Experimental Design and Analysis

Howard J. Seltman

September 8, 2015

Preface

This book is intended as required reading material for my course, Experimental Design for the Behavioral and Social Sciences, a second level statistics course for undergraduate students in the College of Humanities and Social Sciences at Carnegie Mellon University. This course is also cross-listed as a graduate level course for Masters and PhD students (in fields other than Statistics), and supplementary material is included for this level of study.

Over the years the course has grown to include students from dozens of majors beyond Psychology and the Social Sciences and from all of the Colleges of the University. This is appropriate because Experimental Design is fundamentally the same for all fields. This book tends towards examples from behavioral and social sciences, but includes a full range of examples.

In truth, a better title for the course is Experimental Design and Analysis, and that is the title of this book. Experimental Design and Statistical Analysis go hand in hand, and neither can be understood without the other. Only a small fraction of the myriad statistical analytic methods are covered in this book, but my rough guess is that these methods cover 60%-80% of what you will read in the literature and what is needed for analysis of your own experiments. In other words, I am guessing that the first 10% of all methods available are applicable to about 80% of analyses. Of course, it is well known that 87% of statisticians make up probabilities on the spot when they don't know the true values. :)

Real examples are usually better than contrived ones, but real experimental data is of limited availability. Therefore, in addition to some contrived examples and some real examples, the majority of the examples in this book are based on simulation of data designed to match real experiments.

I need to say a few things about the **difficulties of learning** about experimental design and analysis. A practical working knowledge requires understanding many concepts and their relationships. Luckily much of what you need to learn agrees with common sense, once you sort out the terminology. On the other hand, there is no ideal logical order for learning what you need to know, because everything relates to, and in some ways depends on, everything else. So be aware: many concepts are only loosely defined when first mentioned, then further clarified later when you have been introduced to other related material. Please try not to get frustrated with some incomplete knowledge as the course progresses. If you work hard, everything should tie together by the end of the course.

In that light, I recommend that you create your own "concept maps" as the course progresses. A concept map is usually drawn as a set of ovals with the names of various concepts written inside and with arrows showing relationships among the concepts. Often it helps to label the arrows. Concept maps are a great learning tool that help almost every student who tries them. They are particularly useful for a course like this for which the main goal is to learn the relationships among many concepts so that you can learn to carry out specific tasks (design and analysis in this case). A second best alternative to making your own concept maps is to further annotate the ones that I include in this text.

This book is on the world wide web at http://www.stat.cmu.edu/∼hseltman/309/Book/Book.pdf and any associated data files are at http://www.stat.cmu.edu/∼hseltman/309/Book/data/.

One key idea in this course is that you cannot really learn statistics without doing statistics. Even if you will never analyze data again, the hands-on experience you will gain from analyzing data in labs, homework and exams will take your understanding of and ability to read about other peoples experiments and data analyses 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 SPSS, mostly because it is one of the packages most likely to be available to you in your future schooling and work. You will find a chapter on learning to use SPSS in this book. In addition, many of the other chapters end with "How to do it in SPSS" sections.

There are some typographical conventions you should know about. First, in a non-standard way, I use capitalized versions of Normal and Normality because I don't want you to think that the Normal distribution has anything to do with the ordinary conversational meaning of "normal".

Another convention is that optional material has a gray background:

I have tried to use only the minimally required theory and mathematics for a reasonable understanding of the material, but many students want a deeper understanding of what they are doing statistically. Therefore material in a gray box like this one should be considered optional extra theory and/or math.

Periodically I will summarize key points (i.e., that which is roughly sufficient to achieve a B in the course) 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 re-read.

Less often I will 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.

Chapter 1 is an overview of what you should expect to learn in this course. Chapters 2 through 4 are a review of what you should have learned in a previous course. Depending on how much you remember, you should skim it or read through it carefully. Chapter 5 is a quick start to SPSS. Chapter 6 presents the statistical foundations of experimental design and analysis in the case of a very simple experiment, with emphasis on the theory that needs to be understood to use statistics appropriately in practice. Chapter 7 covers experimental design principles in terms of preventable threats to the acceptability of your experimental conclusions. Most of the remainder of the book discusses specific experimental designs and corresponding analyses, with continued emphasis on appropriate design, analysis and interpretation. Special emphasis chapters include those on power, multiple comparisons, and model selection.

You may be interested in my background. I obtained my M.D. in 1979 and practiced clinical pathology for 15 years before returning to school to obtain my PhD in Statistics in 1999. As an undergraduate and as an academic pathologist, I carried

out my own experiments and analyzed the results of other people's experiments in a wide variety of settings. My hands on experience ranges from techniques such as cell culture, electron auto-radiography, gas chromatography-mass spectrometry, and determination of cellular enzyme levels to topics such as evaluating new radioimmunoassays, determining predictors of success in in-vitro fertilization and evaluating the quality of care in clinics vs. doctor's offices, to name a few. Many of my opinions and hints about the actual conduct of experiments come from these experiences.

As an Associate Research Professor in Statistics, I continue to analyze data for many different clients as well as trying to expand the frontiers of statistics. I have also tried hard to understand the spectrum of causes of confusion in students as I have taught this course repeatedly over the years. I hope that this experience will benefit you. I know that I continue to greatly enjoy teaching, and I am continuing to learn from my students.

Howard Seltman
August 2008

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*

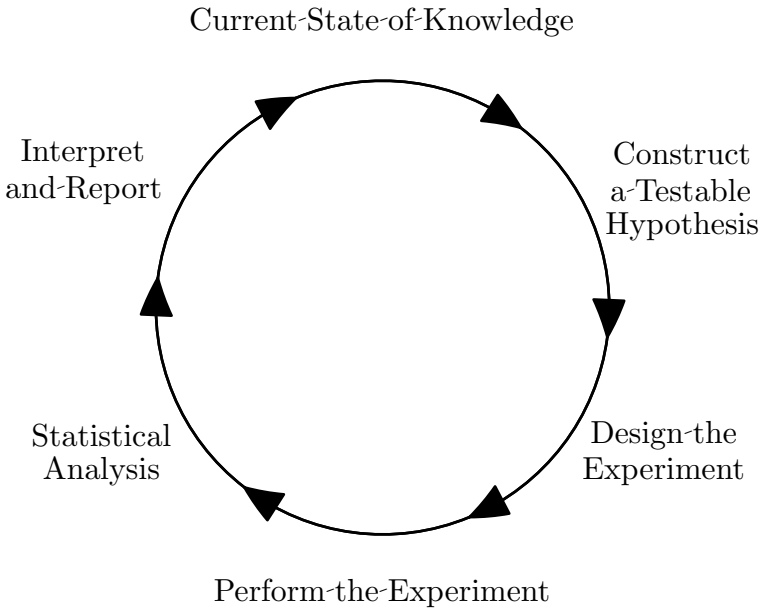


Figure 1.1: The circular flow of scientific learning

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 "nothing interesting going on" usually in terms of some effect 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", "Perform Informal and Formal Statistical Analyses", 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. It is only rarely in these circumstances that statistical analysis can rescue the experimenter. This is an example of the old maxim "an ounce of prevention is worth a pound of cure".

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.

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 assumption violation, 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.

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

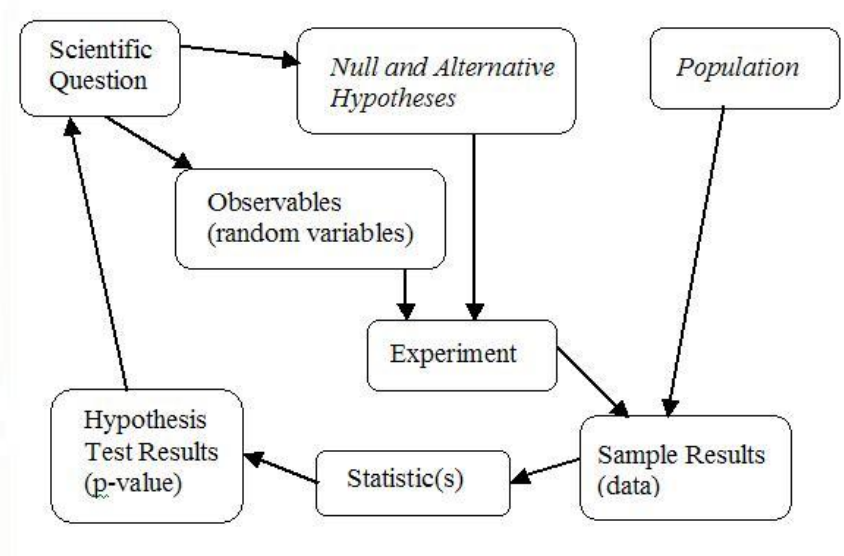


Figure 1.2: An oversimplified concept map.

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 the experience that becomes "data-sense". This requires learning to use a specific statistical computer package. Many excellent packages exist and are suitable for this purpose. Examples here come

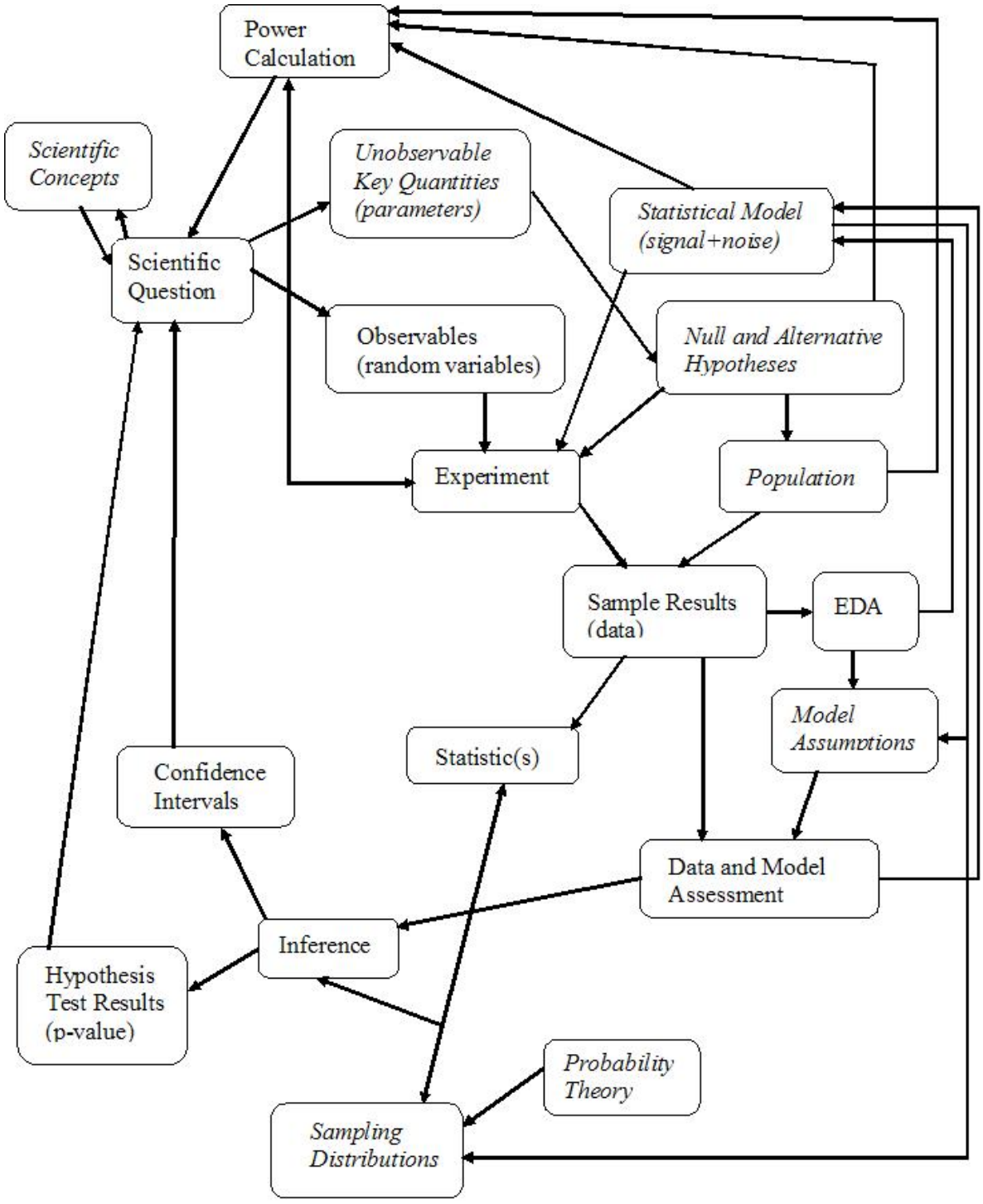


Figure 1.3: A reasonably complete concept map for this course.

from SPSS, but this is in no way an endorsement of SPSS over other packages.

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.

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.

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".

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 devis-

ing 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 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.

Warning: SPSS uses "type" to refer to the storage mode (as in computer science) of a variable. In a somewhat non-standard way it uses "measure" for what we are calling statistical type here.

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

Classification of 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. 32, 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. Other 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).

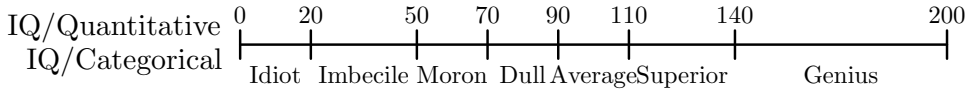


Figure 2.1: Old IQ categorization

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.

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 making predictions about 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, 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 out the five different probabilities and use the first form for the pmf. 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 the list form. (Again the mathematical details are optional.)

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 continuous 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 of all possible outcomes for a given experiment 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 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 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 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.

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

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 for disjoint unions is really a special case of the general rule for the probability that the outcome of an experiment will fall in a set that is the union of two other sets. Using the above 5-question test example, we can define 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 $\Pr(E \cap F) = \Pr(E)\Pr(F)$, 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 the chapter come in very 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 8.3 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 might 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 must use samples in science, it is very important to remember that we are interested in learning about populations, not samples. Inference from samples to populations is the goal of statistical analysis.

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 symbol always represents 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 include the mean and median parameters. The mean (expected value) of a random variable can be thought of as the "balance point" of the distribution if the pdf is cut out of cardboard. Or if the outcome is some monetary payout, the mean is the appropriate amount to bet to come out even in the long term. Another interpretation of mean is the "fair distribution of outcome" in the sense that if we sample many values and think of them as one outcome per subject, the mean is result of a fair redistribution of whatever the outcome represents among all of the subjects. 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 that 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) and the population kurtosis is a measure of peakedness or flatness compared to a Gaussian distribution, which has $\gamma_2 = 0$. 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, 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 called **moment** estimators. They are respectively the 1^{st} through 4^{th} (central) moments. Even simpler are the non-central moments: the r^{th} non-central moment of X is the expected value of X^r. There are formulas for calculating central moments from non-central moments. E.g., $\sigma^2 = E(X^2) - E(X)^2$.

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.

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.

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

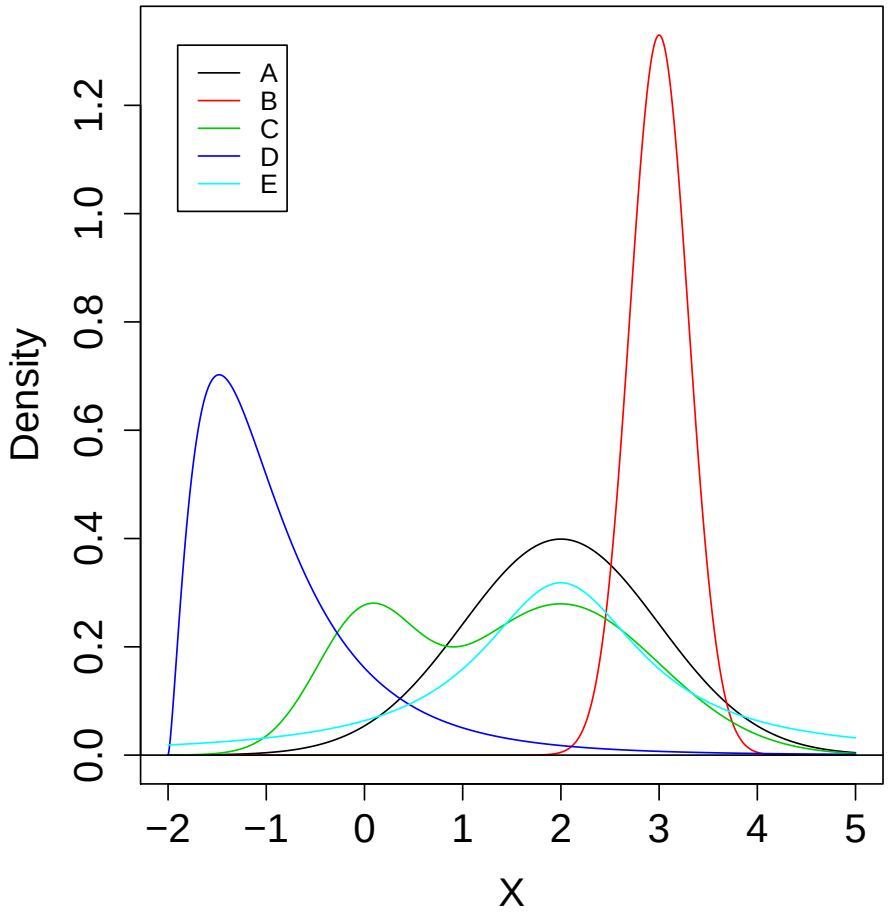


Figure 3.1: Various probability density function

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 obvious margins to the previous table, by adding the rows and columns and putting the sums in the margins (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).

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 31). As shown in table 3.5, the conditional distribution refers to fixing the level of one variable, then "re-normalizing" to find the probability level of the other variable when we only focus on or consider those units or subjects that meeting the condition of interest.

So if we focus on Japanese cars only (technically, we condition on cars be-

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.

ing 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 "the probability of a car being small given that the car is Japanese equals 0.571".

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 joint distribution, but the details will not be given here. Suffice it to say the 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 direct relationship while negative numbers indicate an indirect 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. (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 correlations 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 simpler 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.

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.

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.

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 8.3.) 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 statistics 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, 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 gender, and we will call this X in general and X_1, X_2 and X_3 in particular. Lets say that we know the parameter that represent 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 3.7. For example, the

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			

Table 3.9: Multivariate pmf for animal gender.

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 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 above), 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. 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.

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:

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

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 (out of 2) 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 $X \sim \mathrm{MN}(n, p_1, \ldots, p_k)$.

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 a subject clicks on a moving target on a computer screen in one minute, 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 is 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

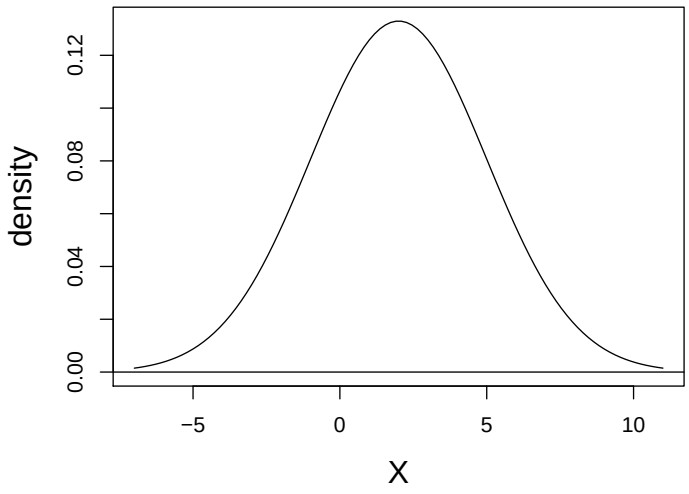


Figure 3.2: Gaussian bell-shaped probability density function

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, above). Also the Normal distribution is quite flexible in that the center and spread can be set to any values independently. On the other hand, every distribution that subjectively looks "bell-shaped" is not a Normal distribution. Some distributions are flatter than Normal, with "thin tails" (negative kurtosis). Some distributions are more "peaked" than a true 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 4.6. The support is all of the real numbers. The t-distributions have fatter tails than the normal distribution, but approach the shape of the normal distribution as the df increase. The t-distribution arises most commonly when evaluating how far a sample mean is from a population mean when the standard deviation of the sampling distribution is estimated from the data rather than known. It is the fact that the standard deviation is an estimate (i.e., a standard error) rather than the true value 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, commonly arises from the sum of squares of a independent N(0,1) random variables. 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 $X \sim \mathrm{F}(r, s)$.

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.

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 calculation of summary statistics, while graphical methods obviously 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.*

Beyond the four categories created by the above cross-classification, each of the categories of EDA have further divisions based on the role (outcome or explanatory) and type (categorical or quantitative) of the variable(s) being examined.

