assumption.

A second reason that regression and ANCOVA are unsuitable for categorical outcomes is that they are based on the model equation $E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, which both is inherently quantitative, and can give numbers out of range of the category codes. The least unreasonable case is when the categorical outcome is ordinal with many possible values, e.g., coded 1 to 10. Then for any particular explanatory variable, say, β_i, a one-unit increase in x_i is associated with a β_i unit change in outcome. This works only over a limited range of x_i values, and then predictions are outside the range of the outcome values.

For binary outcomes where the coding is 0=failure and 1=success, a mean outcome of, say, 0.75 corresponds to 75% successes and 25% failures, so we can think of the prediction as being the probability of success. But again, outside of some limited range of x_i values, the predictions will correspond to the absurdity

of probabilities less than 0 or greater than 1.

Using statistical methods designed for Normal, quantitative outcomes when the outcomes are really categorical gives wrong p-values due to violation of the Normality and equal variance assumptions, and also gives meaningless out-of-range predictions for some levels of the explanatory variables.

15.2 Testing independence in contingency tables

15.2.1 Contingency and independence

A contingency table counts the number of cases (subjects) for each combination of levels of two or more categorical variables. An equivalent term is cross-tabulation (see Section 4.4.1). Among the definitions for "contingent" in the The Oxford English Dictionary is "Dependent for its occurrence or character on or upon some prior occurrence or condition". Most commonly when we have two categorical measures on each unit of study, we are interested in the question of whether the probability distribution (see section 3.2) of the levels of one measure depends on the level of the other measure, or if it is independent of the level of the second measure. For example, if we have three treatments for a disease as one variable, and two outcomes (cured and not cured) as the other outcome, then we are interested in the probabilities of these two outcomes for each treatment, and we want to know if the observed data are consistent with a null hypothesis that the true underlying probability of a cure is the same for all three treatments.

In the case of a clear identification of one variable as explanatory and the other as outcome, we focus on the probability distribution of the outcome and how it changes or does not change when we look separately at each level of the explanatory variable. The "no change" case is called independence, and indicates that knowing the level of the (purported) explanatory variable tells us no more about the possible outcomes than ignoring or not knowing it. In other words, if the variables are independent, then the "explanatory" variable doesn't really explain anything. But if we find evidence to reject the null hypothesis of independence,

then we do have a true explanatory variable, and knowing its value allows us to refine our predictions about the level of the other variable.

Even if both variables are outcomes, we can test their association in the same way as just mentioned. In fact, the conclusions are always the same when the roles of the explanatory and outcome variables are reversed, so for this type of analysis, choosing which variable is outcome vs. explanatory is immaterial.

The usual statistical test in the case of a categorical outcome and a categorical explanatory variable is whether or not the two variables are independent, which is equivalent to saying that the probability distribution of one variable is the same for each level of the other variable.

15.2.2 Contingency tables

It is a common situation to measure two categorical variables, say X (with k levels) and Y (with m levels) on each subject in a study. For example, if we measure gender and eye color, then we record the level of the gender variable and the level of the eye color variable for each subject. Usually the first task after collecting the data is to present it in an understandable form such as a **contingency table** (also known as a cross-tabulation).

For two measurements, one with k levels and the other with m levels, the contingency table is a $k \times m$ table with cells for each combination of one level from each variable, and each cell is filled with the corresponding count (also called **frequency**) of units that have that pair of levels for the two categorical variables.

For example, Table 15.1 is a (fake) contingency table showing the results of asking 271 college students what their favorite music is and what their favorite ice cream flavor is. This table was created using the `table()` function in R. In this simple form of a contingency table, we see the **cell counts** and the **marginal counts**. The margins are the extra column on the right and the extra row at the bottom. The cells are the rest of the numbers in the table. Each cell tells us how many subjects gave a particular pair of answers to the two questions. For example, 23 students said both that strawberry is their favorite ice cream flavor and that jazz is their favorite type of music. The right margin sums over ice cream types to

		favorite ice cream				
		chocolate	vanilla	strawberry	other	total
favorite music	rap	5	10	7	38	60
	jazz	8	9	23	6	46
	classical	12	3	4	3	22
	rock	39	10	15	9	73
	folk	10	22	8	8	48
	other	4	7	5	6	22
	total	78	61	62	70	271

Table 15.1: Basic ice cream and music contingency table.

show that, e.g., a total of 60 students say that rap is their favorite music type. The bottom margin sums over music types to show that, e.g., 70 students report that their favorite flavor of ice cream is neither chocolate, vanilla, nor strawberry. The total of either margin, 271, is sometimes called the "grand total" and represent the total number of subjects.

We can also see, from the margins, that rock is the best liked music genre, and classical is least liked, though there is an important degree of arbitrariness in this conclusion because the experimenter was free to choose which genres were in or not in the "other" group. (The best practice is to allow a "fill-in" if someone's choice is not listed, and then to be sure that the "other" group has no choices with larger frequencies than any of the explicit non-other categories.) Similarly, chocolate is the most liked ice cream flavor, and subject to the concern about defining "other", vanilla and strawberry are nearly tied for second.

Before continuing to discuss the form and content of contingency tables, it is good to stop and realize that the information in a contingency table represents results from a sample, and other samples would give somewhat different results. As usual, any differences that we see in the sample may or may not reflect real differences in the population, so you should be careful not to over-interpret the information in the contingency table. In this sense it is best to think of the contingency table as just a form of EDA. We will need formal statistical analyses to test hypotheses about the population based on the information in our sample.

Other information that may be present in a contingency table includes various percentages. So-called **row percents** add to 100% (in the right margin) for each

		favorite ice cream				
		chocolate	vanilla	strawberry	other	total
favorite music	rap	5 8.3%	10 17.7%	7 11.7%	38 63.3%	60 100%
	jazz	8 17.4%	9 19.6%	23 50.0%	6 13.0%	46 100%
	classical	12 54.5%	3 13.6%	4 18.2%	3 13.6%	22 100%
	rock	39 53.4%	10 13.7%	15 20.5%	9 12.3%	73 100%
	folk	10 20.8%	22 45.8%	8 16.7%	8 16.7%	48 100%
	other	4 18.2%	7 31.8%	5 22.7%	6 27.3%	22 100%
	total	78 28.8%	61 22.5%	62 22.9%	70 25.8%	271 100%

Table 15.2: Basic ice cream and music contingency table with row percents.

row of the table, and **column percents** add to 100% (in the bottom margin) for each column of the table.