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, mostly to investigate the relationships between the 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 the only 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 H&SS, MCS, SCS and "other", then there is a true population of all students enrolled in the 2007 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 H&SS, H&SS, MCS, other, other, SCS, MCS, other, H&SS, MCS, SCS, SCS, other, MCS, MCS, H&SS, MCS, other, H&SS, SCS. Our EDA would look like this:

Statistic/College	**H&SS**	**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 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.*

What we observe in the **sample** of measurements for a particular variable that we select for our particular experiment is the "sample distribution". We need to recognize that this would be different each time we might 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 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 distribution 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 distribution of the population 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 here with the quantitative 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 outside of the areas of a distribution that would commonly occur. 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 4.3.3). Another common definition of "outlier" consider 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 looking at the central tendency, spread, skewness, and kurtosis of the data for a particular variable from an experiment. *But for categorical variables, none of these make any sense.*

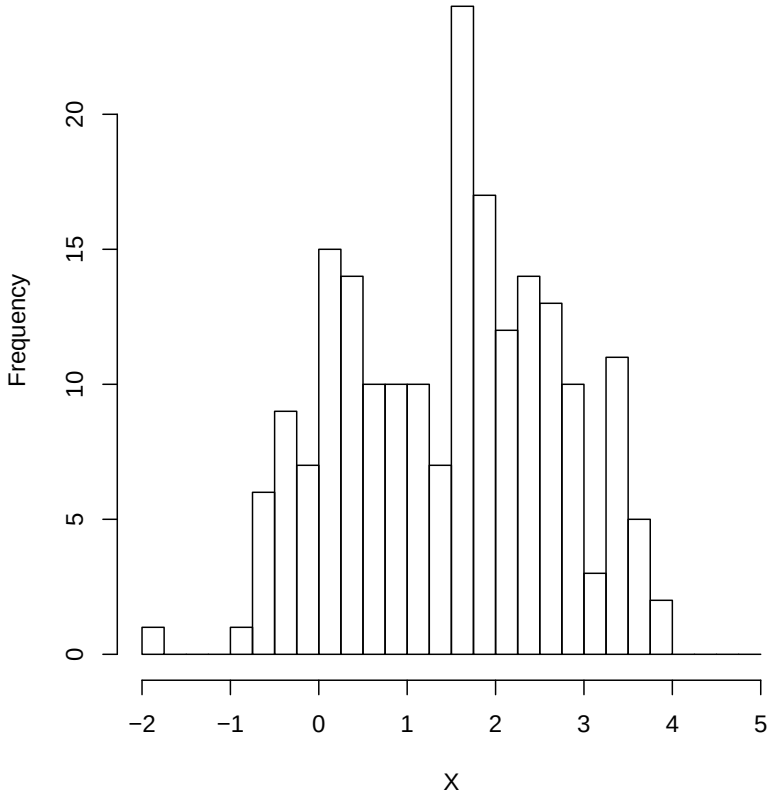


Figure 4.1: Histogram from distribution C.

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 "balance point": 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 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. Often we can use probability theory to work out the exact distribution of the sample statistic, at least under certain assumptions.

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 "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. Nevertheless, both numbers are "correct", as long as you understand their meanings.

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. More practically, a few very high values or very low values usually have no effect on the median.

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 4.6).

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, half of the values (and specifically the middle half) 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 compared to a Gaussian distribution.

4.3 Univariate graphical EDA

If we are focusing on data from observation of 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 required.

4.3.1 Histograms

The only one of these techniques that makes sense for categorical data is the histogram (basically just a barplot of the tabulation of the data). A pie chart

is equivalent, but not often used. The concepts of central tendency, spread and skew have no meaning for nominal categorical data. For ordinal categorical data, it sometimes makes sense to treat the data as quantitative for EDA purposes; you need to use your judgment here.

The most basic graph is the **histogram**, which is a barplot in which each bar represents 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 SPSS are pretty good, they are certainly not always adequate.

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

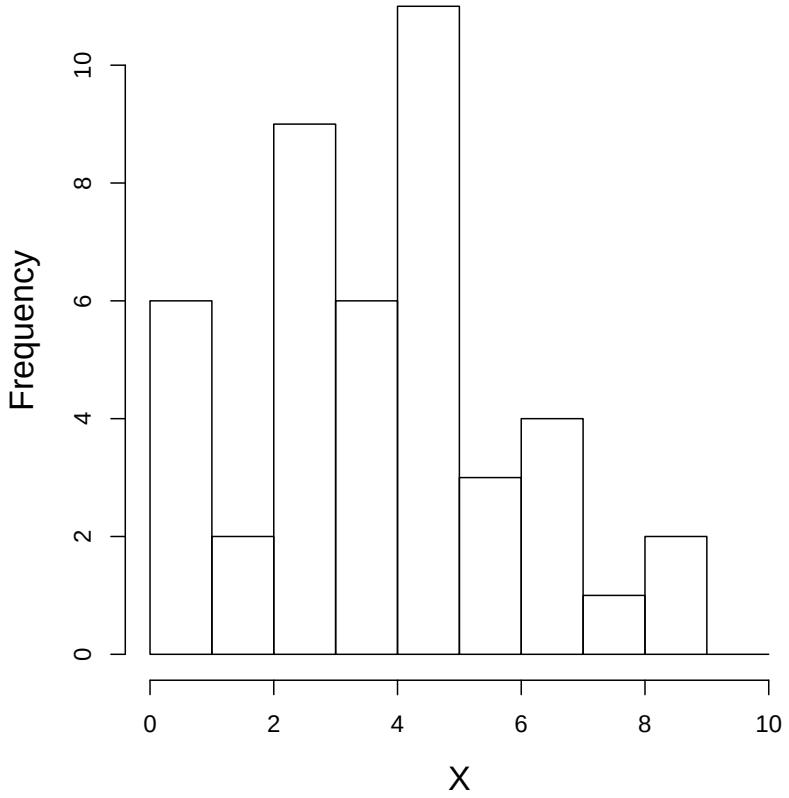


Figure 4.2: Histogram of EDA1.dat.

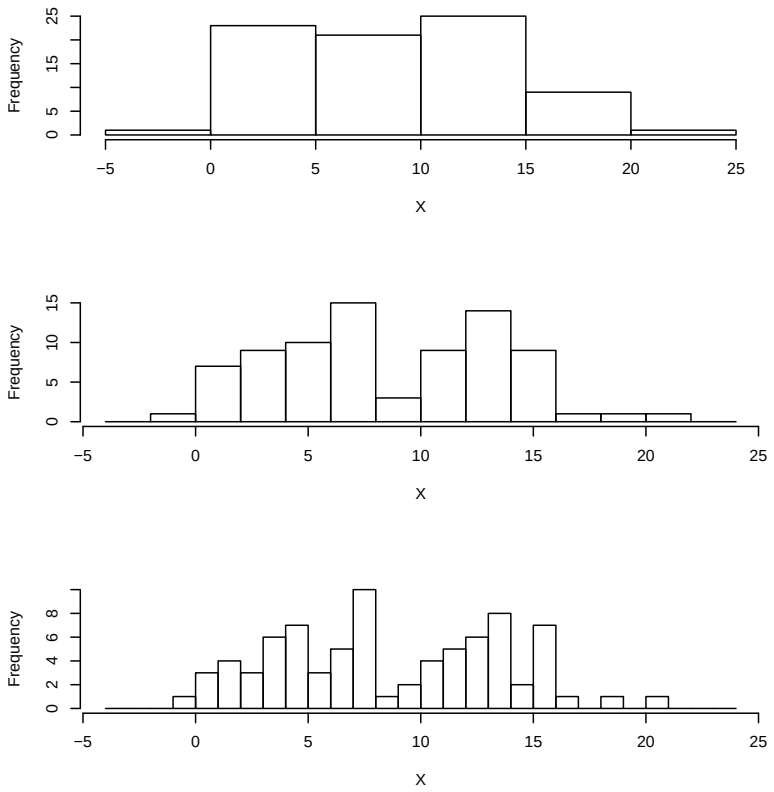


Figure 4.3: Histograms of EDA2.dat with different bin widths.

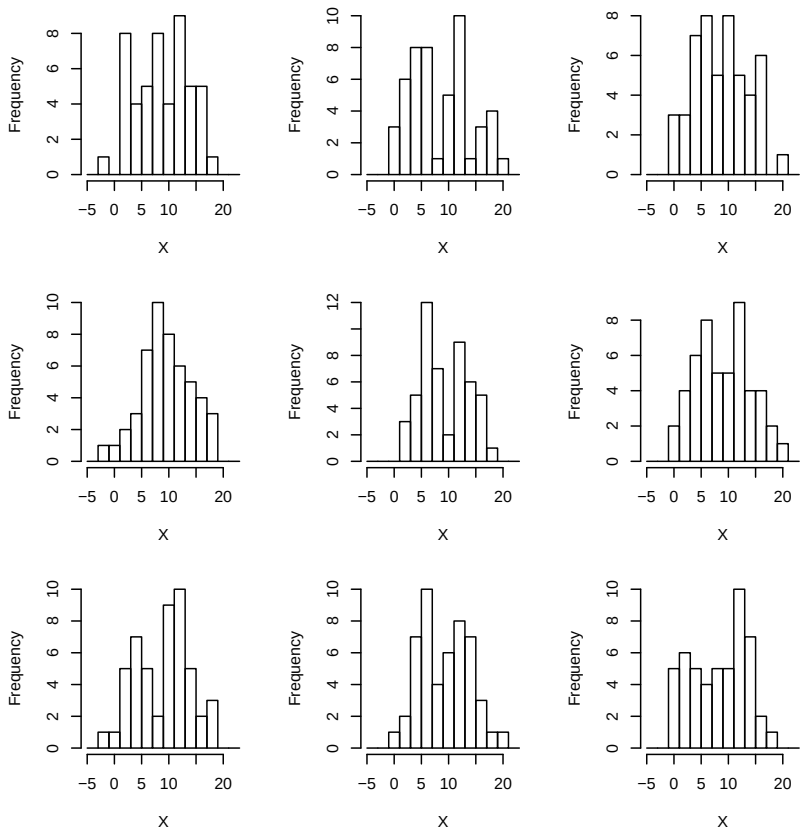


Figure 4.4: Histograms of multiple samples of size 50.

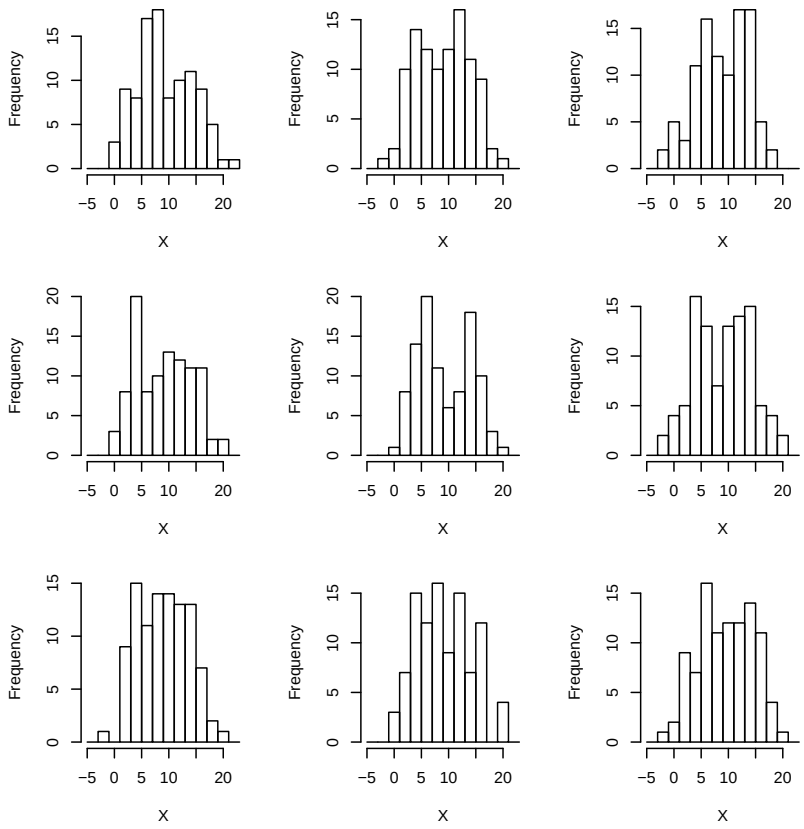


Figure 4.5: Histograms of multiple samples of size 100.

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.

4.3.2 Stem-and-leaf plots

A simple substitute for a histogram is a **stem and leaf plot**. A stem and leaf plot is sometimes easier to make by hand than a histogram, and it tends not to hide any information. Nevertheless, a histogram is generally considered better for appreciating the shape of a sample distribution than is the stem and leaf plot. Here is a stem and leaf plot for the data of figure 4.2:

```
The decimal place is at the "|".
1|000000
2|00
3|0000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00
```

Because this particular stem and leaf plot has the decimal place at the stem, each of the 0's in the first line represent 1.0, and each zero in the second line represents 2.0, etc. So we can see that there are six 1's, two 2's etc. in our data.

A stem and leaf plot shows all data values and the shape of the distribution.

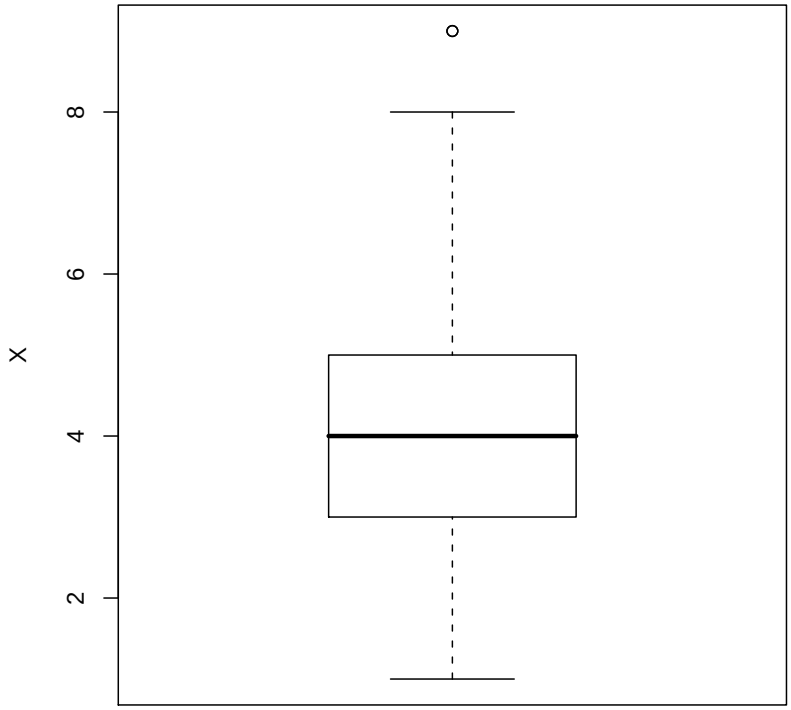


Figure 4.6: A boxplot of the data from EDA1.dat.

4.3.3 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 (see multivariate graphical analysis below).

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 represent the quartiles Q3 and Q1 respectively, and with a horizontal "median"

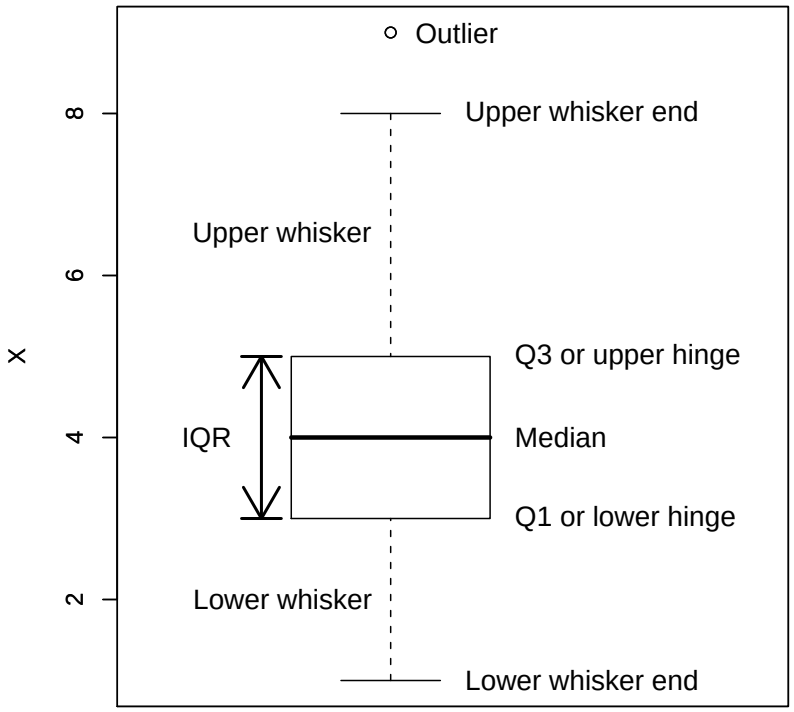


Figure 4.7: Annotated boxplot.

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 or half less than 4, 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 just a bit 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 starplots 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 mistakes or otherwise unusual. Also, points not designated as boxplot outliers may also be mistakes. 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 is 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, below) with a high degree of reliability because of the use of these robust statistics.

It is worth noting that some (few) programs produce boxplots that do not conform to the definitions given here.

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

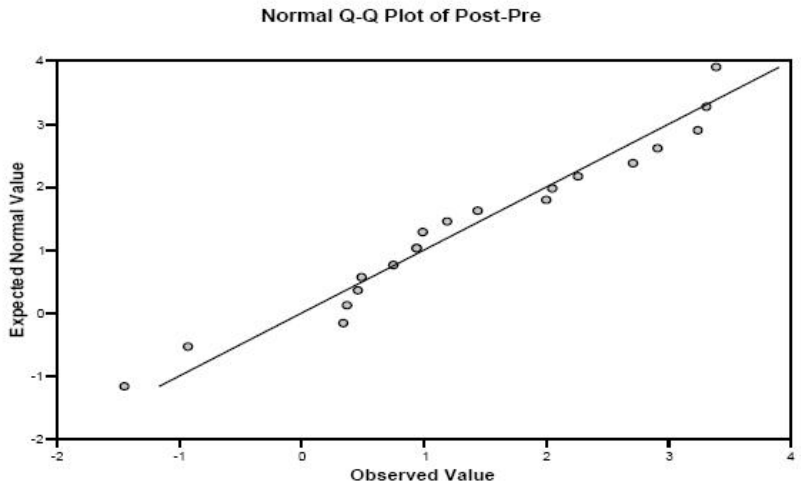


Figure 4.8: A quantile-normal plot.

4.3.4 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.

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 wrong. As we will see in the ANOVA and regression chapters, the most important situation where we use a QN plot is not for EDA, but for examining something called "residuals" (see section 9.4). For

basic interpretation of the QN plot you just need to be able to distinguish 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 area 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 actual data values are plotted against "Expected Normal Values", and some kind of 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.

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 4.10 shows all of the points basically on the reference line, but 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. And all of the observed values are at whole numbers. So either the data are rounded or we are looking at a discrete quantitative (counting) variable. Either way, the data appear

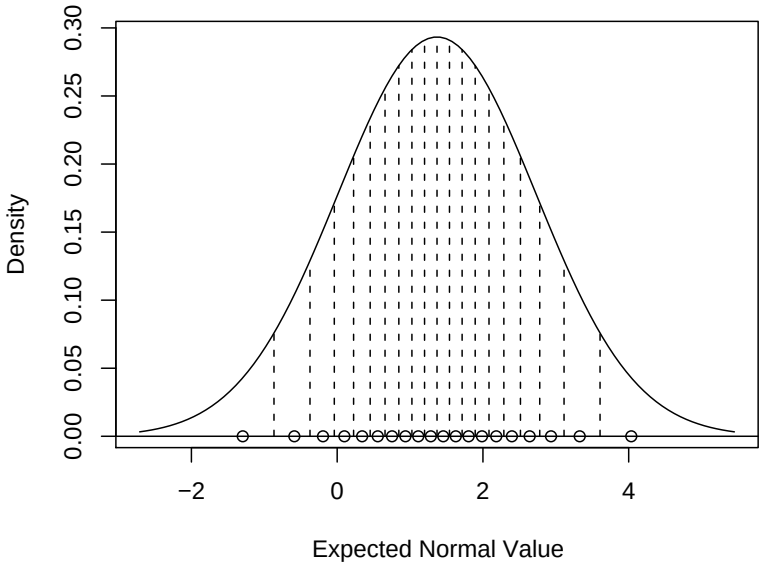


Figure 4.9: A way to think about QN plots.