For example, Table 15.2 shows the ice cream and music data with row percents. In R, you can obtain row and column percents using the `prop.table()` function, where specifying `margin = 1` within that function provides row percents, and specifying `margin = 2` provides column percents. If one variable is clearly an outcome variable, then the most useful and readable version of the table is the one with cell counts plus percentages that add up to 100% across all levels of the outcome for each level of the explanatory variable. This makes it easy to compare the outcome distribution across levels of the explanatory variable. In this example there is no clear distinction of the roles of the two measurements, so arbitrarily picking one to sum to 100% is a good approach.

Many important things can be observed from this table. First, we should look for the 100% numbers to see which way the percents go. Here we see 100% on the right side of each row. So for any music type we can see the frequency of each flavor answer, and those frequencies add up to 100%. We should think of those row percents as estimates of the true population probabilities of the flavors for each

given music type. In other words, they are estimates of conditional probabilities (i.e., the probability of flavors, conditional on music type).

Looking at the bottom (marginal) row, we know that, e.g., averaging over all music types, approximately 26% of students like "other" flavors best, and approximately 29% like chocolate best. Of course, if we repeat the study, we would get somewhat different results because each study looks at a different random sample from the population of interest.

In terms of the main hypothesis of interest, which is whether or not the two questions are independent of each other, it is equivalent to ask if the probabilities are evenly (or uniformly) distributed across the table. For example, although we will use statistical methods to assess independence, it is worthwhile to examine the row (or column) percentages for equality. In this table, we see rather large differences, e.g., chocolate is high for classical and rock music fans, but low for rap music fans, suggesting lack of independence.

A contingency table summarizes the data from an experiment or observational study with two or more categorical variables. Comparing a set of marginal percentages to the corresponding row or column percentages at each level of one variable is good EDA for checking independence.

15.2.3 Chi-square test of Independence

The most commonly used test of independence for the data in a contingency table is the **chi-square test of independence**. In this test the data from a k by m contingency table are reduced to a single statistic usually called either X^2 or χ^2 (chi-squared), although X^2 is better because statistics usually have Latin, not Greek letters. The null hypothesis is that the two categorical variables are independent, or equivalently that the distribution of either variable is the same at each level of the other variable. The alternative hypothesis is that the two variables are not independent, or equivalently that the distribution of one variable depends on (varies with) the level of the other.

If the null hypothesis of independence is true, then the X^2 statistic is **asymptotically distributed** as a chi-square distribution (see Section 3.9.6) with $(k -$

$1)(m-1)$ df. Under the alternative hypothesis of non-independence, the X^2 statistic will be larger on average. The p-value is the area under the null sampling distribution larger than the observed X^2 statistic. The term "asymptotically distributed" simply means "the shape of the distribution when the sample size is very, very large (i.e., infinitely large)." For example, if we were lucky enough to have a dataset that literally had an infinite number of rows, then the X^2 statistic for that dataset would have a chi-square distribution. As a result, the p-values are reliable for "large" sample sizes, but not for small sample sizes. Most textbooks quote a rule that no cell of the expected counts table (see below) can have less than five counts for the X^2 test to be reliable. This rule is conservative, and somewhat smaller counts also give reliable p-values.

Several alternative statistics are sometimes used instead of the chi-square statistic (e.g., likelihood ratio statistic or Fisher exact test), but these will not be covered here. It is important to realize that these various tests may disagree for small sample sizes and it is not necessarily clear which one is "correct".

The calculation of the X^2 statistic is based on the formula

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}}$$

where k and m are the number of rows and columns in the contingency table (i.e., the number of levels of the categorical variables), Observed_{ij} is the observed count for the cell with one variable at level i and the other at level j, and Expected_{ij} is the expected count based on independence. The basic idea here is that each cell contributes a non-negative amount to the sum, that a cell with an observed count very different from expected contributes a lot, and that "a lot" is relative to the expected count (denominator).

Although a computer program is ordinarily used for the calculation, an understanding of the principles is worthwhile. An "expected counts" table can be constructed by looking at either of the marginal percentages, and then computing the expected counts by multiplying each of these percentages by the total counts in the other margin. Table 15.3 shows the expected counts for the ice cream example. For example, using the percents in the bottom margin of Table 15.2, if the two variables are independent, then we expect 22.9% of people to like strawberry best among each group of people defined by their favorite music. Because 73 people like rock best, under the null hypothesis of independence, we expect (on average) $0.229 * 73 = 16.7$ people to like rock and strawberry best, as shown in Table 15.3.

		favorite ice cream				
		chocolate	vanilla	strawberry	other	total
favorite music	rap	17.3	13.5	13.7	15.5	60
	jazz	13.2	10.4	10.5	11.9	46
	classical	6.3	5.0	5.0	5.7	22
	rock	21.0	16.4	16.7	18.9	73
	folk	13.8	10.8	11.0	12.4	48
	other	6.3	5.0	5.0	5.7	22
	total	78	61	62	70	271

Table 15.3: Expected counts for ice cream and music contingency table.

Note that there is no reason that the expected counts should be whole numbers, even though observed counts must be.

By combining the observed data of Table 15.1 with the expected values of Table 15.3, we have the information we need to calculate the X^2 statistic. For the ice cream data we find that

$$X^2 = \left(\frac{(5-17.3)^2}{5}\right) + \left(\frac{(10-13.5)^2}{10}\right) + \cdots + \left(\frac{(6-5.7)^2}{6}\right) = 112.86.$$

So for the ice cream example, jazz paired with chocolate shows a big deviation from independence and of the 24 terms of the X^2 sum, that cell contributes $(5-17.3)^2/5 = 30.258$ to the total of 112.86. There are far fewer people who like that particular combination than would be expected under independence. To test if all of the deviations are consistent with chance variation around the expected values, we compare the X^2 statistic to the χ^2 distribution with $(6-1)(4-1) = 15$ df. This distribution has 95% of its probability below 25.0, so with $X^2 = 112.86$, we reject H_0 at the usual $\alpha = 0.05$ significance level. In fact, only 0.00001 of the probability is above 50.5, so the p-value is far less than 0.05. We reject the null hypothesis of independence of ice cream and music preferences in favor of the conclusion that the distribution of preference of either variable *does* depend on preference for the other variable.