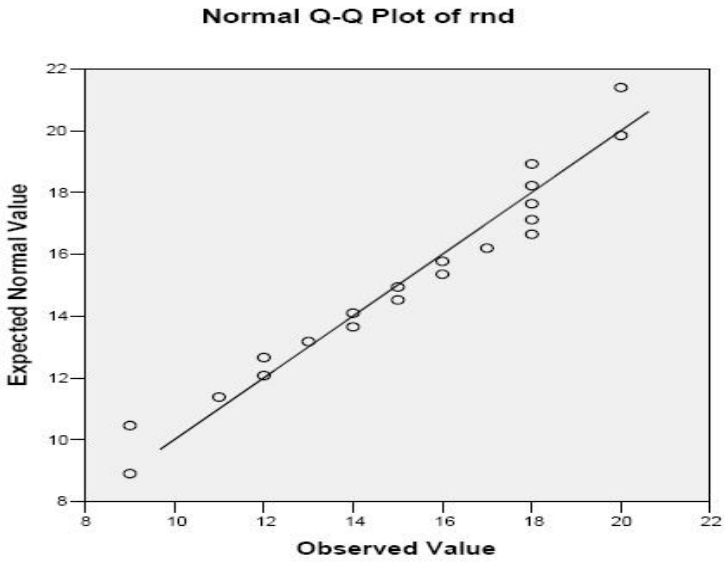


Figure 4.10: Quantile-normal plot with ties.

to be nearly 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 "x" (Observed Value) direction the points appear to have moved relative to 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.

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.

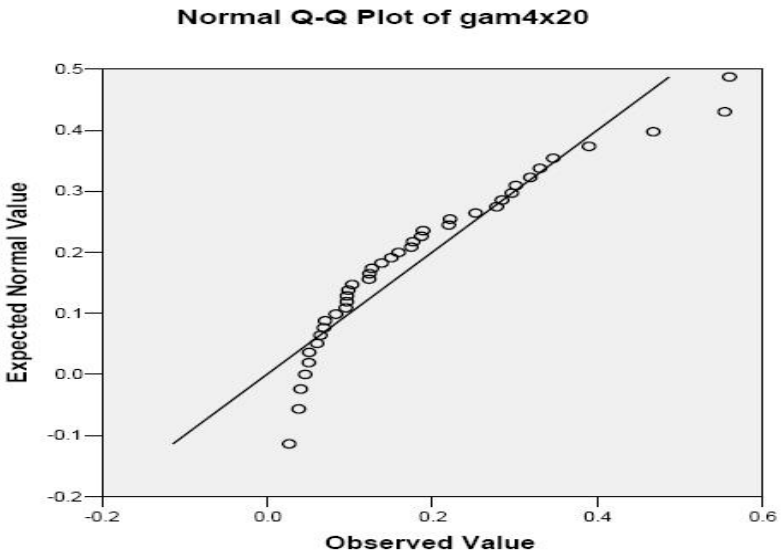


Figure 4.11: Quantile-normal plot showing right skew.

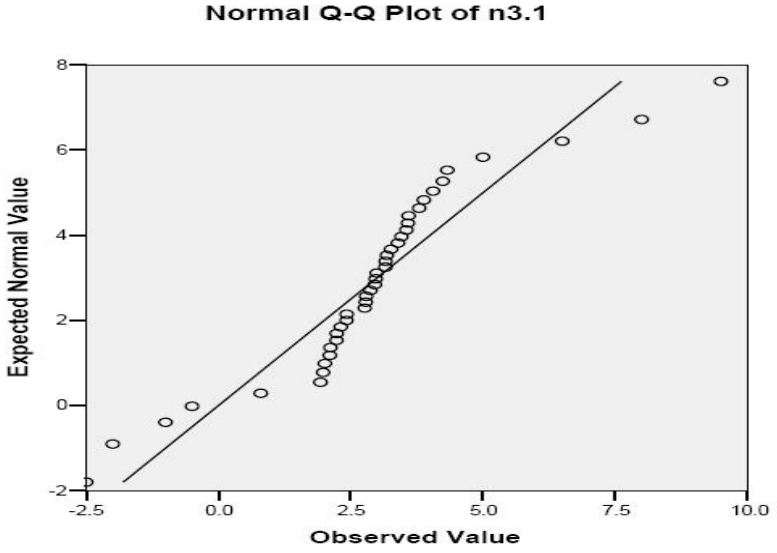


Figure 4.12: Quantile-normal plot showing fat tails.

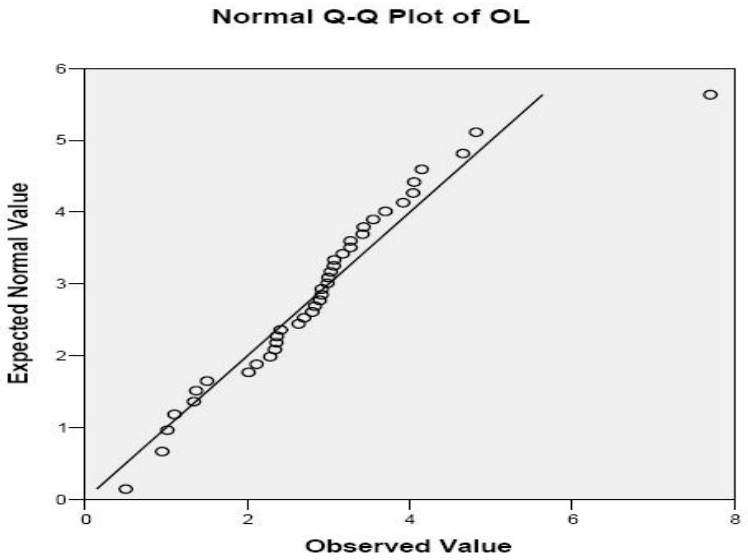


Figure 4.13: Quantile-normal plot showing a high outlier.

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.

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.

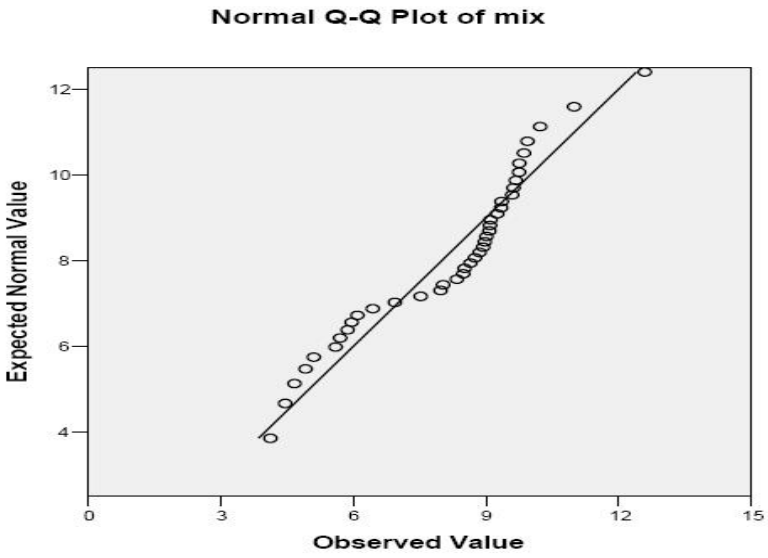


Figure 4.14: Quantile-normal plot showing bimodality.

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 and age as categorical variables.

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 $2/7 \times 100 = 28.6\%$.

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.

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

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

For example, we could make separate age by gender tables for each education level.

Cross-tabulation is the basic 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. Comparing medians is a robust informal version of one-way ANOVA. Comparing measures of spread is a good informal test of the assumption of equal variances needed for valid analysis of variance.

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 3.5. 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.

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). Note that the units on sample covariance are the products of the units of the two variables.

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. The correlation has the nice property that it is always 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 general formula for sample covariance is

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

It is worth noting that $\text{Cov}(X,X) = \text{Var}(X)$.

If you want to see a "manual example" of calculation of 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).

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.

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.

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

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 1.0. 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 figure 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.

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

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.

4.5 Multivariate graphical EDA

There are few useful techniques for graphical EDA of two categorical random variables. The only one used commonly is a grouped barplot with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

	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

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 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 of these are **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. The spreads (IQRs) are similar for the three groups. And 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.

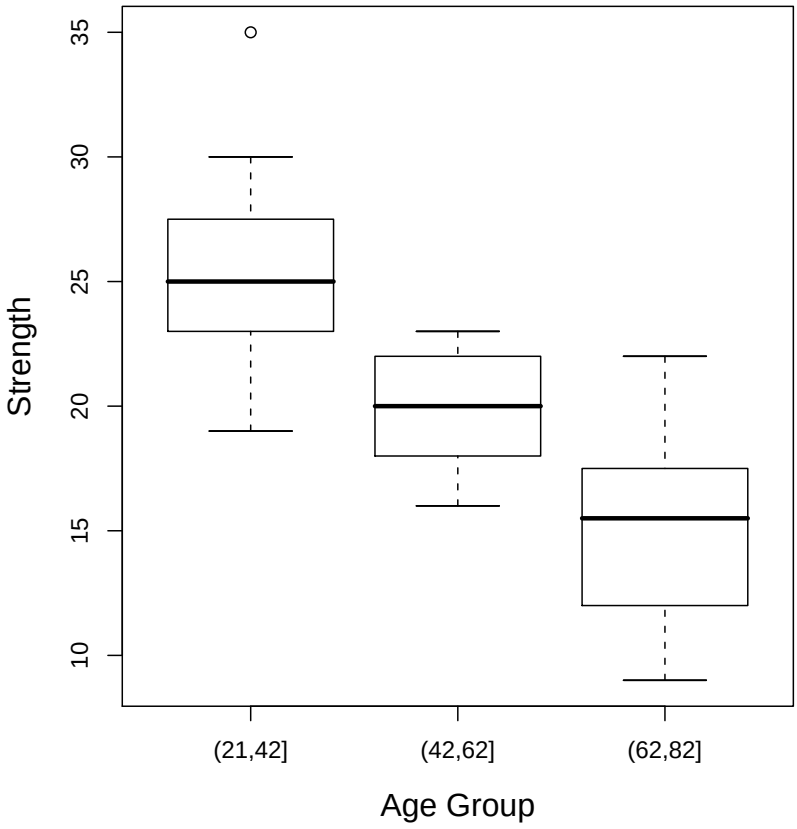


Figure 4.15: Side-by-side Boxplot of EDA3.dat.

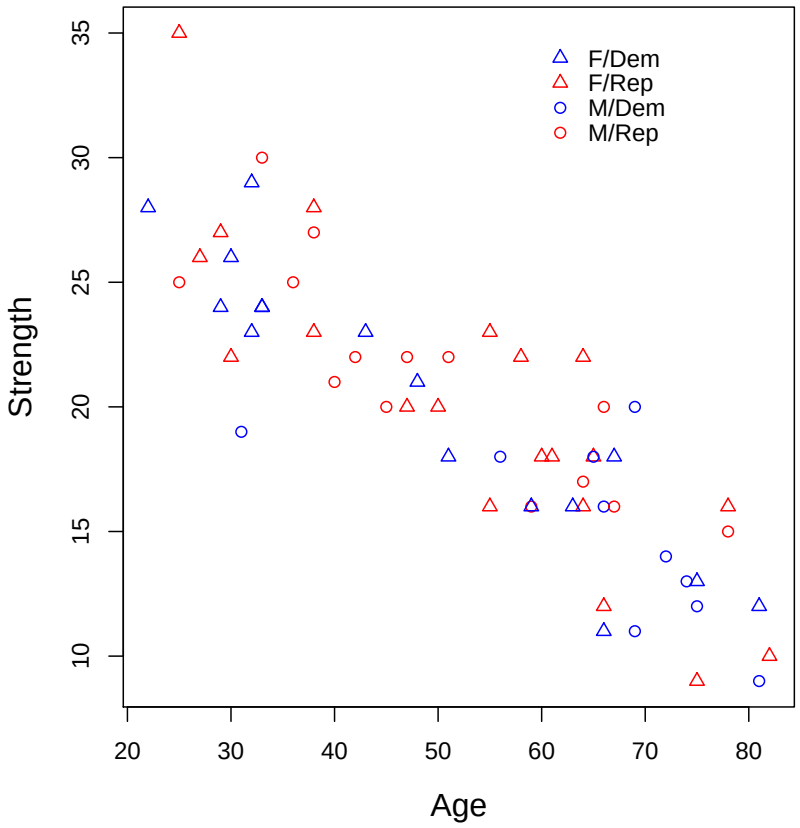


Figure 4.16: scatterplot with two additional variables.

An example is shown in figure 4.16. Age vs. strength is shown, and different colors and symbols are used to code political party and gender.

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!

4.6 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 in a some general situation, 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. One way to think about degrees of freedom for a statistic is that they represent the number of independent pieces of information that go into the calculation of the statistic,

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 $\mathrm{df}_1 + \cdots +$

df_k degrees of freedom, which suggests how many independent pieces of information are available for the calculation.

Chapter 5

Learning SPSS: Data and EDA

An introduction to SPSS with emphasis on EDA.

SPSS (now called PASW Statistics, but still referred to in this document as SPSS) is a perfectly adequate tool for entering data, creating new variables, performing EDA, and performing formal statistical analyses. I don't have any special endorsement for SPSS, other than the fact that its market dominance in the social sciences means that there is a good chance that it will be available to you wherever you work or study in the future. As of 2009, the current version is 17.0, and class datasets stored in native SPSS format in version 17.0 may not be usable with older versions of SPSS. (Some screen shots shown here are not updated from previous versions, but all changed procedures have been updated.)

For very large datasets, SAS tends to be the best program. For creating custom graphs and analyses R, which is free, or the commercial version, S-Plus, are best, but R is not menu-driven. The one program I strongly advise against is Excel (or any other spreadsheet). These programs have quite limited statistical facilities, discourage structured storage of data, and have no facility for documenting your work. This latter deficit is critical! For any serious analysis you must have a complete record of how you created new variables and produced all of your graphical and statistical output.

It is *very* common that you will find some error in your data at some point. So it is highly likely that you will need to repeat all of your analyses, and that is painful without exact records, but easy or automatic with most good software. Also, because it takes a long time from analysis to publishing, you will need these

records to remind yourself of exactly which steps you performed.

As hinted above, the basic steps you will take with most experimental data are:

1. Enter the data into SPSS, or load it into SPSS after entering it into another program.
2. Create new variables from old variables, if needed.
3. Perform exploratory data analyses.
4. Perform confirmatory analyses (formal statistical procedures).
5. Perform model checking and model comparisons.
6. Go back to step 4 (or even 2), if step 5 indicates any problems.
7. Create additional graphs to communicate results.

Most people will find this chapter easier to read when SPSS is running in front of them. There is a lot of detail on getting started and basic data management. This is followed by a brief compilation of instructions for EDA. The details of performing other statistical analyses are at the end of the appropriate chapters throughout this book.

Even if you are someone who is good at jumping in to a computer program without reading the instructions, I urge you to read this chapter because otherwise you are likely to miss some of the important guiding principles of SPSS.

Additional SPSS resources may be found at
http://www.stat.cmu.edu/∼hseltman/SPSSTips.html.

5.1 Overview of SPSS

SPSS is a multipurpose data storage, graphical, and statistical system. At (almost) all times there are two window types available, the Data Editor window(s) which each hold a single data "spreadsheet", and the Viewer window from which analyses are carried out and results are viewed.

The Data Editor has two views, selected by tabs at the bottom of the window. The Data View is a spreadsheet which holds the data in a rectangular format with

cases as rows and variables as columns. Data can be directly entered or imported from another program using menu commands. (Cut-and-paste is possible, but not advised.) Errors in data entry can also be directly corrected here.

You can also use menu commands in the Data View to create new variables, such as the log of an existing variable or the ratio of two variables.

The Variable View tab of the Data Editor is used to customize the information about each variable and the way it is displayed, such as the number of decimal places for numeric variables, and the labels for categorical variables coded as numbers.

The Viewer window shows the results of EDA, including graph production, formal statistical analyses, and model checking. Most data analyses can be carried out using the menu system (starting in either window), but some uncommon analyses and some options for common analyses are only accessible through "Syntax" (native SPSS commands). Often a special option is accessed by using the Paste button found in most main dialog boxes, and then typing in a small addition. (More details on these variations is given under the specific analyses that require them.)

All throughout SPSS, each time you carry out a task through a menu, the underlying non-menu syntax of that command is stored by SPSS, and can be examined, modified and saved for documentation or reuse. In many situations, there is a "Paste" button which takes you to a "syntax window" where you can see the underlying commands that would have been executed had you pressed OK.

SPSS also has a complete help system and an advanced scripting system.

You can save data, syntax, and graphical and statistical output separately, in various formats whenever you wish. (Generally anything created in an earlier program version is readable by later versions, but not vice versa.) Data is normally saved in a special SPSS format which few other programs can understand, but universal formats like "comma separated values" are also available for data interchange. You will be warned if you try to quit without saving changes to your data, or if if you forget to save the output from data analyses.

As usual with large, complex programs, the huge number of menu items available can be overwhelming. For most users, you will only need to learn the basics of interaction with the system and a small subset of the menu options.

Some commonly used menu items can be quickly accessed from a toolbar, and learning these will make you more efficient in your use of SPSS.

SPSS has a few quirks; most notably there are several places where you can make selections, and then are supposed to click Change before clicking OK. If you forget to click Change your changes are often silently forgotten. Another quirk that is well worth remembering is this: *SPSS uses the term Factor to refer to any categorical explanatory variable.* One good "quirk" is the Dialog Recall toolbar button. It is a quick way to re-access previous data analysis dialogs instead of going through the menu system again.

5.2 Starting SPSS

Note: SPSS runs on Windows and Mac operating systems, but the focus of these notes is Windows. If you are unfamiliar with Windows, the link
Top 10 tips for Mac users getting started with Windows may help.

Assuming that SPSS is already installed on your computer system, just choose it from the Windows Start menu or double click its icon to begin. The first screen you will see is shown in figure 5.1 and gives several choices including a tutorial and three choices that we will mainly use: "Type in data", "Open an existing data source", and "Open another type of file". "Type in data" is useful for analyzing small data sets not available in electronic form. "Open an existing data source" is used for opening data files created in SPSS. "Open another type of file" is used for importing data stored in files not created by SPSS. After making your choice, click OK. Clicking Cancel instead of OK is the same as choosing "Type in data".

Use Exit from the File menu whenever you are ready to quit SPSS.

5.3 Typing in data

To enter your data directly into SPSS, choose "Type in data" from the opening screen, or, if you are not at the opening screen, choose New then Data from the File menu.

The window titled "Untitled SPSS Data Editor" is the Data Editor window which is used to enter, view and modify data. You can also start statistical analyses from this window. Note the tabs at the bottom of the window labeled "Data View" and "Variable View". In Data View (5.2), you can view, enter, and edit data for all of your cases, while in Variable View (5.3), you can view, enter, and edit

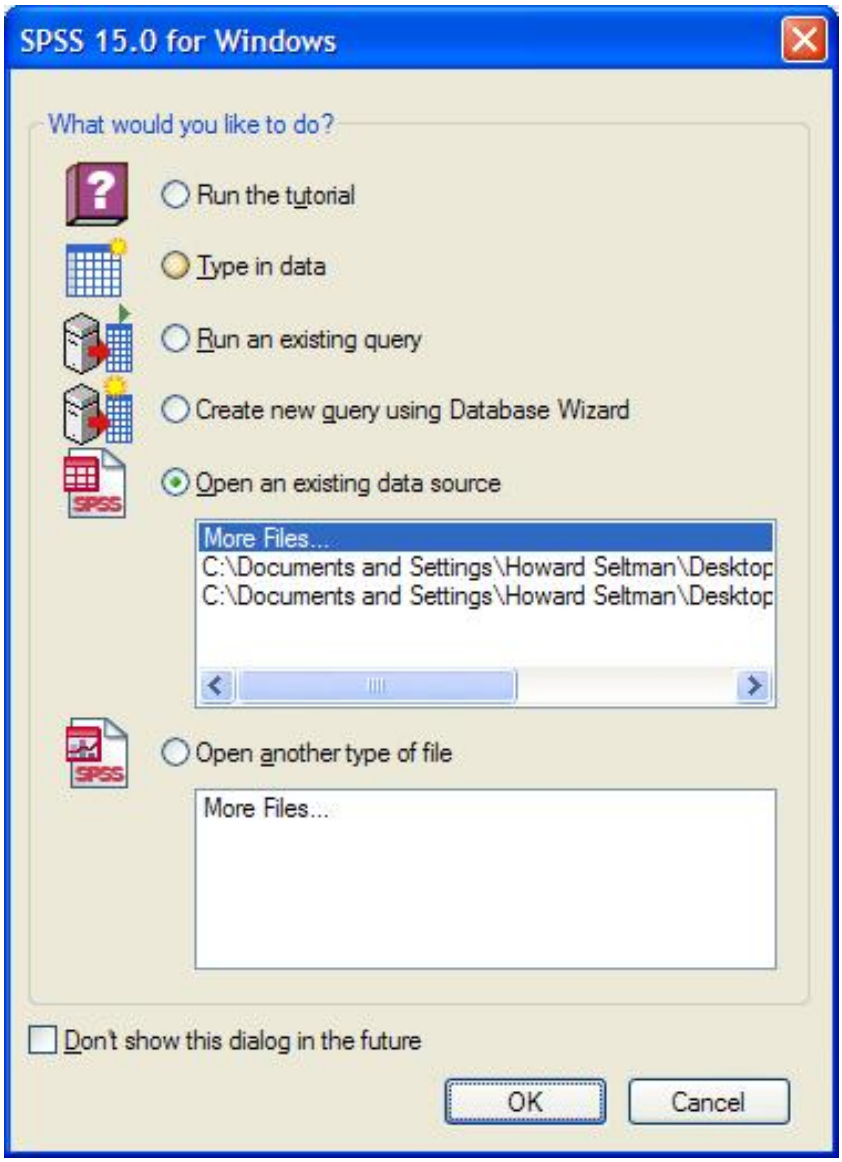


Figure 5.1: SPSS intro screen.