You can choose among several ways to express violation (or non-violation) of the null hypothesis for a "chi-square test of independence" of two categorical variables. You should use the context of the problem to decide which one best expresses the relationship (or lack of relationship) between the variables. In this problem it

is correct to say any of the following: ice cream preference is not independent of music preference, or ice cream preference depends on or differs by music preference, or music preference depends on or differs by ice cream preference, or knowing a person's ice cream preference helps in predicting their music preference, or knowing a person's music preference helps in predicting their ice cream preference. However, note that a causal statement (e.g., "Liking rock music causes you to like chocolate ice cream,") is usually not warranted, unless some kind of randomized experiment was conducted. (In this example, it's very difficult to imagine how we could possibly randomize music and/or ice cream preferences.)

The chi-square test is based on a statistic that is large when the observed cell counts differ markedly from the expected counts under the null hypothesis condition of independence. The corresponding null sampling distribution is a chi-square distribution if no expected cell counts are too small.

Two additional points are worth mentioning in this abbreviated discussion of testing independence among categorical variables. First, because we want to avoid very small expected cell counts to assure the validity of the chi-square test of independence, it is common practice to combine categories with small counts into combined categories. Of course, this must be done in some way that makes sense in the context of the problem.

Second, when the contingency table is larger than 2 by 2, we need a way to perform the equivalent of contrast tests. One simple solution is to create subtables corresponding to the question of interest, and then to perform a chi-square test of independence on the new table. To avoid a high Type 1 error rate we need to make an adjustment, e.g., by using a Bonferroni correction, if this is post-hoc testing. For example to see if chocolate preference is higher for classical than jazz, we could compute chocolate vs. non-chocolate counts for the two music types to get table 15.4. This gives a X^2 statistic of 9.9 with 1 df, and a p-value of 0.0016. If this is a post-hoc test, we need to consider that there are 15 music pairs and 4 flavors plus 6 flavor pairs and 6 music types giving 4*15+6*6=96 similar tests, that might just as easily have been noticed as "interesting". The Bonferroni correction implies using a new alpha value of 0.05/96=0.00052, so because $0.0016 > 0.00052$, we cannot make the post-hoc conclusion that chocolate preference differs for jazz vs. classical. In other words, if the null hypothesis of independence is true, and we

		favorite ice cream		
		chocolate	not chocolate	total
favorite music	jazz	8 17.4%	38 82.6%	46 100%
	classical	12 54.5%	10 45.5%	22 100%
	total	20 29.4%	48 70.6%	68 100%

Table 15.4: Cross-tabulation of chocolate for jazz vs. classical.

data snoop looking for pairs of categories of one factor being different for presence vs. absence of a particular category of the other factor, finding that one of the 96 different p-values is 0.0016 is not very surprising or unlikely.

As a final point, we'll note that, to implement the chi-squared test in R, you use the `chisq.test()` function. This function takes a contingency table as an input; so, use the `table()` function in conjunction with the `chisq.test()` function.

15.3 Logistic regression

15.3.1 Introduction

Logistic regression is a flexible method for modeling and testing the relationships between one or more quantitative and/or categorical explanatory variables and one **binary** (i.e., two level) categorical outcome. The two levels of the outcome can represent anything, but generically we label one outcome "success" (usually coded as 1) and the other "failure" (usually coded as 0). Then, logistic regression attempts to estimate the probability of success conditional on the explanatory variables.

Logistic regression resembles ordinary linear regression in many ways. Besides allowing any combination of quantitative and categorical explanatory variables (with the latter in indicator variable form), it is appropriate to include functions of the explanatory variables such as log(x) when needed, as well as products of pairs of explanatory variables (or more) to represent interactions. In addition, there

is usually an intercept parameter (β_0) plus one parameter for each explanatory variable (β_1 through β_k), and these are used in the linear combination form: $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. We will call this sum **eta** (written η) for convenience.

Logistic regression differs from ordinary linear regression because its outcome is binary rather than quantitative. In ordinary linear regression the structural (means) model is that $E(Y) = \eta$. This is inappropriate for logistic regression because, among other reasons, the outcome can only take two arbitrary values (coded as 0 or 1), while eta can take any value. The solution to this dilemma is to use the means model

$$\log\left(\frac{E(Y)}{1-E(Y)}\right) = \log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \eta.$$

After some basic-but-tedious algebra, we can write the above equality as:

$$E(Y) = \frac{\exp(\eta)}{1+\exp(\eta)}$$

Here, $\exp(\cdot)$ denotes the natural exponential function, i.e., $\exp(A) = e^A$, where e is the natural constant. Because of the 0/1 coding, $E(Y)$, read as the "expected value of Y" is equivalent to the probability of success, and $1 - E(Y)$ is the probability of failure. Thus, logistic regression models the probability of success as a transformation of a linear combination of explanatory variables (where this linear combination is represented as eta).

The main benefit of using this transformation of eta is that it forces $E(Y)$ to be between 0 and 1; this is fundamentally different from Normal linear regression, where $E(Y)$ can be any value. By plugging in values for η, we can find that as η is made more and more negative, $E(Y)$ gets closer and closer to 0 (this is because $\exp(\eta)$ gets closer and closer to 0 as η is made more and more negative). Similarly, we can find that as η is made more and more positive, $E(Y)$ gets closer and closer to 1. Also, by noting that $\exp(0) = 1$, $\eta = 0$ corresponds to $E(Y) = \frac{1}{2}$, i.e., success and failure are equally likely. Finally, it is important to note that, for the jth explanatory variable, $\beta_j > 0$ corresponds to that variable being positively associated with the probability of success (i.e., increasing the jth explanatory variable will increase the probability of success, just not by β_j, due to the above transformation of eta). Thus, the signs of the coefficients (i.e., positive or negative) in logistic regression can still be qualitatively interpreted the same way they are interpreted in Normal linear regression.