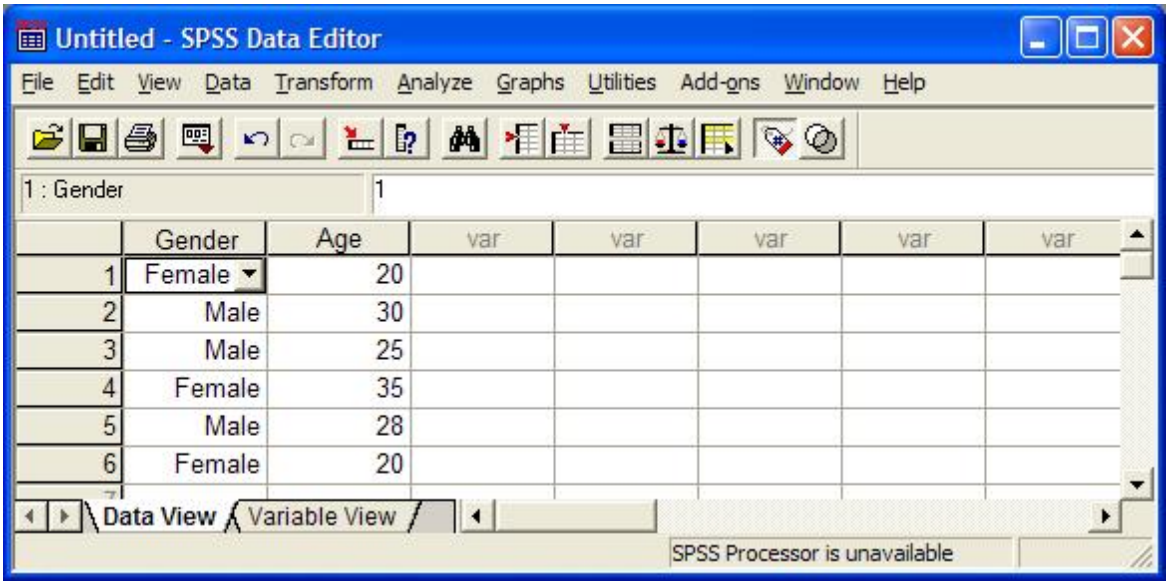


Figure 5.2: Data Editor window: Data View.

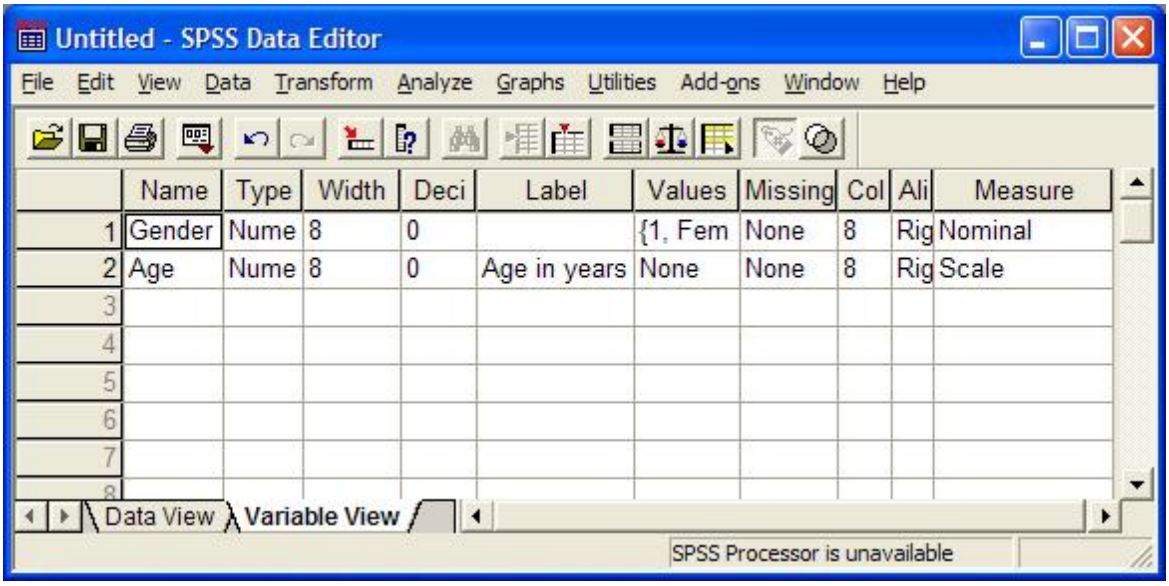


Figure 5.3: Data Editor window: Variable View.

information about the variables themselves (see below). Also note the menu and toolbar at the top of the window. You will use these to carry out various tasks related to data entry and analysis. There are many more choices than needed by a typical user, so don't get overwhelmed! You can hover the mouse pointer over any toolbar button to get a pop-up message naming its function. This chapter will mention useful toolbar items as we go along. (Note: Toolbar items that are inappropriate for the current context are grayed out.)

Before manually entering data, you should tell SPSS about the individual variables, which means that you should think about variable types and coding before entering the data. Remember that the two data types are categorical and quantitative and their respective subtypes are nominal and ordinal, and discrete and continuous. These data type correspond to the Measure column in the Variable View tab. SPSS does not distinguish between discrete and continuous, so it calls all quantitative variables "scale". Ordinal and nominal variables are the other options for Measure. In many parts of SPSS, you will see a visual reminder of the Measure of your variables in the form of icons. A small diagonal yellow rule indicates a "scale" variable (with a superimposed calendar or clock if the data hold dates or times). A small three level bar graph with increasing bar heights indicates an "ordinal" variable. Three colored balls with one on top and two below indicates nominal data (with a superimposed "a" if the data are stored as "strings" instead of numbers).

Somewhat confusingly SPSS Variable View has a column called Type which is the "computer science" type rather than the "statistics" data type. The choices are basically numeric, date and string with various numeric formats. This course does not cover time series, so we won't use the "date" Type. Probably the only use for the "string" Type is for alphanumeric subject identifiers (which should be assigned "nominal" Measure). All standard variables should be entered as numbers (quantitative variables) or numeric codes (categorical variables). Then, for categorical variables, we always want to use the Values column to assign meaningful labels to the numeric codes.

Note that, in general, to set or change something in the Data Editor, you first click in the cell whose row and column correspond to what you want to change, then type the new information. To modify, rather than fully re-type an entry, press the key labeled "F2".

When entering a variable name, note that periods and underscores are allowed in variable names, but spaces and most other punctuation marks are not. The

variable name must start with a letter, may contain digits, and must not end with a period. Variable names can be at most 64 characters long, are *not* case sensitive, and must be unique. The case that you enter is preserved, so it may be useful to mix case, e.g., hotDogsPerHour to improve readability.

In either View of the Data Editor, you can neaten your work by dragging the vertical bar between columns to adjust column widths.

After entering the variable name, change whichever other column(s) need to be changed in the Variable View. For many variables this includes entering a Label, which is a human-readable alternate name for each variable. It may be up to 255 characters long with no restrictions on what you type. The labels replace the variable names on much of the output, but the names are still used for specifying variables for analyses.

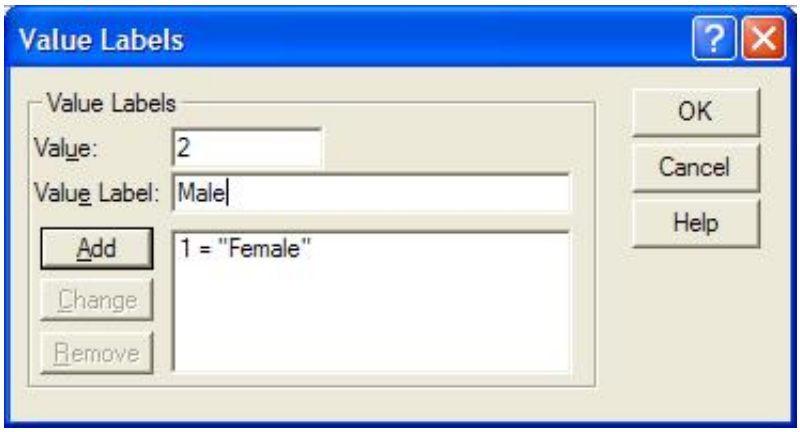


Figure 5.4: Values dialog box.

For categorical variables, you will almost always enter the data as numeric codes (Type "numeric"), and then enter Labels for each code. The Value Labels dialog box (5.4) is typical of many dialog boxes in SPSS. To enter Values for a variable, click in the box at the intersection of the variable's row and the Value column in the Variable View. Then click on the "..." icon that appears. This will open the "Value Labels" dialog box, into which you enter the words or phrases that label each level of your categorical variable. Value labels can contain anything you like up to 255 characters long. Enter a level code number in the Value box, press Tab, then enter the text for that level in the Value Label box. Finally you *must* click the Add button for your entry to be registered. Repeat the process as many times as needed to code all of the levels of the variable. When you are finished, verify

that all of the information in the large unlabeled box is correct, then click OK to complete the process. At any time while in the Value Label box (initially or in the future), you can add more labels; delete old labels by clicking on the variable in the large box, then clicking the Delete button; or change level values or labels by selecting the variable in the large box, making the change, then clicking the Change button. Version 16 has a spell check button, too.

If your data has missing values, you should use the Missing column of the Variable View to let SPSS know the missing value code(s) for each variable.

The only other commonly used column in Variable View is the Measure column mentioned above. SPSS uses the information in the column sporadically. Sometimes, but certainly not always, you will not be able carry out the analysis you want if you enter the Measure incorrectly (or forget to set it). In addition, setting the Measure assures that you appropriately think about the type of variable you are entering, so it is a really, really good idea to always set it.

Once you have entered all of the variable information in Variable View, you will switch to Data View to enter the actual data. At it's simplest, you can just click on a cell and type the information, possibly using the "F2" key to edit previously entered information. But there are several ways to make data entry easier and more accurate. The tab key moves you through your data case by case, covering all of the variables of one case before moving on to the next. Leave a cell blank (or delete its contents) to indicate "missing data"; missing data are displayed with a dot in the spreadsheet (but don't type a dot).

The Value Labels setting, accessed either through its toolbar button (which looks like a gift tag) or through the View menu, controls both whether columns with Value Labels display the value or the label, and the behavior of those columns during data entry. If Value Labels is turned on, a "..." button appears when you enter a cell in the Data View spreadsheet that has Value Labels. You can click the button to select labels for entry from a drop down box. Also, when Value Labels is on, you can enter data either as the code or by typing out the label. (In any case the code is what is stored.)

You should use Save (or Save as) from the File menu to save your data after every data entry session and after any edits to your data. Note that in the "Save Data As" dialog box (5.5) you should be careful that the "Save in:" box is set to save your data in the location you want (so that you can find it later). Enter a file name and click "Save" to save your data for future use. Under "Save as type:" the default is "SPSS" with a ".sav" extension. This is a special format that

can be read quickly by SPSS, but not at all by most other programs. For data exchange between programs, several other export formats are allowed, with Excel with "Comma separated values" being the most useful.

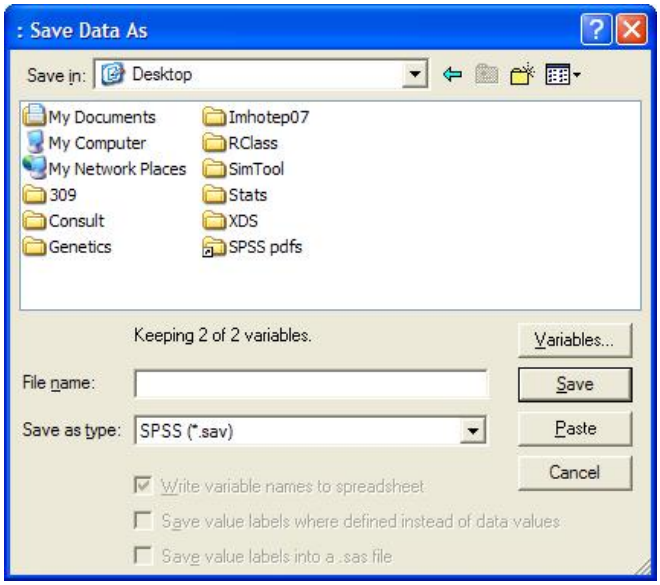


Figure 5.5: Save Data As dialog box.

5.4 Loading data

To load in data when you first start SPSS, your can select a file in one of the two lower boxes of the "Intro Screen". At any other time you can load data from the File menu by selecting Open, then Data. This opens the "Open File" dialog box (5.6).

It's a good idea to save any changes to any open data set before opening a new file. In the Open File dialog box, you need to find the file by making appropriate choices for "Look in:" and "Files of type:". If your file has a ".txt" extension and you are looking for files of type ".dat", you will not be able to find your file. As a last resort, try looking for files of type "all files(*.*)". Click Open after finding your file.

If your file is a native SPSS ".sav" file, it will open immediately. If it is of another type, you will have to go through some import dialogs. For example, if

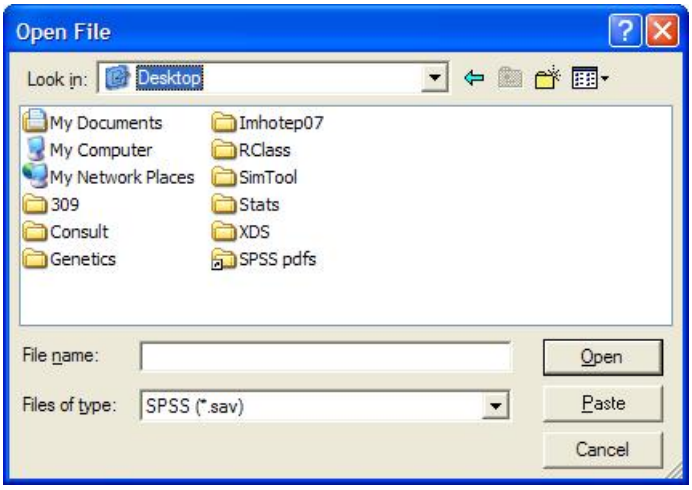


Figure 5.6: Open File dialog box.

you open an Excel file (.xls), you will see the "Opening Excel Data Source" dialog box (5.7). Here you use a check box to tell SPSS whether or not your data has variable names in the first row. If your Excel workbook has multiple worksheets you must select the one you want to work with. Then, optionally enter a Range of rows and columns if your data does not occupy the entire range of used cells in the worksheet. Finish by clicking OK.

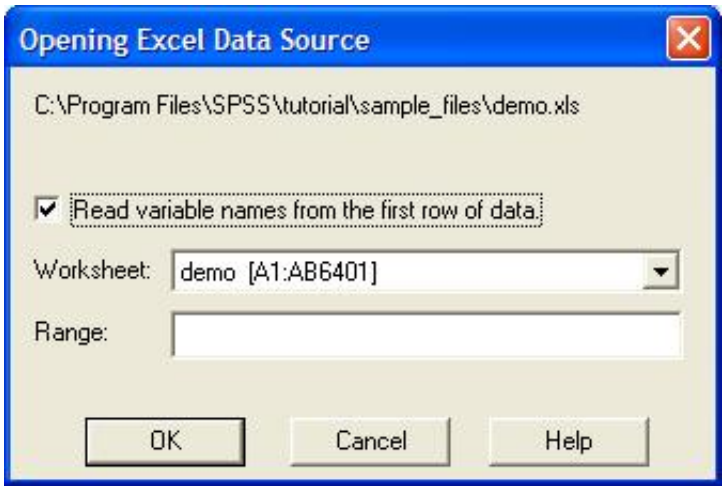


Figure 5.7: Open Excel Data Source dialog box.

The other useful type of data import is one of the simple forms of human-readable text such as space or tab delimited text (usually .dat or .txt) or comma separated values (.csv). If you open one of these files, the "Text Import Wizard" dialog box will open. The rest of this section describes the use of the text import wizard.

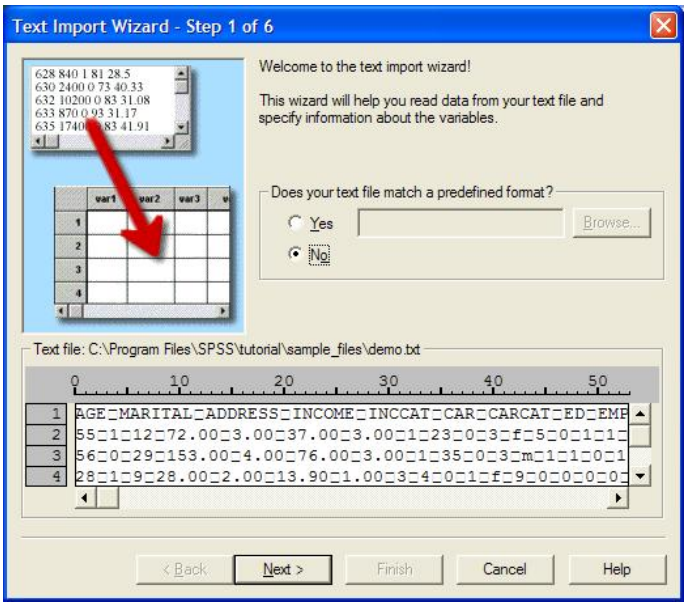


Figure 5.8: Text Import Wizard - Step 1 of 6.

In "Step 1 of 6" (5.8) you will see a question about predefined formats which we will skip (as being beyond the scope of this course), and below you will see some form of the first four lines of your file (and you can scroll down or across to see the whole file). (If you see strange characters, such as open squares, your file probably has non-printable characters such as tab character in it.) Click Next to continue.

In "Step 2 of 6" (5.9) you will see two *very important* questions that you must answer accurately. The first is whether your file is arranged so that each data column always starts in exactly the same column for every line of data (called "Fixed width") or whether there are so-called delimiters between the variable columns (also called "fields"). Delimiters are usually either commas, tab characters or one or more spaces, but other delimiters occasionally are seen. The second question is "Are variable names include at the top of the file?" Answer "no" if the first

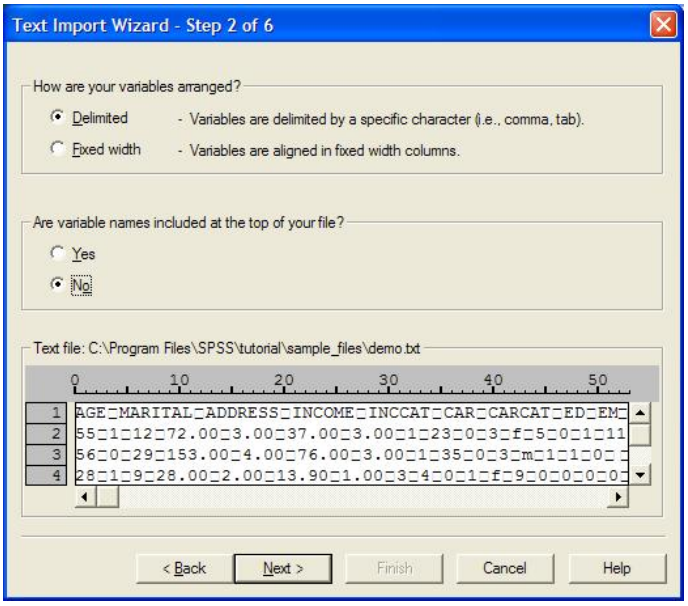


Figure 5.9: Text Import Wizard - Step 2 of 6.

line of the file is data, and "yes" if the first line is made of column headers. After answering these questions, click Next to continue.

In "Step 3 of 6" (5.10) your first task is to input the line number of the file that has the first real data (as opposed to header lines or blank lines). Usually this is line 2 if there is a header line and line 1 otherwise. Next is "How are your cases represented?" Usually the default situation of "Each line represents a case" is true. Under "How many cases do you want to import?" you will usually use the default of "All of the cases", but occasionally, for very large data sets, you may want to play around with only a subset of the data at first.

In "Step 4 of 6" (5.11) you must answer the questions in such a way as to make the "Data preview" correctly represent your data. Often the defaults are OK, but not always. Your main task is to set the delimiters between the data fields. Usually you will make a single choice among "Tab", "Space", "Comma", and "Semicolon". You may also need to specify what sets off text, e.g. there may be quoted multi-word phrases in a space separated file.

If your file has fixed width format instead of delimiters, "Step 4 of 6" has an alternate format (5.12). Here you set the divisions between data columns.

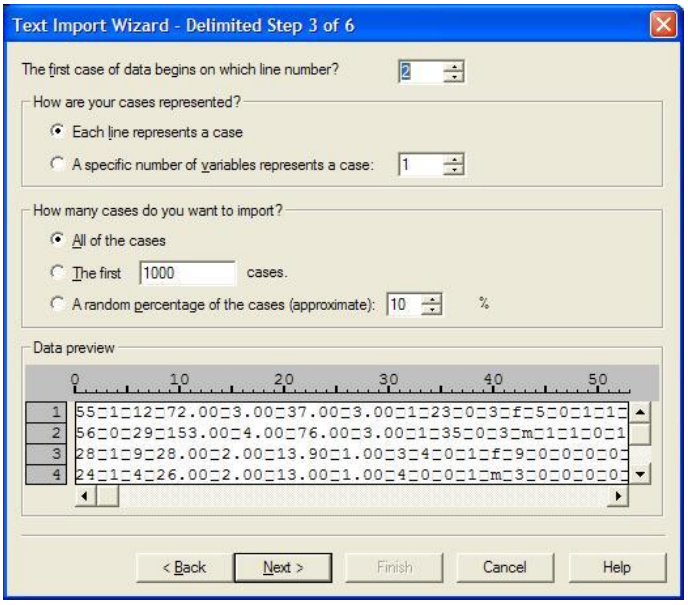


Figure 5.10: Text Import Wizard - Step 3 of 6.

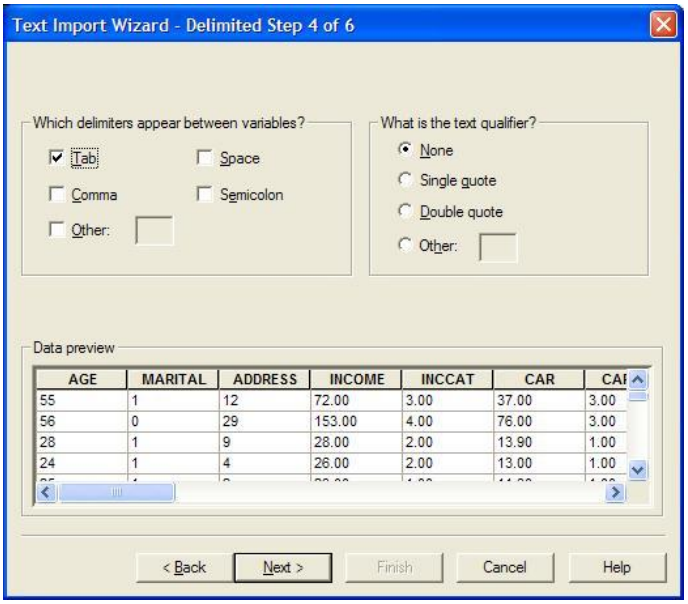


Figure 5.11: Text Import Wizard - Step 4 of 6.

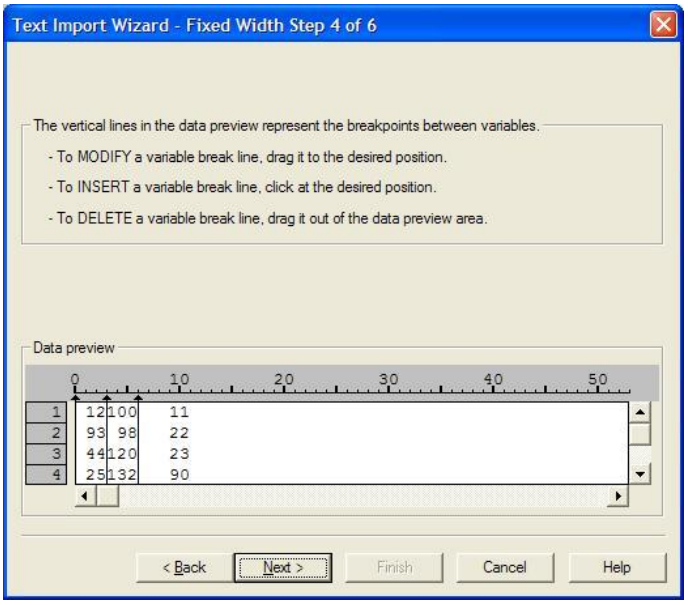


Figure 5.12: Text Import Wizard - Alternate Step 4 of 6.

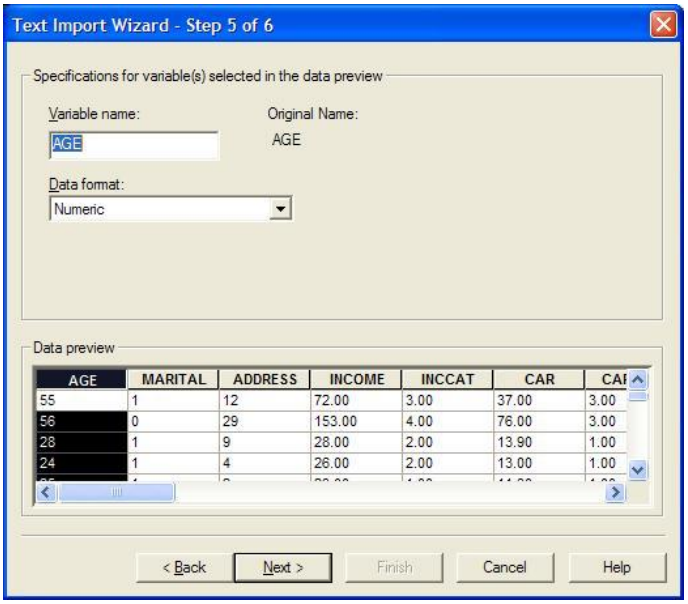


Figure 5.13: Text Import Wizard - Step 5 of 6.

In "Step 5 of 6" (5.13) you will have the chance to change the names of variables and/or the data format (numeric, data or string). Ordinarily you don't need to do anything at this step.

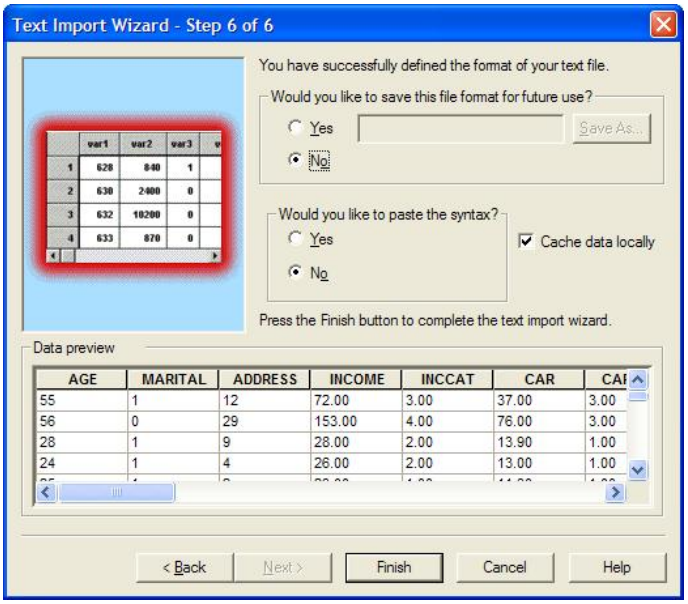


Figure 5.14: Text Import Wizard - Step 6 of 6.

In "Step 6 of 6" (5.14) you will have the chance to save all of your previous choices to simplify future loading of a similar file. We won't use this feature in this course, so you can just click the Finish button.

The most common error in loading data is forgetting to specify the presence of column headers in step 2. In that case the column header (variable names) appear as data rather than variable names.

5.5 Creating new variables

Creating new variables (data transformation) is commonly needed, and can be somewhat complicated. Depending on what you are trying to do, one of several menu options starts the process.

For creating of a simple data **transformation**, which is the result of applying a mathematical formula to one or more existing variables, use the ComputeVariable

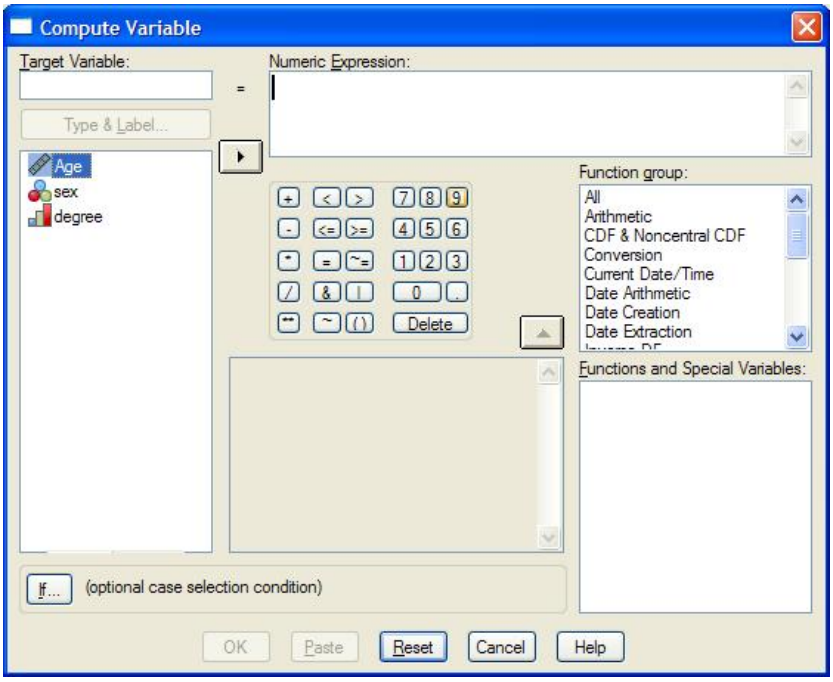


Figure 5.15: Compute Variable dialog box.

item on the Transform menu of the Data Editor. This open the Compute Variable dialog box (5.15). First enter a new variable name in the Target Variable box (remembering the naming rules discussed above). Usually you will want to click the "Type & Label" box to open another dialog box which allows you to enter a longer, more readable Label for the variable. (You will almost never want to change the type to "String".) Click Continue after entering the Label. Next you will enter the "Numeric Expression" in the Compute Variable dialog box. Two typical expressions are "log(weight)" which creates the new variable by the taking the log of the existing variable "weight", and "weight/height**2" which computes the body mass index from height and weight by dividing weight by the square (second power) of the height. (Don't enter the quotation marks.)

To create a transformation, use whatever method you can to get the required Numeric Expression into the box. You can either type a variable name or double click it in the variable list to the left, or single click it and click the right arrow. Spaces don't matter (except within variable names), and standard order of operations are used, but can be overridden with parentheses as needed. Numbers, operators (including * for times), and function names can be entered by clicking the mouse, but direct typing is usually faster. In addition to the help system, the list of functions may be helpful for finding the spelling of a function, e.g., sqrt for square root.

Comparison operators (such as =, <. and >) can be used with the understanding that the result of any comparison is either "true", coded as 1, or "false", coded as 0. E.g., if one variable called "vfee" has numbers indicating the size of a fee, and a variable called "surcharge" is 0 for no surcharge and 1 for a $25 surcharge, then we could create a new variable called "total" with the expression "vfee+25*(surcharge=1)". In that case either 25 (25*1) or 0 (25*0) is added to "vfee" depending of the value of "surcharge".

Advanced: To transform only some cases and leave others as "missing data" use the "If" button to specify an expression that is true only for the cases that need to be transformed.

Some other functions worth knowing about are ln, exp, missing, mean, min, max, rnd, and sum. The function ln() takes the natural log, as opposed to log(), which is common log. The function exp() is the anti-log of the natural log, as opposed to 10**x which is the common log's anti-log. The function missing() returns 1 if the variable has missing data for the case in question or 0 otherwise. The functions min(), max(), mean() and sum(), used with several variables separated with

commas inside the parentheses, computes a new value for each case from several existing variables for that case. The function rnd() rounds to a whole number.

5.5.1 Recoding

In addition to simple transformations, we often need to create a new variable that is a **recoding** of an old variable. This is usually used either to "collapse" categories in a categorical variable or to create a categorical version of a quantitative variable by "binning". Although it is possible to over-write the existing variable with the new one, I strongly suggest that you always preserve the old variable (for record keeping and in case you make an error in the encoding), and therefore you should use the 'into Different Variables" item under "Recode" on the "Transform" menu, which opens the "Recode into Different Variables" dialog box (5.16).

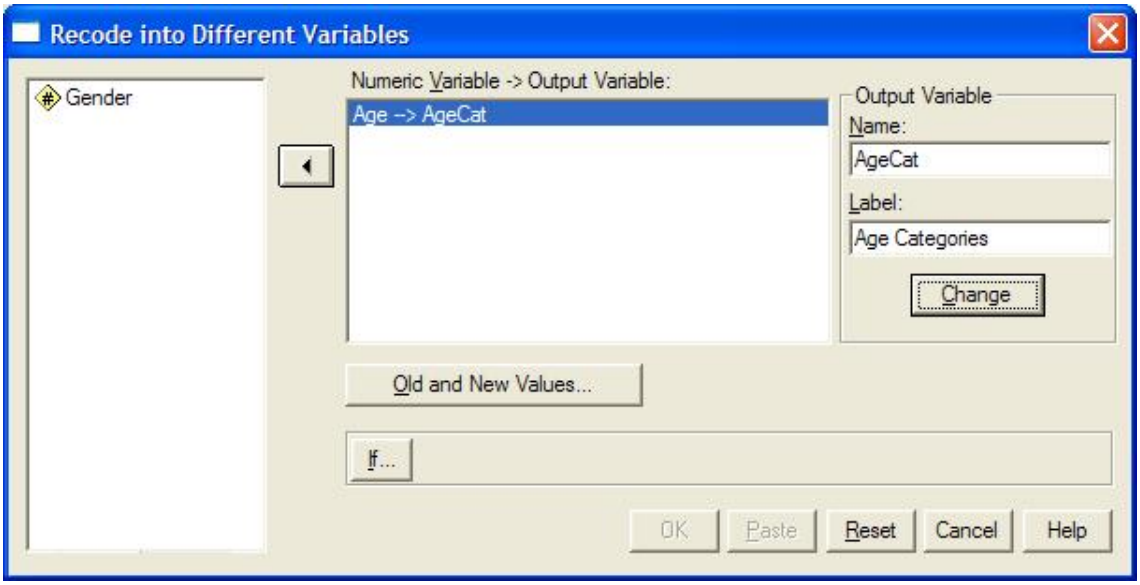


Figure 5.16: Recode into Different Variables Dialog Box.

First enter the existing variable name into the "Numeric Variable -> Output Variable" box. If you have several variables that need the same recoding scheme, enter each of them before proceeding. Then, for each existing variable, go to the "Output Variable" box and enter a variable Name and Label for the new recoded variable, and confirm the entry with the Change button.

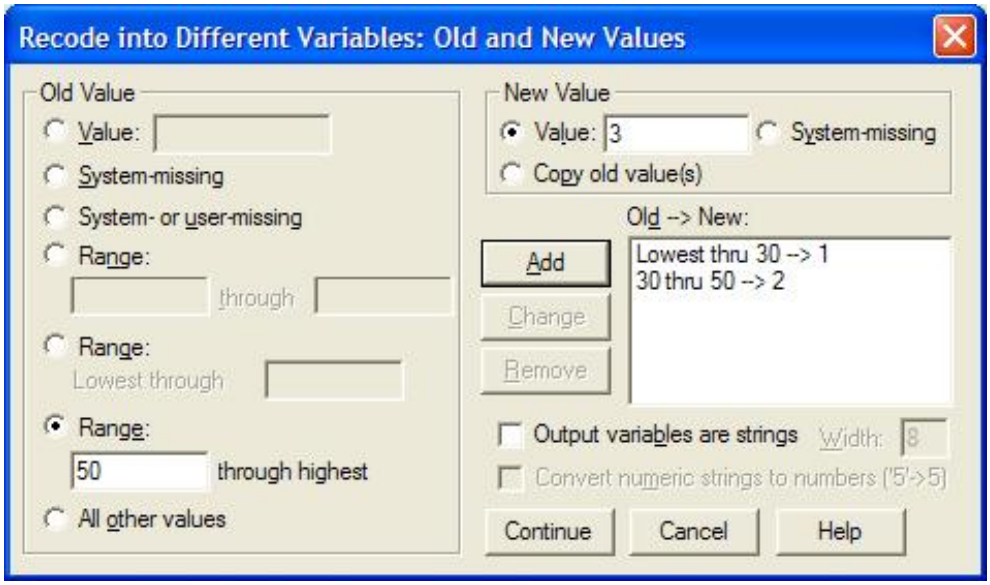


Figure 5.17: Recode into Different Variables: Old and New Values Dialog Box.

Then click the "Old and New Values" button to open the "Recode into Different Variables: Old and New Values" dialog box (5.17). Your goal is to specify as many "rules" as needed to create a new value for every possible old value so that the "Old–>New" box is complete and correct. For each one or several old values that will be recoded to a particular new value, enter the value or range of values on the left side of the dialog box, then enter the new value that represents the recoding of the old value(s) in the "New Value" box. Click Add to register each particular recoding, and repeat until finished. Often the "All other value" choice is the last choice for the "Old value". You can also use the Change and Remove buttons as needed to get a final correct "Old–>New" box. Click Continue to finalize the coding scheme and return to the "Recode into Different Values" box. Then click OK to create the new variable(s). If you want to go directly on to recode another variable, I strongly suggest that you click the Reset button first to avoid confusion.

5.5.2 Automatic recoding

Automatic recode is used in SPSS when you have strings (words) as the actual data levels and you want to convert to numbers (usually with Value labels). Among other reasons, this conversion saves computer memory space.

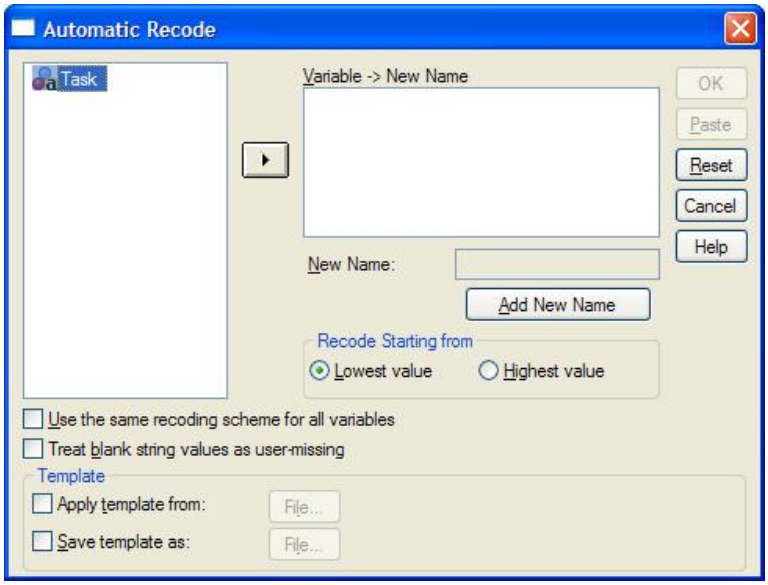


Figure 5.18: Automatic Recode Dialog Box.

From the Transform menu of the Data Editor menu, select "Automatic Recode" to get the "Automatic Recode" dialog box as shown in figure 5.18. Choose a variable, enter a new variable name in the "New Name" box and click "Add New Name". Repeat if desired for more variables. If there are missing data values in the variable and they are coded as blanks, click "Treat blank string values as user-missing". Click OK to create the new variable. You will get some output in the Output window showing the recoding scheme. A new variable will appear in the Data Window. If you click the Value Labels toolbar button, you will see that the new variable is really numeric with automatically created value labels.

5.5.3 Visual binning

SPSS has a option called "Visual Binning", accessed through the Visual Binning item on the Transformation menu, which allows you to interactively choose how to create a categorical variable from a quantitative (scale) variable. In the "Visual Binning" dialog box you select one or more quantitative (or ordinal) variables to work with, then click Continue. The next dialog box is also called "Visual Binning" and is shown in figure 5.19. Here you select a variable from the one(s) you previously chose, then enter a new name for the categorical variable you want

to create in the "Binned Variable" box (and optionally change its Label). A histogram of the variable appears. Now you have several choices for creating the "bins" that define the categories. One choice is to enter numbers in the Value column (and optionally Labels). For the example in the figure, I entered 33 as Value for line 1 and 50 for line 2, and the computer entered HIGH for line 3. I also entered the labels. When I click "OK" the quantitative variable "Age" will be recoded into a three level categorical variable based on my cutpoints.