The term $\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right)$ is often referred as the "log odds," where the odds is the ratio between the probability of success and the probability of failure. Thus, in logistic regression, the log odds is set to be equal to eta.

The means model for logistic regression is that the log odds of success equals a linear combination of the parameters and explanatory variables.

A shortcut term that is often used is **logit** of success, which is equivalent to the log odds of success. With this terminology the means model is logit(P(Y = 1))=η.

It takes some explaining and practice to get used to working with odds and log odds, but because this form of the means model is most appropriate for modeling the relationship between a set of explanatory variables and a binary categorical outcome, it's worth the effort.

First consider the term **odds**, which will always indicate the odds of success for us. By definition

$$\text{odds}(Y=1) = \frac{\Pr(Y=1)}{1-\Pr(Y=1)} = \frac{\Pr(Y=1)}{\Pr(Y=0)}.$$

The odds of success is defined as the ratio of the probability of success to the probability of failure. The odds of success (where Y=1 indicates success) contains the same information as the probability of success, but is on a different scale. Probability runs from 0 to 1 with 0.5 in the middle. Odds runs from 0 to ∞ with 1.0 in the middle. A few simple examples, shown in Table 15.5, make this clear. Note how the odds equal 1 when the probability of success and failure are equal. The fact that, e.g., the odds are 1/9 vs. 9 for success probabilities of 0.1 and 0.9 respectively demonstrates how 1.0 can be the "center" of the odds range of 0 to infinity.

Here is one way to think about odds. If the odds are 9 or 9/1, which is often written as 9:1 and read 9 to 1, then this tells us that for every nine successes there is one failure on average. For odds of 3:1, for every 3 successes there is one failure on average. For odds equal to 1:1, there is one failure for each success on average. For odds of less than 1, e.g., 0.25, write it as 0.25:1 then multiply the numerator

$\Pr(Y=1)$	$\Pr(Y=0)$	Odds	Log Odds
0	1	0	-∞
0.1	0.9	1/9	-2.197
0.2	0.8	0.25	-1.383
0.25	0.75	1/3	-1.099
1/3	2/3	0.5	-0.693
0.5	0.5	1	0.000
2/3	1/3	2	0.693
0.75	0.25	3	1.099
0.8	0.2	4	1.386
0.9	0.1	9	2.197
1	0	∞	∞

Table 15.5: Relationship between probability, odds and log odds.

and denominator by whatever number gives whole numbers in the answer. In this case, we could multiply by 4 to get 1:4, which indicates that for every one success there are four failures on average. As a final example, if the odds are 0.4, then this is 0.4:1 or 2:5 when I multiply by 5/5, so on average there will be five failures for every two successes.

To calculate probability, p, when you know the odds use the formula

$$p = \frac{\text{odds}}{1 + \text{odds}}.$$

The odds of success is defined as the ratio of the probability of success to the probability of failure. It ranges from 0 to infinity.

The **log odds** of success is defined as the natural (i.e., base e, not base 10) log of the odds of success. The concept of log odds is very hard for humans to understand, so we often "undo" the log odds to get odds, which are then more interpretable. Because the log is a natural log, we undo log odds by taking Euler's constant (e), which is approximately 2.718, to the power of the log odds. For example, if the log odds are 1.099, then we can find $e^{1.099}$ as exp(1.099) in most computer languages or in Google search to find that the odds are 3.0 (or 3:1).

The log odds scale runs from $-\infty$ to $+\infty$ with 0.0 in the middle. So zero represents the situation where success and failure are equally likely, positive log odds values represent a greater probability of success than failure, and negative log odds values represent a greater probability of failure than success. Importantly, because log odds of $-\infty$ corresponds to probability of success of 0, and log odds of $+\infty$ corresponds to probability of success of 1, the model "log odds of success equal eta" cannot give invalid probabilities as predictions for any combination of explanatory variables.

It is important to note that in addition to population parameter values for an ideal model, odds and log odds are also used for observed percent success. E.g., if we observe 5/25=20% successes, then we say that the (observed) odds of success is 0.2/0.8=0.25.

The log odds of success is simply the natural log of the odds of success. It ranges from minus infinity to plus infinity, and zero indicates that success and failure are equally likely.

As usual, any model prediction, which is the probability of success in this situation, applies for all subjects with the same levels of all of the explanatory variables. In logistic regression, we are assuming that for any such group of subjects the probability of success, which we can call p, applies individually and independently to each of the set of similar subjects. These are the conditions that define a binomial distribution (see Section 3.9.1). If we have n subjects all with the same level of the explanatory variables and with predicted success probability p, then our error model is that the outcomes will follow a random binomial distribution written as Binomial(n,p). The mean number of successes will be the product np, and the variance of the number of successes will be $np(1-p)$. Note that this indicates that there is no separate variance parameter (σ^2) in a logistic regression model; instead the variance varies with the mean and is determined by the mean. In fact, this is the case for pretty much every non-Normal distribution. The Normal distribution is unique in that the mean and variance are independent (the only exceptions being the t distribution and double exponential distribution; we've discussed the former in this class, but not the latter).

The error model for logistic regression is that for each fixed combination of explanatory variables, the distribution of success follows a binomial distribution with success probability $p = \frac{\exp(\eta)}{1+\exp(\eta)}$, where η denotes the sum $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ that we typically see in linear regression.

15.3.2 Example and EDA for logistic regression

The example that we will use for logistic regression is a simulated dataset (LRex.dat) based on a real experiment where the experimental units are posts to an Internet forum and the outcome is whether or not the message received a reply within the first hour of being posted. The outcome variable is called "reply" with 0 as the failure code and 1 as the success code. The posts are all to a single high volume forum and are computer generated. The time of posting is considered unimportant to the designers of the experiment. The explanatory variables are the length of the message (20 to 100 words), whether it is in the passive or active voice (coded as an indicator variable for the "passive" condition), and the gender of the fake first name signed by the computer (coded as a "male" indicator variable).