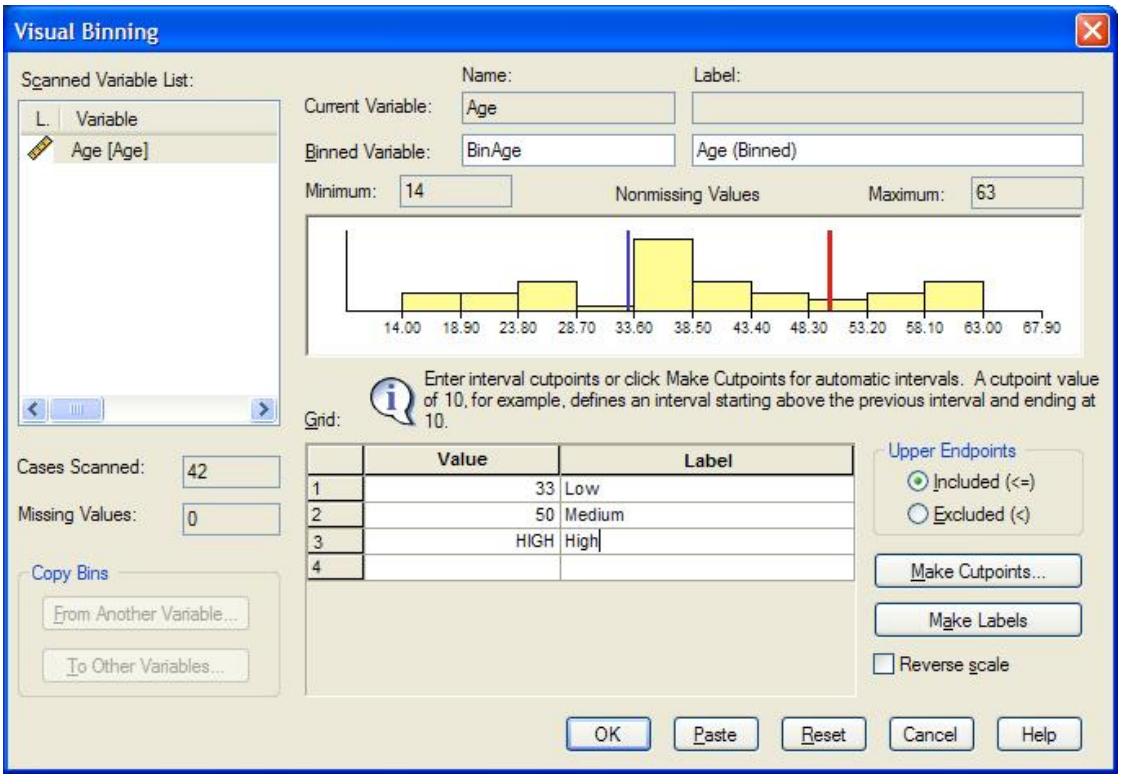


Figure 5.19: Visual Binning dialog box: Entered interval cutpoints.

The alternative to directly entering interval cutpoints is to click "Make Cutpoints" to open the "Make Cutpoints" dialog box shown in figure 5.20. Here your choices are to define some equal width intervals, equal percent intervals, or make cutpoints at fixed standard deviation intervals around the mean. After defining your cutpoints, click Apply to return to the histogram, which is now annotated based on your definition. (If you don't like the cutpoints edit them manually or return to Make Cutpoints.) You should manually enter meaningful labels for

the bins you have chosen or click "Make Labels" to get some computer generated labels. Then click OK to make your new variable.

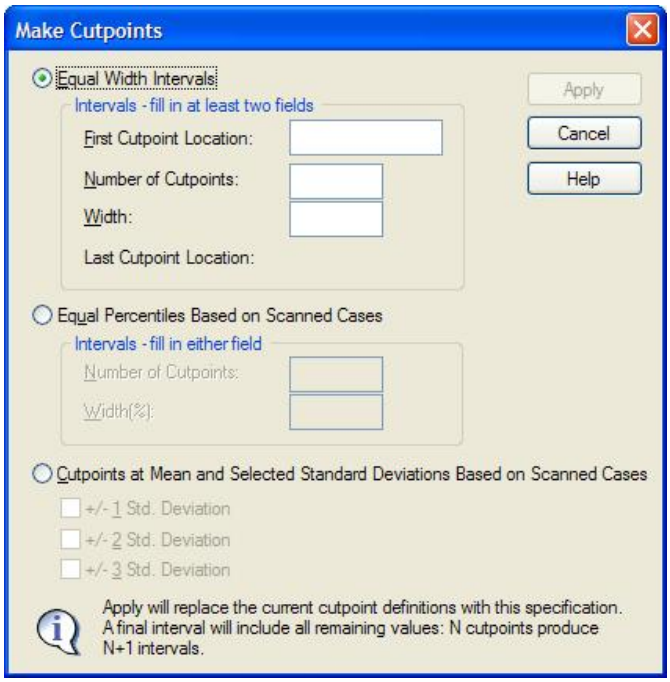


Figure 5.20: Visual Binning dialog box: Make cutpoints.

5.6 Non-graphical EDA

To **tabulate a single categorical variable**, i.e., get the numbers and percent of cases at each level of the variable, use the Frequencies subitem under the Descriptive Statistics item of the Analyze menu. This is also useful for quantitative variables with not too many unique values. When you choose your variable(s) and click OK, the Frequency table will appear in the Output Window. The default output (e.g., figure 5.21) shows each unique value, and its frequency and percent. The "Valid Percent" column calculates percents for only the non-missing data, while the "Percent" column only adds to 100% when you include the percent missing. Cumulative Percent can be useful for ordinal data. It adds all of the Valid Percent numbers for any row plus all rows above in the table, i.e. for any data value it shows what percent of cases are less than or equal to that value.

degree

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	14	33.3	33.3	33.3
	2	15	35.7	35.7	69.0
	3	13	31.0	31.0	100.0
	Total	42	100.0	100.0	

Figure 5.21: SPSS frequency table.

To **cross-tabulate two or more categorical variables** use the Crosstabs subitem under the Descriptive Statistics item of the Analyze menu. This is also useful for quantitative variables with not too many unique values. Enter one variable under "Rows" and one under "Columns". If you have a third variable, enter it under "Layer". (You can use the "Next" Layer button if you have more than three variables to cross-tabulate, but that may be too hard to interpret. Click OK to get the cross-tabulation of the variables. The default is to show only the counts for each combination of levels of the variables. If you want percents, click the "Cells" button before clicking OK; this gives the "Crosstabs: Cell Display" dialog box from which you can select percentages that add to 100% across each Row, down each "Column" or in "Total" across the whole cross-tabulation. Try to think about which of these makes the most sense for understanding your dataset it each particular case. Example output is shown in figure 5.22.

sex * degree Crosstabulation

			degree			Total
			1	2	3	
sex	Male	Count	7	5	8	20
		% within degree	50.0%	33.3%	61.5%	47.6%
	Female	Count	7	10	5	22
		% within degree	50.0%	66.7%	38.5%	52.4%
Total		Count	14	15	13	42
		% within degree	100.0%	100.0%	100.0%	100.0%

Figure 5.22: SPSS cross-tabulation.

For various **univariate quantitative variable sample statistics** use the

Descriptives subitem under the Descriptive Statistics item of the Analyze menu. Ordinarily you should use "Descriptives" for quantitative and possibly ordinal variables. (It words, but rarely makes sense for nominal variables.) The default is to calculate the sample mean, sample "Std. deviation", sample minimum and sample maximum. You can click on "Options" to access other sample statistics such as sum, variance, range, kurtosis, skewness, and standard error of the mean. Example output is show in figure 5.23. The sample size (and indication of any missing values) is always given. Note that for skewness and kurtosis standard errors are given. The rough rule-of-thumb for interpreting the skewness and kurtosis statistics is to see if the absolute value of the statistic is smaller than twice the standard error (labeled Std. Error) of the corresponding statistic. If so, there is no good evidence of skewness (asymmetry) or kurtosis. If the absolute value is large (compared to twice the standard error), then a positive number indicates right skew or positive kurtosis respectively, and a negative number indicates left skew or negative kurtosis.

Rule of thumb: Interpret skewness and kurtosis sample statistics by comparing the absolute value of the statistic to twice the standard error of the statistic. Small statistic value are consistent with the zero skew and kurtosis of a Gaussian distribution.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	41	14	63	38.59	13.240
degree	42	1	3	1.98	.811
Valid N (listwise)	41				

Figure 5.23: SPSS descriptive statistics.

To get the **correlation of two quantitative variables** in SPSS, from the Analyze menu item choose Correlate/Bivariate. Enter two (or more) quantitative variables into the Variables box, then click OK. The output will show correlations and a p-value for the test of zero correlation for each pair of variables. You may also want to turn on calculation of means and standard deviations using the Options button.Example output is show in figure 5.24. The "Pearson Correlation" statis-

tic is the one that best estimates the population correlation of two quantitative variables discussed in section 3.5.

Correlations

		Age	Response time (ms)
Age	Pearson Correlation	1.000	-.252
	Sig. (2-tailed)		.283
	N	20.000	20
Response time (ms)	Pearson Correlation	-.252	1.000
	Sig. (2-tailed)	.283	
	N	20	20.000

Figure 5.24: SPSS correlation.

(To calculate the various types of correlation for categorical variables, run the crosstabs, but click on the "Statistics" button and check "Correlations".)

To calculate **median or quartiles for a quantitative variable** (or possibly an ordinal variable) use Analyze/Frequencies (which is normally used just for categorical data), click the Statistics button, and click median and/or quartiles. Normally you would also uncheck "Display frequency tables" in the main Frequencies dialog box to avoid voluminous, unenlightening output. Example output is show in figure 5.25.

Statistics

degree

N	Valid	42
	Missing	0
Median		2.00
Percentiles	25	1.00
	50	2.00
	75	3.00

Figure 5.25: SPSS median and quantiles.

5.7 Graphical EDA

5.7.1 Overview of SPSS Graphs

The Graphs menu item in SPSS version 16.0 has two sub-items: ChartBuilder and LegacyDialogs. As you might guess, the legacy dialogs item access older ways to create graphs. Here we will focus on the interactive Chart Builder approach. Note that graph, chart, and plot are interchangeable terms.

There is a great deal of flexibility in building graphs, so only the principles are given here.

When you select the Chart Builder menu item, it will bring up the Chart Builder dialog box. Note the three main areas: the variable box at top left, the chart preview area (also called the "canvas") at top right, and the (unnamed) lower area from which you can select a tab out of this group of tabs: Gallery, Basic Elements, Groups/PointID, and Titles/Footnotes.

A view of the (empty) Chart Builder is shown in 5.26.

To create a graph, go to the Gallery tab, select a graph type on the left, then choose a suitable template on the right, i.e. one that looks roughly like the graph you want to create. Note that the templates have names that appear as pop-up labels if you hover the mouse over them. Drag the appropriate template onto the canvas at top right. A preview of your graph (but not based on your actual data) will appear on the canvas.

The use of the Basic Elements tab is beyond the scope of this chapter.

The Groups/PointsID tab (5.27) serves both to add additional information from auxiliary variables (Groups) and to aid in labeling outliers or other interesting points (Point ID). After placing your template on the canvas, select the Groups/PointID tab. Sex check boxes are present in this tab. The top five choices refer to grouping, but only the ones appropriate for the chosen plot will be active. Check whichever ones might be appropriate. For each checked box, a "drop zone" will be added to the canvas, and adding an auxiliary variable into the drop zone (see below) will, in some way that is particular to the kind of graph you are creating, cause the graphing to be split into groups based on each level of the auxiliary variable. The "Point ID label" check box (where appropriate) adds a drop zone which hold the name of the variable that you want to use to label outliers or other special points. (If you don't set this, the row number in the spreadsheet is used

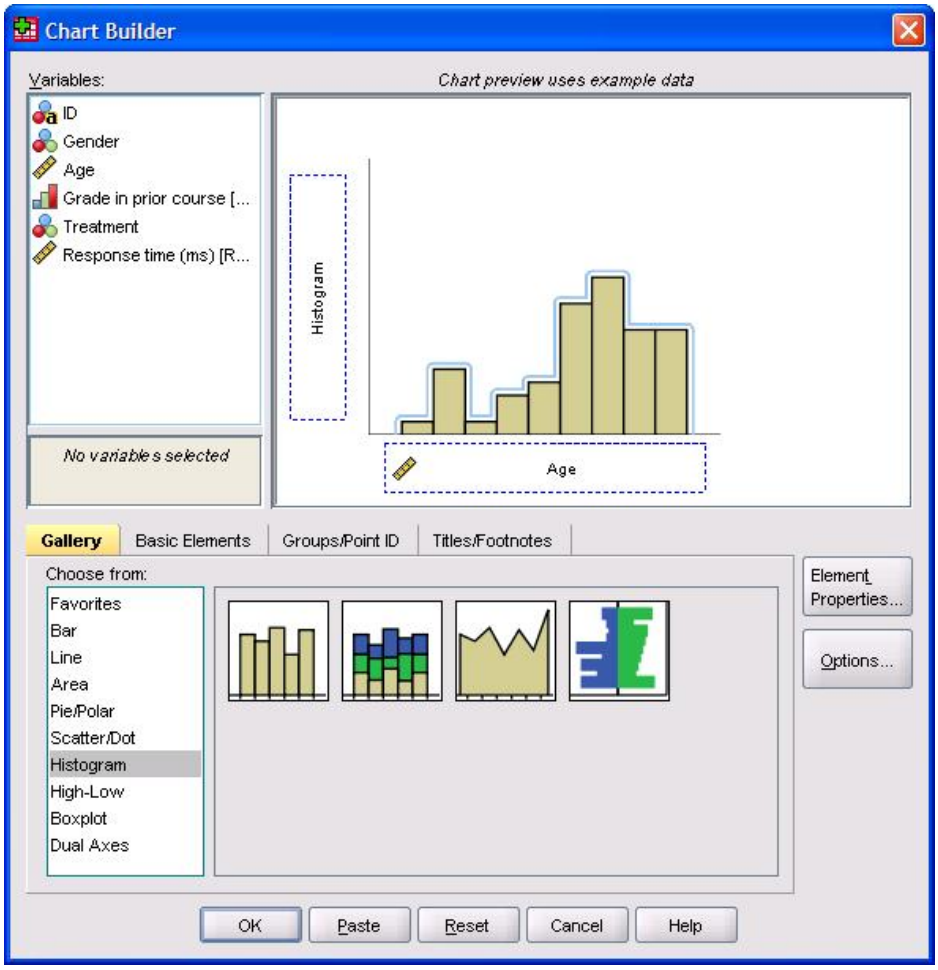


Figure 5.26: SPSS Empty Chart Builder.

for labeling.)

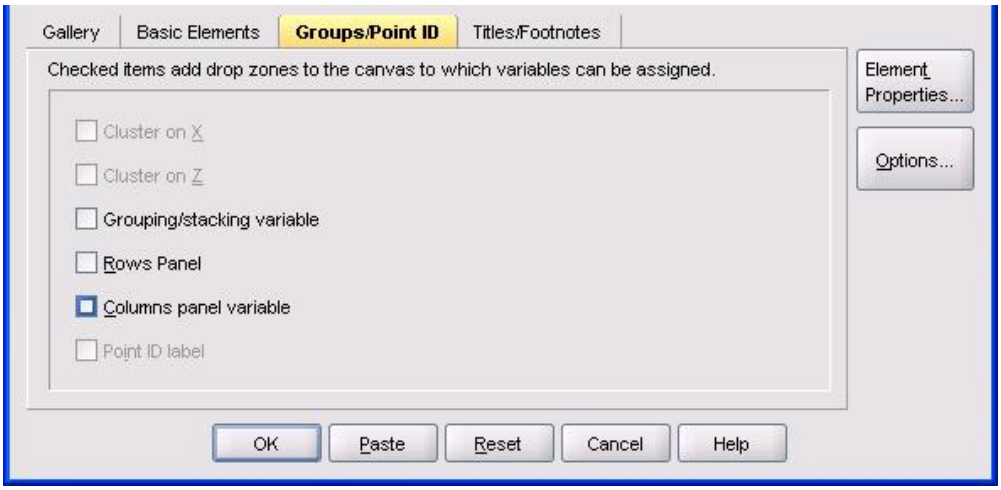


Figure 5.27: SPSS Groups/Point ID tab of Chart Builder.

The Titles/Footnotes tab (5.28) has check boxes for titles and footnotes. Check any that you need to appropriately annotate your graph. When you do so, the Element Properties dialog box (5.29) will open. (You can also open and close this box with the Element Properties button.) In the Element Properties box, select each title and/or footnote, then enter the desired annotation in the "Content" box.

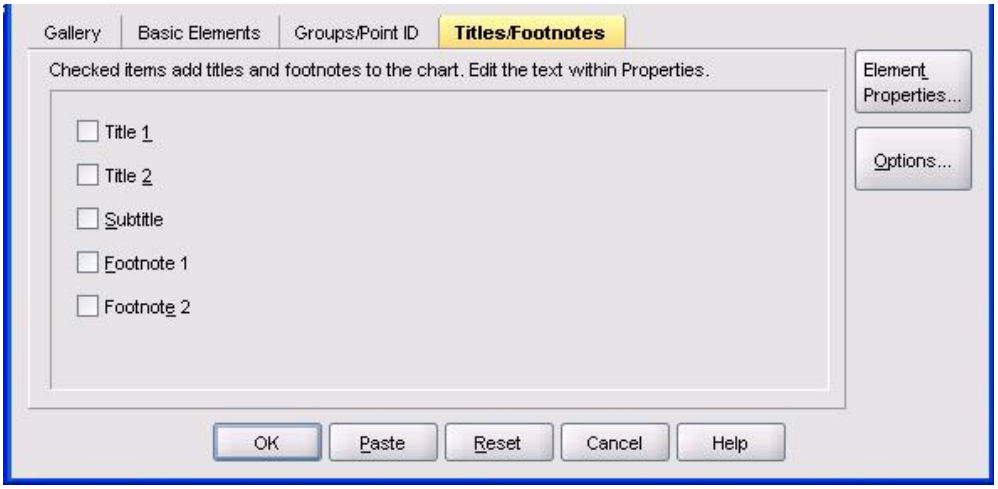


Figure 5.28: SPSS Titles/Footnote tab of Chart Builder.

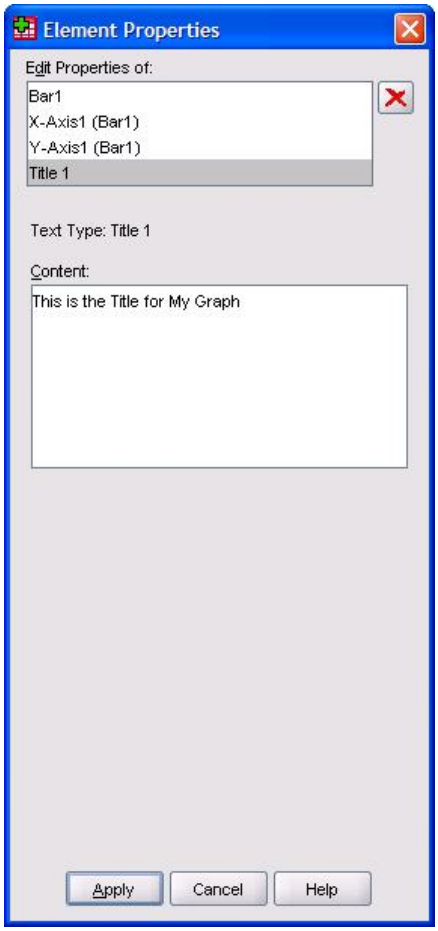


Figure 5.29: SPSS Element Properties dialog box.

Next you will add all of the variables that participate in the production of your graph to the appropriate places on the canvas. Note that when you click on any categorical variable in the Variables box, its categories are listed below the variable box. Drag appropriate variables into the pre-specified drop boxes (which vary with the type of graph chosen, and may include things like the x-axis and y-axis), as well as the drop boxes you created from the Groups/PointID tab.

You may want to revisit the Element Properties box and click through each element of the "Edit Properties of" box to see if there are any properties you might want to alter (e.g., the order of appearance of the levels of a categorical variable, or the scale for a quantitative variable). Be sure to click the Apply button after making any changes and before selecting another element or closing the Element Properties box.

Finally click OK in the Chart Builder dialog box to create your plot. It will appear at the end of your results in the SPSS Viewer window.

When you re-enter the Chart Builder, the old information will still be there, and that is useful to tweak the appearance of a plot. If you want to create a new plot unrelated to the previous plot, you will probably find it easiest to use the Reset button to remove all of the old information.