Plotting the outcome vs. one (or each) explanatory variable is not helpful when there are only two levels of outcome because many data points end up on top of each other. For categorical explanatory variables, cross-tabulating the outcome and explanatory variables is good EDA.

For quantitative explanatory variables, one reasonably good possibility is to break the explanatory variable into several groups (e.g., using the `cut()` function in R), and then to plot the mean of the explanatory variable in each bin vs. the observed fraction of successes in that bin (e.g., using the `aggregate()` function). Figure 15.1 shows a binning of the length variable vs. the fraction of successes with separate marks of "0" for active vs. "1" for passive voice. The curves are from a non-parametric smoother (loess) that helps in identifying the general pattern of any relationship. The main things you should notice are that active voice messages are more likely to get a quick reply, as are shorter messages.

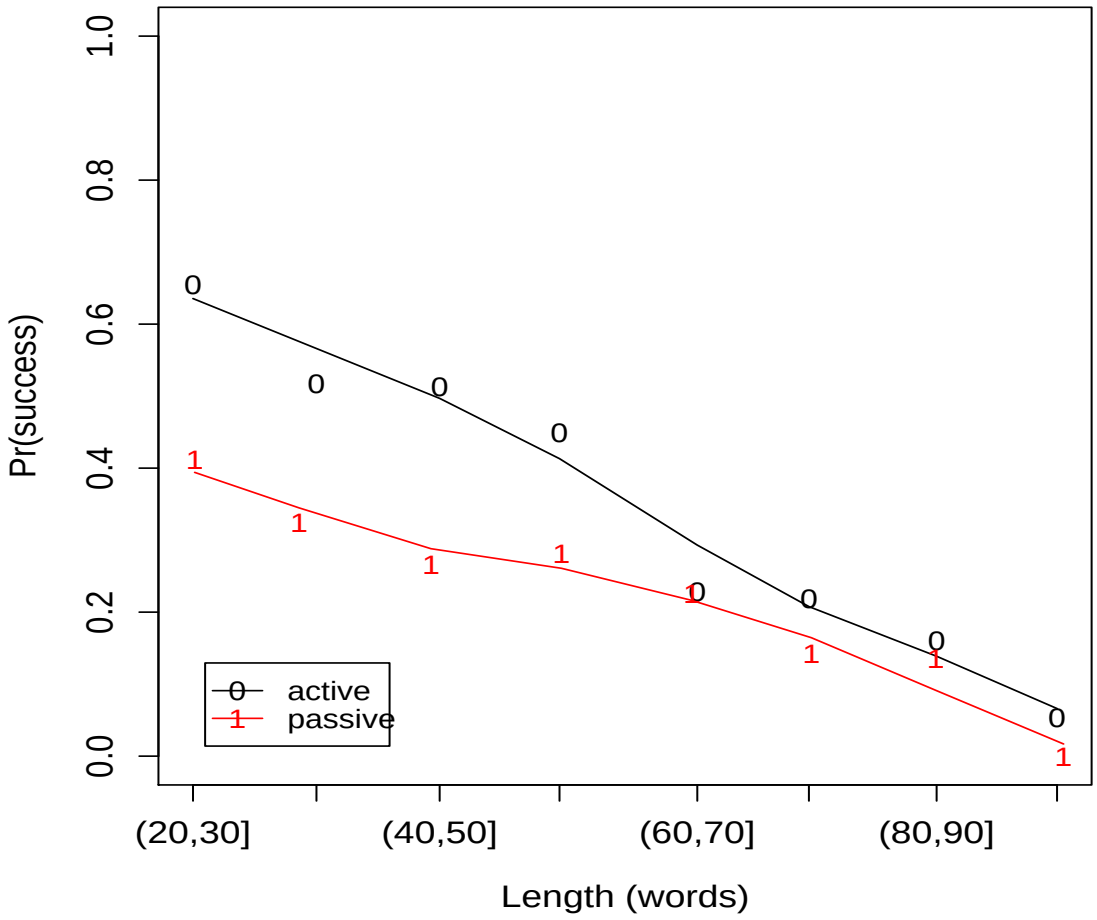


Figure 15.1: EDA for forum message example.

As always, EDA should be performed before fitting a model for binary outcomes. For categorical explanatory variables, it is common practice to cross-tabulate the number of successes and failures for each value of the explanatory variable. For quantitative explanatory variables, it is helpful to use binning techniques to plot the rate of successes across bins of the quantitative explanatory variable.

15.3.3 Fitting a logistic regression model

Remember that the means model in logistic regression is that

$$\text{logit}(\text{P}(\text{Y}=1)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

For any continuous explanatory variable, x_i, at any fixed levels of all of the other explanatory variables, this is linear on the logit scale. What does this correspond to on the more natural probability scale? It represents an "S" shaped curve that either rises or falls (monotonically, without changing direction) as x_i increases. If the curve is rising, as indicated by a positive sign on β_i, then it approaches $\text{Pr}(\text{Y} = 1)=1$ as x_i increases and $\text{Pr}(\text{Y} = 1)=0$ as x_i decreases. For a negative β_i, the curve starts near $\text{Pr}(\text{Y} = 1)=1$ and falls toward $\text{Pr}(\text{Y} = 1)=0$. This gives an intuitive interpretation to the coefficients: Positive coefficients suggest that increasing the explanatory variable tends to increase the probability of success, and negative coefficients suggest that increasing the explanatory variable tends to decrease the probability of success. Therefore, a logistic regression model is only appropriate if the EDA suggest a monotonically rising or falling curve. The curve need not approach 0 and 1 within the observed range of the explanatory variable, although it will at extreme values of that variable (extreme values that may not be observed in the data).

It is worth mentioning here that the magnitude of β_i is related to the steepness of the rise or fall, and the intercept denotes the population-level rate of success when all of the explanatory variables are set to zero. Thus, the interpretation of the intercept is extremely analogous to the interpretation we saw for Normal linear regression; the only difference is that we are now calling the "mean outcome" the "rate of success." Furthermore, similar to Normal linear regression, note that (1) the intercept may correspond to the population-level rate of success for the

"reference group" of a categorical variable, and (2) estimating the intercept may not make sense if setting all of the explanatory variables equal to zero does not make sense.

Fitting a logistic regression model involves the computer finding the best estimates of the β values, i.e., $\hat{\beta}$, which we saw in Normal linear regression. When it comes to interpreting estimated coefficients in logistic regression, there are many parallels to interpreting estimated coefficients in Normal linear regression: The intercept corresponds to a mean when all explanatory variables are equal to zero, slopes correspond to a change in a mean when a particular explanatory variable increases by 1, and coefficients for indicator variables correspond to a change in a mean for one group compared to the reference group. Furthermore, we can consider additive or interactive logistic regression models, where we can use the `anova()` function to determine which model is more appropriate. The key difference is that estimated coefficients for logistic regression are on what is called the *log-odds scale*, which makes interpretation slightly more nuanced and complicated. It is often more intuitive to put interpretations on the *odds* scale or the *probability* scale instead. As an example, Table 15.6 shows the estimated coefficients, their standard errors, and p-values for the null hypotheses that each parameter equals zero. Interpretation of this table is the subject of the next section.

15.3.4 Tests in a logistic regression model

The main interpretations that researchers conduct for a logistic regression model are for the parameters. Because the means model is

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

the interpretations are similar to those of ordinary linear regression, but the linear combination of parameters and explanatory variables gives the log odds of success rather than the expected outcome directly. For human interpretation we usually convert log odds to odds. As shown below, it is best to use the odds scale for interpreting coefficient parameters. For predictions, we can convert to the probability scale for easier interpretation.

Table 15.6 shows the estimated coefficients for the forum message example. Note that this table does not show an estimated coefficient for the gender variable; this is because we found the p-value for the gender variable to be non-significant (p-value=0.783), and thus we excluded it from our final model. This demonstrates

	$\hat{\beta}$	S.E.	Wald	df	Sig.	Exp(B)
Constant (Intercept)	1.384	0.308	20.077	1	<0.005	3.983
length	-0.035	0.005	46.384	1	<0.005	0.966
passive(1)	-0.744	0.212	12.300	1	<0.005	0.475

Table 15.6: Logistic regression coefficient estimates for the forum message example.

that, just like Normal linear regression, you can use statistical tests to determine which variables should be included in a given regression model. Furthermore, we can note that we see a coefficient for a `passive` indicator variable; thus, we know that "active voice" is the reference group for this logistic regression, and we should interpret the coefficient for the `passive` indicator variable accordingly (as we will describe shortly).

This model's prediction equation is

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_{\text{length}}(\text{length}) + \beta_{\text{passive}}(\text{passive})$$

and filling in the estimates we get

$$\text{logit}(\widehat{P(Y = 1)}) = 1.384 - 0.035(\text{length}) - 0.744(\text{passive}).$$

The intercept is the average log odds of success when all of the explanatory variables are zero. In this model this is the meaningless extrapolation to an active voice message with zero words. If this were meaningful, we could say that the estimated log odds for such messages is 1.384. To get to a more human scale we take $\exp(1.384)=e^{1.384}$ which is given in the last column of the table as 3.983 or 3.983:1. We can express this as approximately four successes for every one failure. We can also convert to the probability scale using the formula $p = \frac{3.983}{1+3.983} = 0.799$, i.e., an 80% chance of success.

The intercept estimate in logistic regression is an estimate of the log odds of success when all explanatory variables equal zero. If "all explanatory variables are equal to zero" is meaningful for the problem, you may want to convert the log odds to odds or to probability.

For a k-level categorical explanatory variable like "passive", R creates $k-1$ indicator variables and estimates $k-1$ coefficient parameters labeled $\hat{\beta}_{\text{x(1)}}$ through $\hat{\beta}_{\text{x(k-1)}}$. In this case we only have $\hat{\beta}_{\text{passive(1)}}$ because $k = 2$ for the passive variable. As usual, $\hat{\beta}_{\text{passive(1)}}$ represents the effect of increasing the explanatory variable by one-unit, and for an indicator variable this is a change from baseline to the specified non-baseline condition. The only difference from ordinary linear regression is that the "effect" is a change in the log odd of success.

For our forum message example, the estimate of -0.744 indicates that at any fixed message length, a passive message has a log odds of success 0.744 lower than a corresponding active message. For example, if the log odds of success for active messages for some particular message length is 1.744, then the log odds of success for passive messages of the same length is 1.000.

Because log odds is hard to understand we often rewrite the prediction equation as something like

$$\text{logit}(\widehat{P(Y=1)}) = \hat{\beta}_{0L} - 0.744(\text{passive})$$

where $\hat{\beta}_{0L} = 1.384 - 0.035L$ for some fixed message length, L. Then we exponentiate both sides to get

$$\text{odds}(\widehat{P(Y=1)}) = e^{\hat{\beta}_{0L}} e^{-0.744(\text{passive})}.$$

The left hand side of this equation is the estimate of the odds of success. Because $e^{-0.744} = 0.475$ and $e^0 = 1$, this says that for active voice $\text{odds}(\widehat{P(Y=1)}) = e^{\hat{\beta}_{0L}}$ and for passive voice $\text{odds}(\widehat{P(Y=1)}) = 0.475e^{\hat{\beta}_{0L}}$. In other words, at any message length, compared to active voice, the odds of success are *multiplied* (not added) by 0.475 to get the odds for passive voice. To put it yet another way, the coefficients in logistic regression are additive on the log-odds scale but multiplicative on the odds scale.

So the usual way to interpret the effect of a categorical variable on a binary outcome is to look at $\exp(\hat{\beta})$ and take that as the multiplicative change in odds when comparing the specified level of the indicator variable to the baseline level. If $\hat{\beta} = 0$ and therefore $\exp(\hat{\beta}) = 1$, then there is no effect of that variable on the outcome (and the p-value will be non-significant). If $\exp(\hat{\beta}) > 1$, then the odds increase for the specified level compared to the baseline. If $\exp(\hat{\beta}) < 1$, then the odds decrease for the specified level compared to the baseline. In our example, 0.475 is less than 1, so passive voice, compared to active voice, lowers the odds (and therefore probability) of success at each message length.