5.7.2 Histogram

The basic univariate **histogram** for quantitative or categorical data is generated by using the Simple Histogram template, which is the first one under Histogram in the Gallery. Simply drag your variable onto the x-axis to define your histogram ("Histogram" will appear on the y-axis.). For optionally grouping by a second variable, check "Grouping/stacking variable" in the Groups/PointID tab, then drag the second variable to the "Stack:set color" drop box. The latter is equivalent to choosing the "Stacked Histogram" in the gallery.

A view of the Chart Builder after setting up a histogram is shown in 5.30.

The "Population Pyramid" template (on the right side of the set of Histogram templates) is a nice way to display histograms of one variable at all levels of another (categorical) variable.

To **change the binning of a histogram**, double click on the histogram in the SPSS Viewer, which opens the Chart Editor (5.31), then double click on a histogram bar in the Chart Editor to open the Properties dialog box (5.32). Be

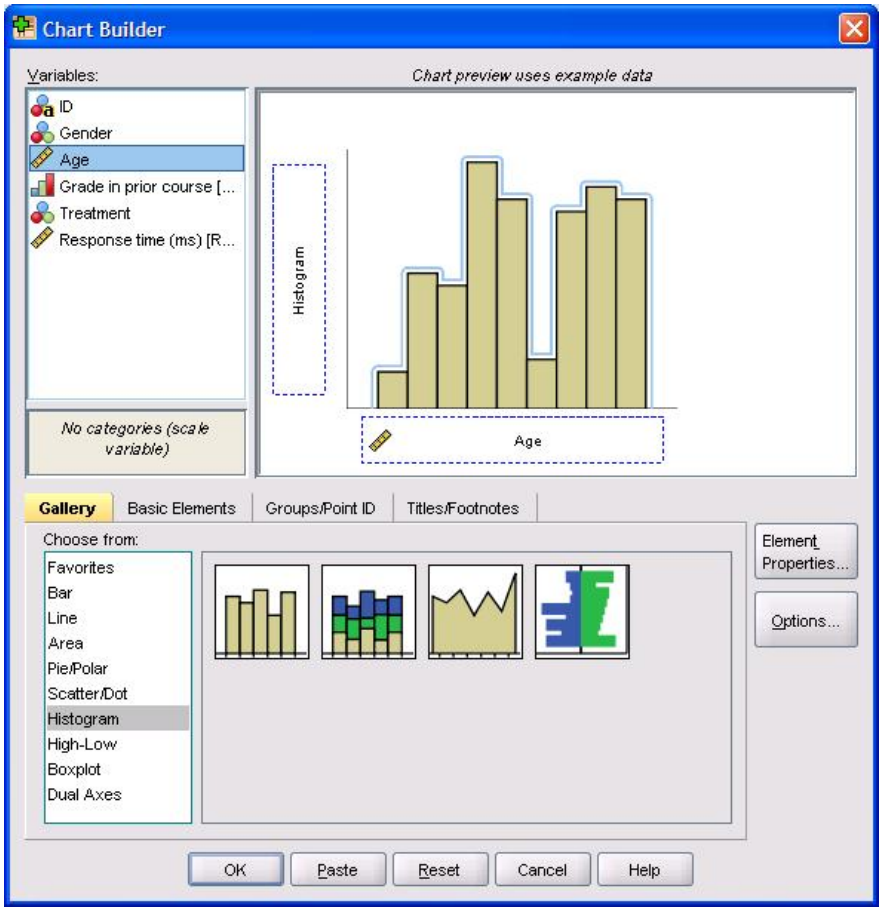


Figure 5.30: SPSS histogram setup.

sure that the Binning tab is active. Under "X Axis" change from Automatic to Custom, then enter either the desired number of intervals of the desired interval width. Click apply to see the result. When you achieve the best result, click Close in the Properties window, then close the Chart Editor window.

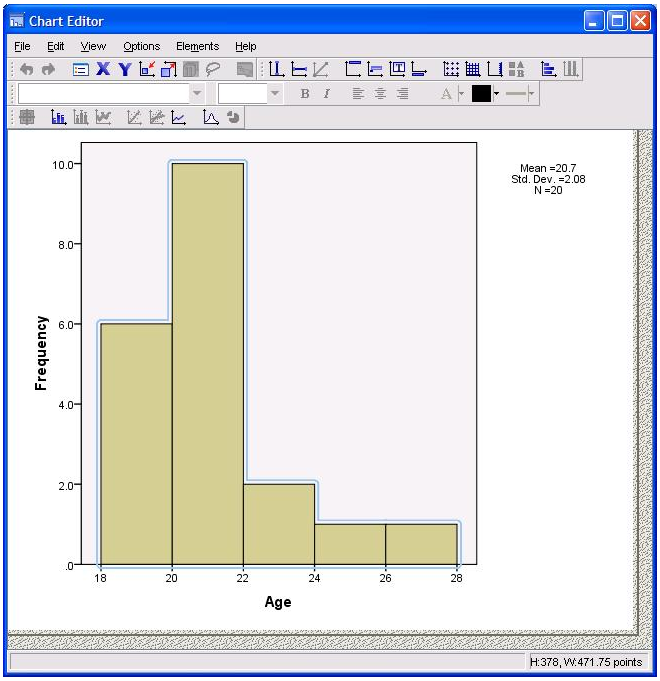


Figure 5.31: SPSS Chart Editor.

An example of a histogram produced in SPSS is shown in figure 5.33.

For histograms or any other graphs, it is a good idea to use the Titles/Footnote tab to set an appropriate title, subtitle and/or footnote.

5.7.3 Boxplot

A **boxplot** for quantitative random variables is generated in SPSS by using one of the three boxplot templates in the Gallery (called simple, clustered, and 1-D, from left to right). The 1-D boxplot shows the distribution of a single variable. The simple boxplot shows the distribution a one (quantitative) variable at each level of another (categorical) variable. The clustered boxplot shows the distribution a one (quantitative) variable at each level of two other categorical variables.

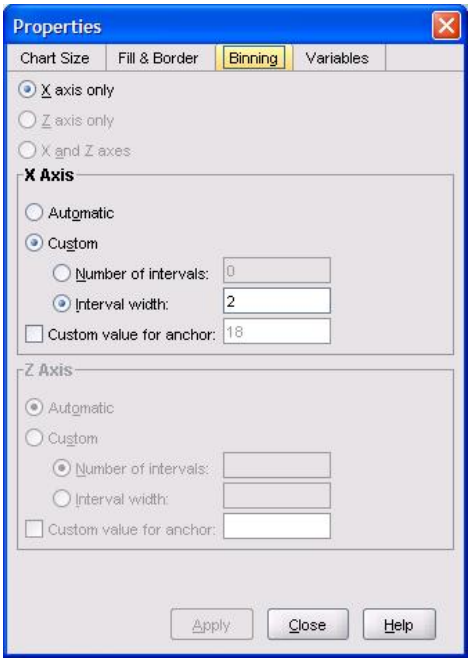


Figure 5.32: Binning in the SPSS Chart Editor.

An example of the Chart Builder setup for a simple boxplot with ID labels is shown in figure 5.34. The corresponding plot is in figure 5.35.

Other univariate graphs, such as pie charts and bar charts are also available through the Chart Builder Gallery.

5.7.4 Scatterplot

A **scatterplot** is the best EDA for examining the relationship between two quantitative variables, with a "point" on the plot for each subject. It is constructed using templates from the Scatter/Dot section of the Chart Builder Gallery. The most useful ones are the first two: Simple Scatter and Grouped Scatter. Grouped Scatter adds the ability to show additional information from some categorical variable, in the form of color or symbol shape.

Once you have placed the template on the canvas, drag the appropriate quantitative variables onto the x- and y-axes. *If one variable is outcome and the other explanatory, be sure to put the outcome on the vertical axis.* A simple example is

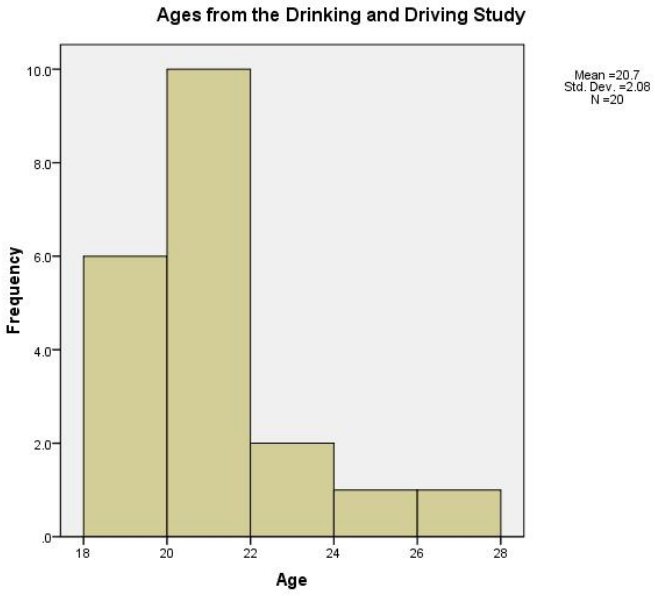


Figure 5.33: SPSS histogram.

shown in figure 5.36. The corresponding plot is in figure 5.37.

You can further modify a scatter plot by adding a best-fit straight line or a "non-parametric" smooth curve. This is done using the Chart Editor rather than the Chart Builder, so it is an addition to a scatterplot already created. Open the Chart Editor by double clicking on the scatterplot in the SPSS Viewer window. Choose "Add Fit Line at Total" by clicking on the toolbar button that looks like a scatterplot with a fit line through it, or by using the menu option Elements/FitLineAtTotal. This brings up the a Properties box with a "Fit Line" tab (5.38). The "Linear" Fit Method adds the best fit linear regression line. The "Loess" Fit Method adds a "smoother" line to your scatterplot. The smoother line is useful for detecting whether there is a non-linear relationship. (Technically it is a kernel smoother.) There is a degree of subjectivity in the overall smoothness vs. wiggliness of the smoother line, and you can adjust the "% of points to fit" to change this. Also note that if you have groups defined with separate point colors for each group, you can substitute "Add Fit Line at Subgroups" for "Add Fit Line at Total" to have separate lines for each subgroup.

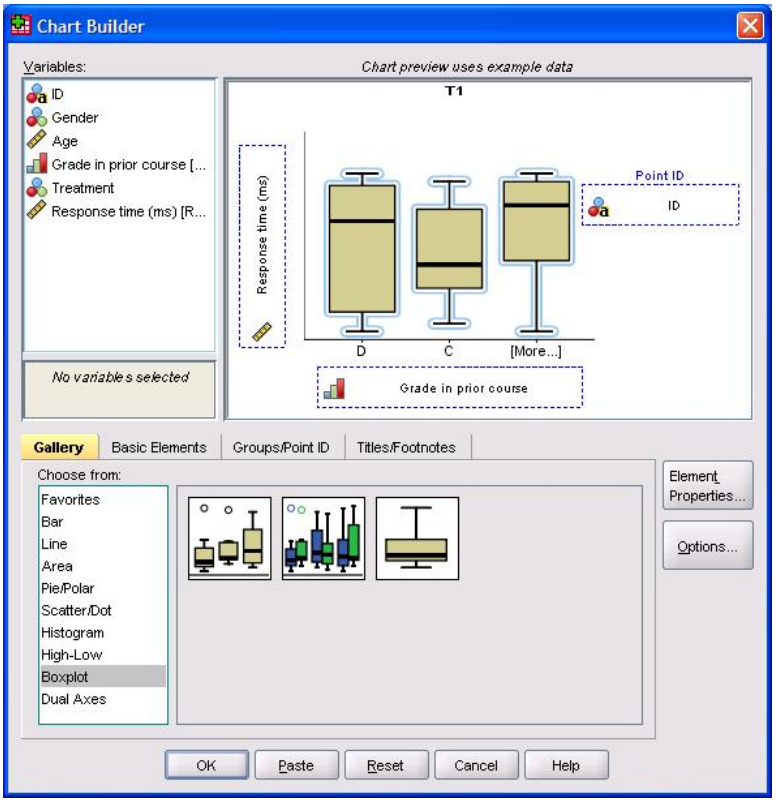


Figure 5.34: SPSS boxplot setup in Chart Builder.

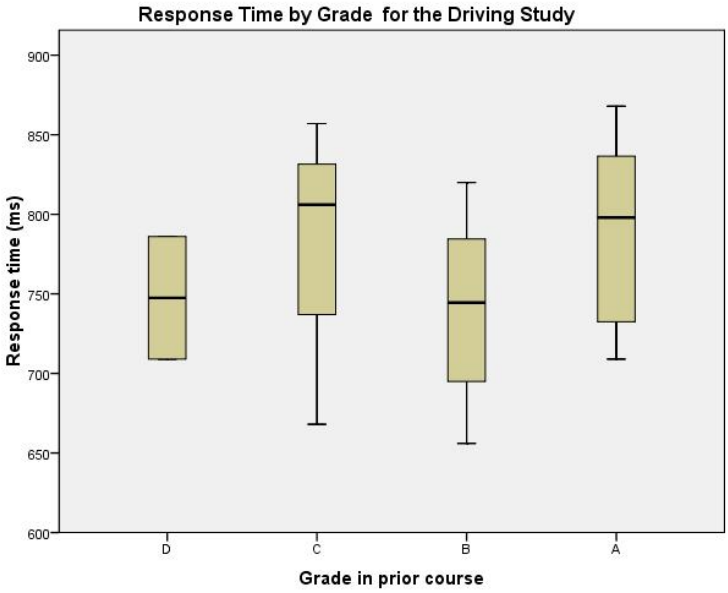


Figure 5.35: SPSS boxplot.

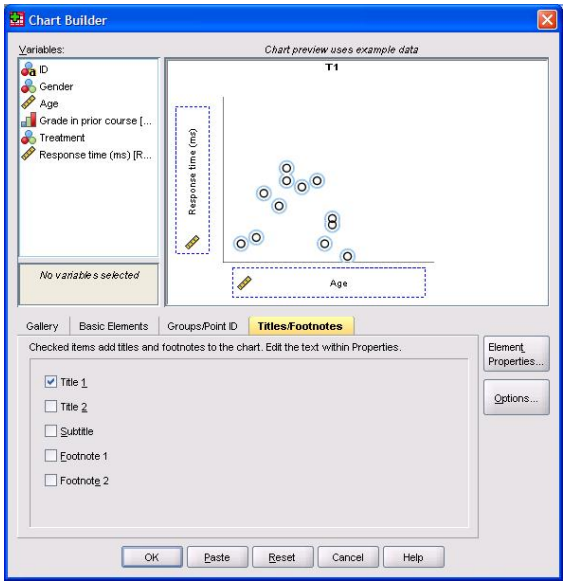


Figure 5.36: SPSS scatterplot setup in Chart Builder.

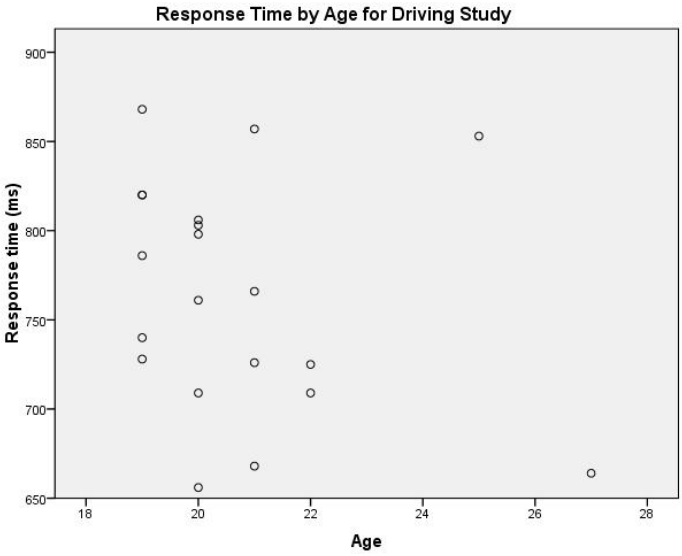


Figure 5.37: SPSS simple scatterplot.

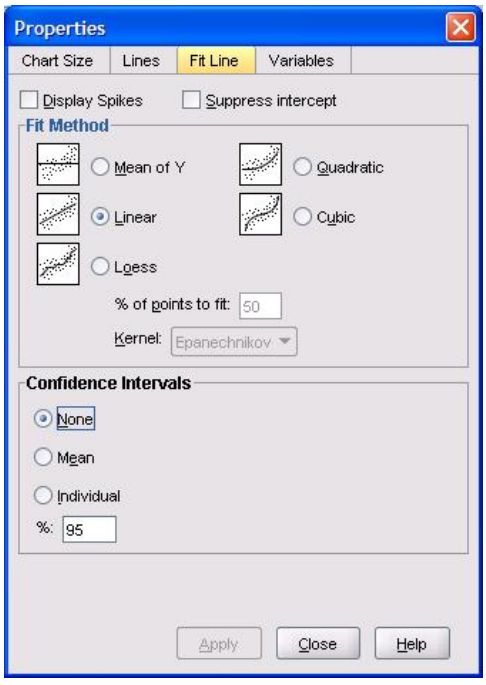


Figure 5.38: SPSS Fit Line tab of Chart Editor.

5.8 SPSS convenience item: Explore

The Analyze/DescriptiveStatistics/Explore menu item in SPSS is a convenience menu item that performs several reasonable EDA steps, both graphical and non-graphical for a quantitative outcome and a categorical explanatory variable (factor). "Explore" is *not* a standard statistical term; it is only an SPSS menu item. So don't use the term in any formal setting!

In the Explore dialog box you can enter one or more quantitative variables in the "Dependent List" box and one or more categorical variables in the "Factor List" box. For each variable in the "Factor List", a complete section of output will be produced. Each section of output examines each of the variables on the "Dependent List" separately. For each outcome variable, graphical and non-graphical EDA are produced that examine the outcome broken down into groups determined by the levels of the "factor". A partial example is given in figure 5.39. In addition to the output shown in the figure, stem-and-leaf plots and side-by-side boxplots are produced by default. The choice of plots and statistics can be changed in the Explore dialog box.

This example has "strength" as the outcome and "sex" as the explanatory variable (factor). The "Case Processing Summary" tells us the number of cases and information about missing data separately for each level of the explanatory variable. The "Descriptives" section gives a variety of statistics for the strength outcome broken down separately for males and females. These statistics include mean and confidence interval on the mean (i.e., the range of means for which we are 95% confident that the true population mean parameter falls in). (The CI is constructed using the "Std. Error" of the mean.) Most of the other statistics should be familiar to you except for the "5% trimmed mean"; this is a "robust" measure of central tendency equal to the mean of the data after throwing away the highest and lowest 5% of the data. As mentioned on page 125, standard errors are calculated for the sample skewness and kurtosis, and these can be used to judge whether the observed values are close or far from zero (which are the expected skewness and kurtosis values for Gaussian data).

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	sex	N	Percent	N	Percent	N	Percent
Strength	Male	20	100.0%	0	.0%	20	100.0%
	Female	21	95.5%	1	4.5%	22	100.0%

Descriptives

	sex			Statistic	Std. Error
Strength	Male	Mean		22.015	.5390
		95% Confidence Interval for Mean	Lower Bound	20.887	
			Upper Bound	23.143	
		5% Trimmed Mean		22.106	
		Median		21.700	
		Variance		5.810	
		Std. Deviation		2.4103	
		Minimum		16.5	
		Maximum		25.9	
		Range		9.4	
		Interquartile Range		3.1	
		Skewness		-.412	.512
		Kurtosis		.016	.992
	Female	Mean		22.119	.4502
		95% Confidence Interval for Mean	Lower Bound	21.180	
			Upper Bound	23.058	
		5% Trimmed Mean		22.191	
		Median		22.300	
		Variance		4.257	
		Std. Deviation		2.0632	
		Minimum		17.8	
		Maximum		25.1	
		Range		7.3	
		Interquartile Range		3.5	
		Skewness		-.408	.501
		Kurtosis		-.727	.972

Figure 5.39: SPSS "Explore" output.

Chapter 6

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. Table 6.1 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, so we will be able to apply the word "causes" rather than just "is associated with" to any statistically significant result. This chapter only discusses so-called "between subjects" explanatory variables, which means that we are assuming that 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 the fMRI experiment).

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 of statistical inference that will be used throughout the book. The understanding of these principles, along with some degree of theoretical underpinning, is key to using statistical results intelligently. Among other things, you need to really understand what a p-value and a confidence interval tell us, and when they can

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

and cannot be trusted.

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.

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 the Threats chapter. For now we will focus on the EDA and confirmatory analyses for a two-group between-subjects experiment with a quantitative outcome. This will give you a general picture of 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.

6.1 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

Note that in SPSS if you enter the "Values" for the two colors and turn on "Value labels", then the color words rather than the numbers will be seen in the second column. Because this data set is not too large, it is possible to examine it to see that 0 and 1 are the only two values for Color and that the time ranges from 291 to 1005 milliseconds (or 0.291 to 1.005 seconds). 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.