It is worth noting that multiplying the odds by a fixed number has very different effects on the probability scale for different baseline odds values. As a result, the interpretation of the coefficients in logistic regression is a bit more nuanced than the interpetation of coefficients in Normal linear regression. For example, in Normal linear regression, we interpreted a given $\hat{\beta}$ as, "For every one-unit increase in x, the mean outcome is expected to increase by $\hat{\beta}$, holding all other explanatory variables fixed." Because the effect of $\hat{\beta}$ is *multiplicative* on the odds (instead of additive), the interpretation of $\hat{\beta}$ for logistic regression is, "For every one-unit increase in x, the odds of success are expected to multiply by $\hat{\beta}$, holding all other explanatory variables fixed." For example, if $\hat{\beta} = 2$, then we are estimating that the odds double for each one unit increase in x. This means that if the baseline odds are 0.5 or 2 or 9 (with probabilities 0.333, 0.667, and 0.9, respectively), then a one-unit increase in x changes the odds to 1, 4, and 18 respectively (with probabilities 0.5, 0.8, and 0.95 respectively). From this example, we can see that this kind of multiplicative effect leads to a bigger change (on the probability scale) for midrange probabilities than for more extreme probabilities. This fundamentally differs from Normal linear regression, where the effect of increasing x by one was the same regardless of the mean outcome. To put the message of this paragraph in visual terms: This is the consequence of working with an S-shaped curve that gradually drifts to 0 and 1 (as in logistic regression) instead of a straight line that steadily shoots off to $-\infty$ and ∞ (as in Normal linear regression).

The estimate of the coefficient for an indicator variable of a categorical explanatory variable in a logistic regression is in terms of $\exp(\hat{\beta})$. This is the *multiplicative* change in the odds of success for the named vs. the baseline condition when all other explanatory variables are held constant.

For a quantitative explanatory variable, the interpretation of the coefficient estimate is quite similar to the case of a categorical explanatory variable. The differences are that there is no baseline, and that x can take on any value, not just 0 and 1. In general, the coefficient for a given continuous explanatory variable represents the additive change in log-odds of success when the explanatory variable increases by one unit (with all other explanatory variables held fixed). This corresponds to an $\exp(\hat{\beta})$ multiplicative change in the odds of success.

For our forum message example, our estimate is that when the voice is fixed at

either active or passive, the log odds of success (getting a reply within one hour) decreases by 0.035 for each additional word. In other words, the odds are multiplied by exp(0.035) = 0.966 (making them slightly smaller) for each additional word.

The p-value for each coefficient is a test of the null hypothesis that $\beta_x = 0$. If $\beta_x = 0$, then when x goes up by 1, the log odds go up by 0 and the odds get multiplied by exp(0)=1. In other words, if the coefficient is not significantly different from zero, then changes in that explanatory variable do not affect the outcome.

For a continuous explanatory variable in logistic regression, $\exp(\hat{\beta})$ **is the multiplicative change in odds of success for a one-unit increase in the explanatory variable.**

15.3.5 Predictions in a logistic regression model

Predictions in logistic regression are analogous to ordinary linear regression. First, create a prediction equation using the intercept (constant) and one coefficient for each explanatory variable (including $k-1$ indicators for a k-level categorical variable). Plug in the estimates of the coefficients and a set of values for the explanatory variables to get what we called η, above. This is your prediction of the log odds of success. Take $\exp(\eta)$ to get the odds of success, then compute $\frac{\text{odds}}{1+\text{odds}}$ to get the probability of success. Graphs of the probability of success vs. levels of a quantitative explanatory variable, with all other explanatory variable fixed at some values, will be S-shaped (or its mirror image), and are a good way to communicate what the means model represents.

As an example, for our forum messages data, we can compute the predicted log odds of success for a 30 word message in passive voice as $\eta = 1.384 - 0.035(30) - 0.744(1) = -0.41$. Then the odds of success for such a message is exp(-0.41)=0.664, and the probability of success is 0.664/1.664=0.40 or 40%. Computing this probability for all message lengths from 20 to 100 words separately for both voices gives Figure 15.2, which is a nice summary of the means model.

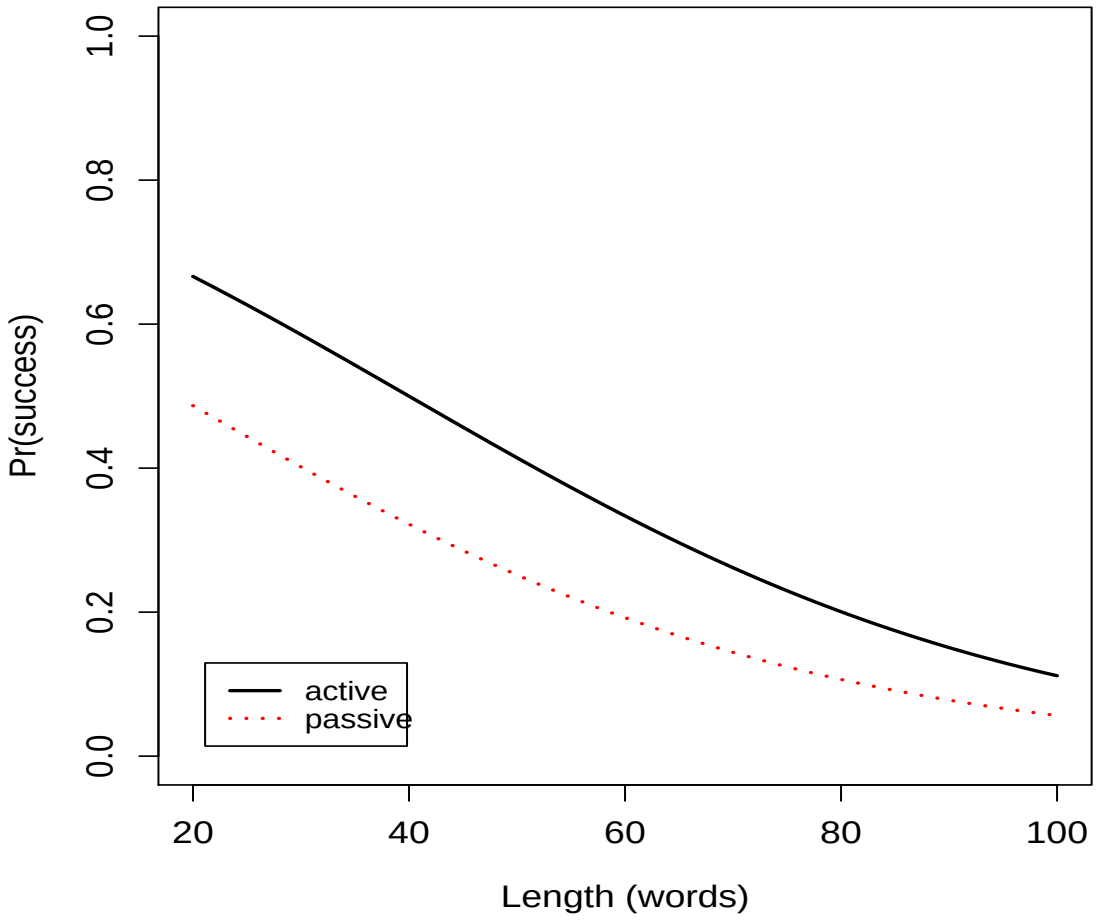


Figure 15.2: Model predictions for forum message example.

Prediction of probabilities for a set of explanatory variables involves calculating log odds from the linear combination of coefficient estimates and explanatory variables, then converting to odds and finally probability.

15.3.6 Do it in R

To implement the chi-squared test in R, you use the `chisq.test()` function, which takes a contingency table as an input. So, use the `table()` function in conjunction with the `chisq.test()` function when implementing the chi-squared test in R.

Implementing logistic regression is nearly identical to implementing Normal linear regression in R; the only differences are:

1. Instead of using the `lm()` function, you use the `glm()` function (where "glm" standards for "generalized linear model").
2. Within the `glm()` function, you need to specify the argument `family = "binomial"`.

As a demonstration, here is the code you would write to implement logistic regression for the messages dataset:

```
> summary(glm(reply ~ length + passive + male, data = messages,
    family = "binomial"))

Call:
glm(formula = reply ~ length + passive + male, family = "binomial"
    ,
    data = messages)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4665  -0.8317  -0.5516   1.0660   2.2173

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.347446   0.332780   4.049 5.14e-05 ***
length      -0.034469   0.005083  -6.781 1.19e-11 ***
```

```
passive      -0.740765   0.212315  -3.489 0.000485 ***
male          0.057886   0.210520   0.275 0.783341
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 605.69  on 499  degrees of freedom
Residual deviance: 540.80  on 496  degrees of freedom
AIC: 548.8

Number of Fisher Scoring iterations: 4
```

Here, `length` is a quantitative explanatory variable (denoting the word length of a message), and `passive` and `male` are indicator variables (i.e., binary categorical explanatory variables). Furthermore, `reply` is an indicator variable, where `reply = 1` denotes a message getting a reply. In other words, `reply` is a binary outcome, suggesting that we should use logistic regression, as we've done above.

When implementing logistic regression, it's fairly intuitive to change `lm()` to `glm()` - instead of performing linear regression, you're performing generalized linear regression. However, when implmenting logistic regression, people very commonly forget to write `family = "binomial"` within the `glm()` function. It's very important that you don't forget this! If we excluded the `family = "binomial"`, the `glm()` function by default runs linear regression instead of logistic regression. There are several specifications for the `family` argument; for example, specifying `family = "poisson"` runs Poisson regression (which is useful when the outcome is a count), and specifying `family = "Gamma"` runs Gamma regression (which is useful when the outcome is continuous but right-skewed). We will not discuss these types of regressions in further detail here.

Notice that the `male` variable is non-significant (p-value=0.783), suggesting that we can remove it from the model. Indeed, we could alternatively have used the `anova()` function to test whether we should include the `male` variable in our model:

```
#model without male
> glm1 = glm(reply ~ length + passive, data = messages, family = "
    binomial")
#model with male
> glm2 = glm(reply ~ length + passive + male, data = messages,
    family = "binomial")
```

```
#use the ANOVA function to determine whether
#we should include male in our model
> anova(glm1, glm2, test = "Chisq")
Analysis of Deviance Table

Model 1: reply ~ length + passive
Model 2: reply ~ length + passive + male
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       497     540.87
2       496     540.80  1 0.075577   0.7834
```

First, note that we had to specify the argument `test = "Chisq"` in order to obtain a p-value from `anova()`; this is something that is unique to generalized linear models when using this function. Furthermore, notice that the p-value from `anova()` is exactly the same as the original `male` p-value. This isn't a coincidence: Just like in linear regression, in logistic regression the `anova()` function assesses if additional coefficients in a more complex model are equal to zero. In this case, the more complex model (`glm2`) only has a single additional coefficient (the coefficient for `male`); thus, the p-value corresponds to testing if the coefficient for `male` is equal to zero.

Because of the above output, we focus on interpreting the `glm1` model, which excludes `male` from the model:

```
> summary(glm1)

Call:
glm(formula = reply ~ length + passive, family = "binomial",
    data = messages)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4813  -0.8386  -0.5580   1.0767   2.2314

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.382103   0.308456   4.481 7.44e-06 ***
length      -0.034561   0.005075  -6.811 9.72e-12 ***
passive     -0.743788   0.212074  -3.507 0.000453 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

(Dispersion parameter for binomial family taken to be 1)
```

```
20
21     Null deviance: 605.69  on 499  degrees of freedom
22 Residual deviance: 540.87  on 497  degrees of freedom
23 AIC: 546.87
24
25 Number of Fisher Scoring iterations: 4
```

This is exactly the output we saw in Table 15.6; we already provided interpretation for this output, so we won't repeat that interpretation here.

Index