Here are some basic univariate exploratory data analyses. 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.

Frequencies

Background Color

		Frequency	Percent Valid	Percent	Cumulative Percent
Valid	yellow	17	48.6	48.6	48.6
	cyan	18	51.4	51.4	100.0
	Total	35	100.0	100.0	

The "Frequency" column gives the basic tabulation of the variable's values. Seventeen subjects were shown a yellow background, and 18 were shown cyan for a total of 35 subjects. The "Percent Valid" vs. "Percent" columns in SPSS differ only if there are missing values. The Percent Valid column always adds to 100% across the categories given, while the Percent column will include a "Missing" category if there are missing data. The Cumulative Percent column accounts for each category *plus* all categories on prior lines of the table; this is not very useful for nominal data.

This is non-graphical EDA. Other non-graphical exploratory analyses of Color, such as calculation of mean, variance, etc. don't make much 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 48.6 vs 51.4 percent numbers.

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

	N	Minimum	Maximum	Mean	Std. Deviation
Reaction Time (ms)	35	291	1005	670.03	180.152

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 180.152. We can calculate that the variance is $180.152^2 = 32454$, but we need to look further at the data to calculate the median or 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 mean plus or minus 2 sd, which is about 310 to 1030. But such an assumption is 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 and possibly find errors.

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. A convenient way to do this in SPSS is with the "Explore" menu option. Abbreviated results are shown in this table and the graphical EDA (side-by-side boxplots) is shown in figure 6.1.

	Background Color			Statistics	Std.Error Std.Error
Reaction Time	Yellow	Mean		679.65	38.657
		95% Confidence Interval for Mean	Lower Bound	587.7	
			Upper Bound	761.60	
		Median		683.05	
		Std. Deviation		159.387	
		Minimum		392	
		Maximum		906	
		Skewness		-0.411	0.550
		Kurtosis		-0.875	1.063
	Cyan	Mean		660.94	47.621
		95% Confidence Interval for Mean	Lower Bound	560.47	
			Upper Bound	761.42	
		Median		662.38	
		Std. Deviation		202.039	
		Minimum		291	
		Maximum		1005	
		Skewness		0.072	0.536
		Kurtosis		-0.897	1.038

Very briefly, the mean reaction times for the subjects shown cyan backgrounds is about 19 ms shorter than the mean for those shown yellow backgrounds. The standard deviation of the reaction times is somewhat larger for the cyan group than it is for the yellow group.

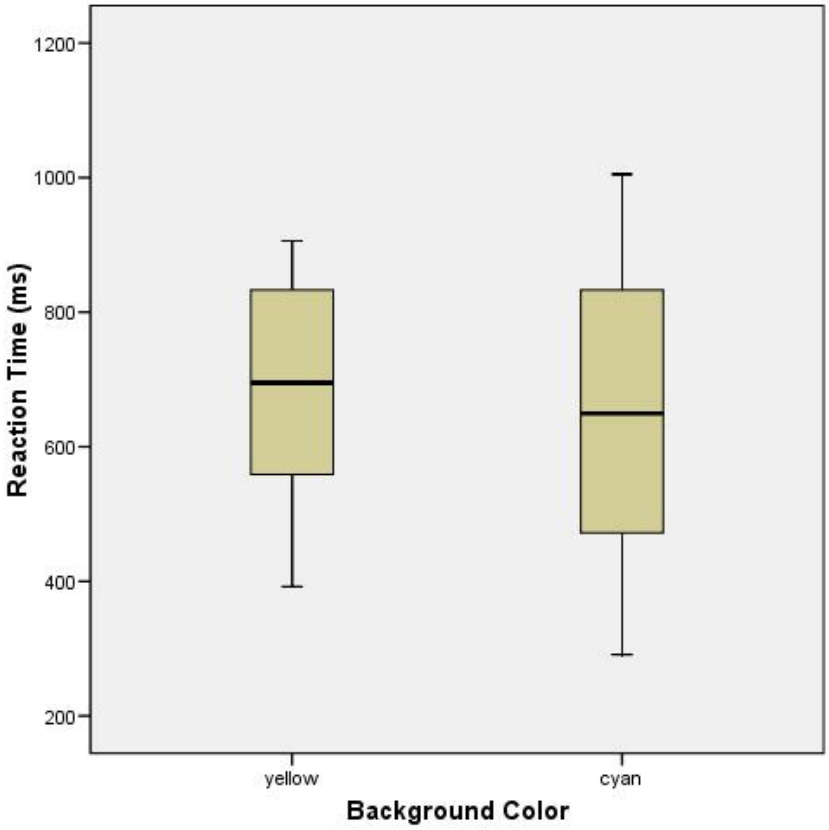


Figure 6.1: Boxplots of reaction time by color.

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.

We should follow up on this EDA with formal statistical testing. But first we need to explore some important concepts underlying such analyses.

6.2 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 8.5) such as measurement variability, environmental variability, treatment application variability, and subject-to-subject variability. The understanding of the concept 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 get if specific conditions are met and if we were to repeat an experiment many times, and then to compare the observed result to these hypothetical results and characterize how "typical" the observed result is.

6.2.1 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.

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.

6.2.2 Model and parameter definition

We start with definition of a model and parameters. 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 "against the facts" a theoretical "counterfactual" result for the treatment they did not get. A useful way to visualize this is to draw each member of the population of interest in the shape of a person. Inside this shape for each actual person (potential subject) are many numbers which are their true values for various outcomes under many different possible conditions (of treatment and environment). If we write the reaction time for a yellow background near the right ear and the reaction time for cyan near the left ear, then the parameter μ_1 is the mean of the right ear numbers over the entire population. It is this parameter, a fixed, unknown "secret of nature" that we want to learn about, **not** the corresponding (noisy) sample quantity for the random sample of subjects randomly assigned to see a yellow background. Put another way, in essentially every experiment that we run, the sample means of the outcomes for the treatment groups 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.

Figure 6.2 shows a diagram demonstrating this way of thinking. The first two subjects of the population are shown along with a few of their attributes. The population mean of any attribute is a parameter that may be of interest in a particular experiment. Obviously we can define many parameters (means, variances, etc.) for many different possible attributes, both marginally and conditionally on other attributes, such as age, gender, etc. (see section 3.6).

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 meaning of "in the same group" is obvious here, but is less so, e.g., in regression models.) We assume that we cannot predict the deviation of individual measurements from the group mean more exactly than saying that they randomly follow the probability distribution of the error model.

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 (step 5). Similarly, the assumption of

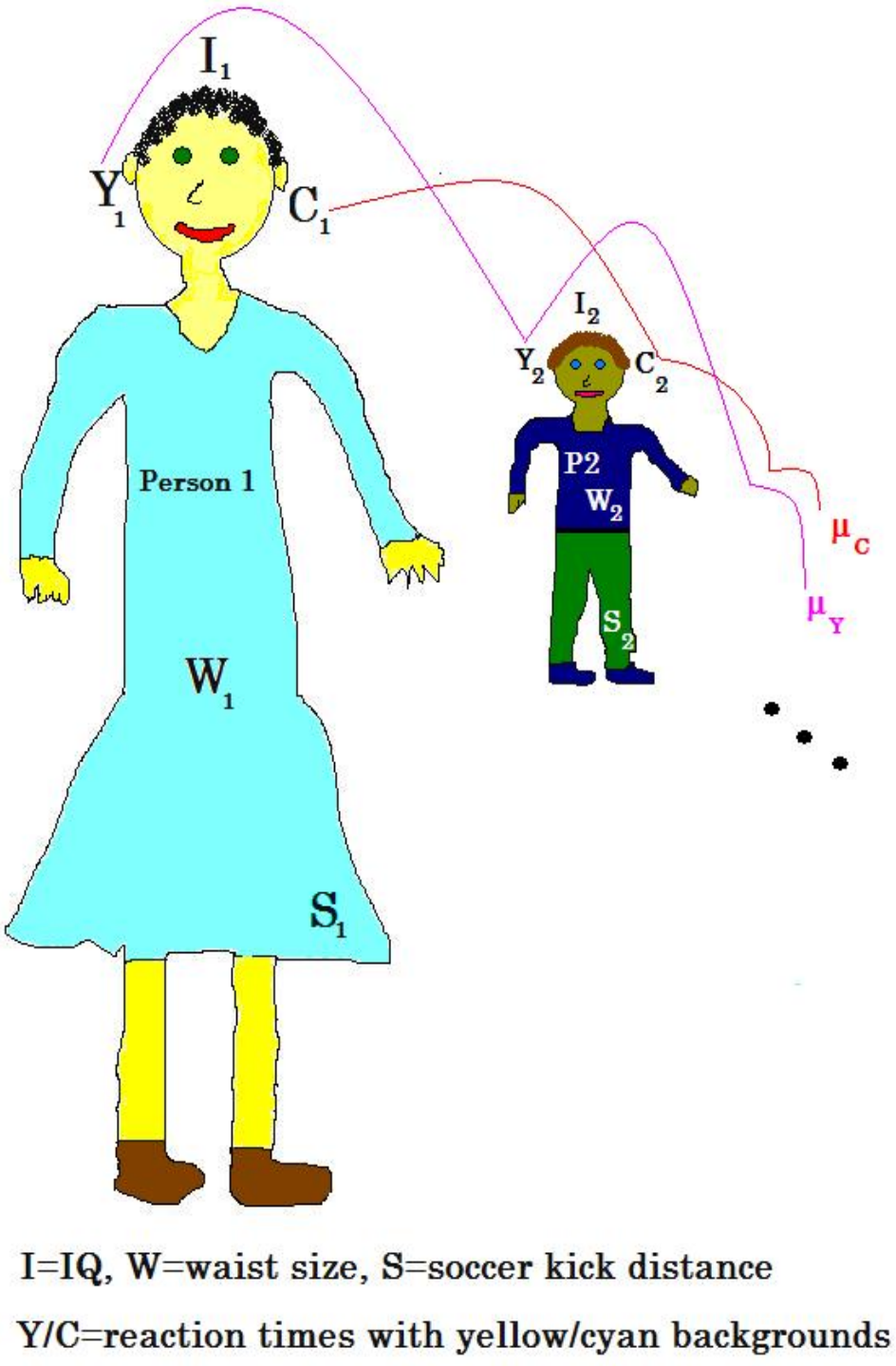


Figure 6.2: A view of a population and parameters.

equal variance is often but not always true.

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.

6.2.3 Null and alternative hypotheses

The null and alternative hypotheses are statements about *the population parameters* that express different possible characterizations of the population which correspond to different scientific hypotheses. 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. Note that the null hypothesis is usually interpretable as "nothing interesting is going on," and that is why the term null is used.

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 we use an alternative hypothesis that states that one population mean is less than the other, but in my opinion 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.

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

6.2.4 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. 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. In fact we will encounter situations where no one statistic is optimal, and different researchers might choose different statistics for their formal statistical analyses.

Inference is usually based on a single statistic whose choice may or may not be obvious or unique.

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

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

where n_1 and n_2 are the group sample sizes. Note that this simplifies to $\sqrt{2}\sigma/\sqrt{n}$ when the sample sizes are equal.

In practice the estimate of the SE that uses an appropriate averaging

of the observed sample variances is used.

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

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.

6.2.5 Computing the null sampling distribution

The next step in the general scheme of formal (classical) statistical analysis 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 calculated over 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.

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.

Probability theory (beyond the scope of this book) comes into play in computing the null sampling distribution of the chosen statistic based on the model assumptions.

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.

6.2.6 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 (experimental) 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. 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 reject the null hypothesis incorrectly when the underlying truth is that the null hypothesis 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.) A synonym for Type 1 error is "false rejection" of the null hypothesis.

The usual way that we make a formal, objective reject vs. retain decision 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 the observed value in our actual experiment or something more extreme (in the sense of less compatible with the null hypotheses), when 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 the observed value in our actual experiment or something more extreme, 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 6.3 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.

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

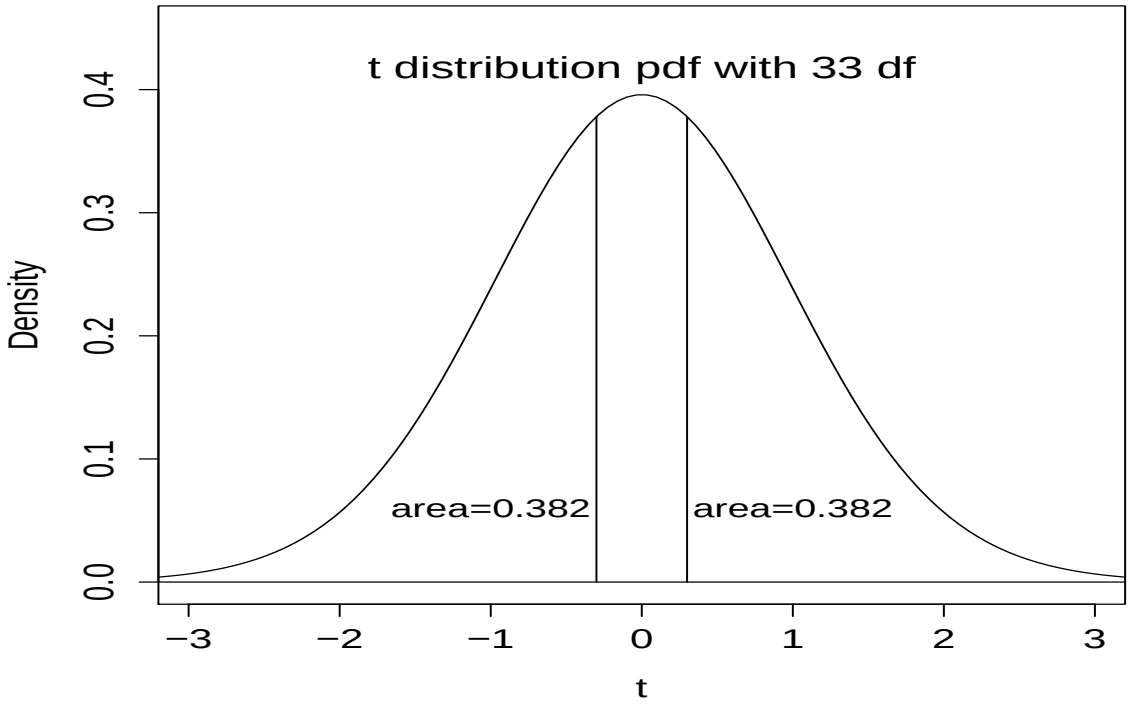


Figure 6.3: Calculation of the p-value for the HCI example

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, 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 12.

You should *never use the word "insignificant"* to indicate a large p-value. Use "not significant" or "non-significant" because "insignificant" implies no substantive significance rather than no statistical significance.

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 is 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, 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 6.2.10 for more on this "power" problem.

We talk about an "incorrect" p-value mostly with regard to the situation where the null hypothesis is the underlying truth. It is really the behavior of the p-value over repeats of the experiment that is incorrect, and we want to identify what can cause that to happen even though we will usually see only a single p-value for an experiment. Because the p-value for an experiment is computed as an area under the pdf of the null sampling distribution of a statistic, the main reason a p-value is "incorrect" (and therefore misleading) is that we are not using the appropriate null sampling distribution. That happens when the model assumptions used in the computation of the null sampling distribution of the statistic are not close to the reality of nature. For the t-test, this can be caused by non-normality of the distributions (though this is not a problem if the sample size is large due to 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 meaning, and it is no longer an interpretable quantity.

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 incorrect p-values.

6.2.7 Confidence intervals

Besides p-values, another way to express what the evidence of an experiment is telling us 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. That is not possible! 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 arbitrary; we can choose another number like 99% or 75%, but note that 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.

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.

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, it would not be substantively significant.

In contrast, if we had a 95% CI of [-30, 200] dozen per month with p=0.15, then even though this 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 add 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.

The slight downside to CIs and substantive significance is that they are hard to interpret if you don't know much about your subject matter. This is usually only a problem for learning exercises, not for real experiments.

6.2.8 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 perform some assumption checking 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 observed value for each outcome observation is deviated higher or lower than the true mean. The error component of our model describes the distribution of these deviations, which are called **errors**. The residuals, which are

defined as observed minus expected value for each outcome measurement, are our best estimates of the unknowable, true errors for each subject. We will examine the distribution of the residuals to allow us to make a judgment about whether or not the distribution of the errors is consistent with the error model.

Assumption checking is needed to verify that the assumptions involved in the initial model construction were good enough to allow us to believe our inferences.

Defining groups among which all subjects have identical predictions may be complicated for some models, but is simple for the 2-treatment-group model. For this situation, all subjects in either one of the two treatment groups appear to be identical in the model, so they must have the same prediction based on the model. 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).

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. 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 say "all models are wrong, but some are useful." 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 vari-

ance 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.

6.2.9 Subject matter conclusions

Applying subject matter knowledge to the confidence interval is one key form of relating statistical conclusions back to the subject matter of the experiment. For p-values, you do something similar with the reject/retain result of your decision rule. In either case, an analysis is incomplete if you stop at reporting the p-value and/or CI without returning to the original scientific question(s).

6.2.10 Power

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%). 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 100-100β%. 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.

It should be fairly obvious that for any given experiment you have more power to detect a large effect than a small one.

You should use the methods of chapter 12 to estimate the power of any experiment before running it. This is only an estimate or educated guess because some

needed information is usually not known. Many, many experiments are performed which have insufficient power, often in the 20-30% range. This is horrible! It means that even if you are studying effective treatments, you only have a 20-30% chance of getting a statistically significant result. Combining power analysis with intelligent experimental design to alter the conduct of the experiment to maximize its power is a quality of a good scientist.

Poor power is a common problem. It cannot be fixed by statistical analysis. It must be dealt with before running your experiment.

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, you know that a true null hypothesis and a Type 2 error are not distinguishable. But if you have good power for some reasonably important sized effect, then a large p-value is good evidence that no important sized effect exists, although a Type 2 error is still possible.

A non-significant p-value and a low power combine to make an experiment totally uninformative.

In a nutshell: All classical statistical inference is based on the same set of steps in which a sample statistic is compared to the kinds of values we would expect it to have if nothing interesting is going on, i.e., if the null hypothesis is true.

6.3 Do it in SPSS

Figure 6.4 shows the Independent Samples T-test dialog box.

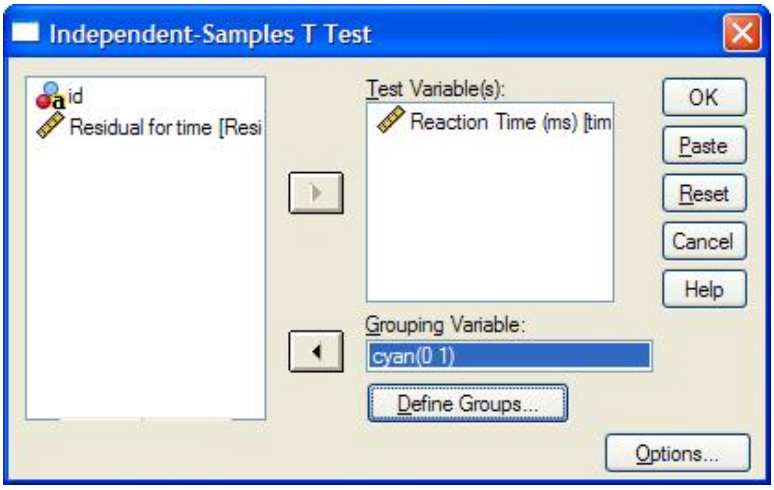


Figure 6.4: SPSS "Explore" output.

Before performing the t-test, check that your outcome variable has Measure "scale" and that you know the numeric codes for the two levels of your categorical (nominal) explanatory variable.

To perform an independent samples t-test in SPSS, use the menu item "Independent Samples T-Test" found under Analyze/CompareMeans. Enter the outcome (dependent) variable into the Test Variables box. Enter the categorical explanatory variable into the Grouping Variable box. Click "Define Groups" and enter the numeric codes for the two levels of the explanatory variable and click Continue. Then click OK to produce the output. (The t-statistic will be calculated in the direction that subtracts the level you enter second from the level you enter first.)

For the HCI example, put Reaction Time in the Test Variables box, and Background Color in the Grouping Variable box. For Define Groups enter the codes 0 and 1.

6.4 Return to the HCI example

The SPSS output for the independent samples (two-sample) t-test for the HCI text background color example is shown in figure 6.5.

The group statistics are very important. In addition to verifying that all of

Group Statistics

	Background Color	N	Mean	Std. Deviation	Std. Error Mean
Reaction Time (ms)	yellow	17	679.65	159.387	38.657
	cyan	18	660.94	202.039	47.621

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reaction Time (ms)	Equal variances assumed	1.250	.272	.30	33	.764	18.703	61.758	-106.94	144.35
	Equal variances not assumed			.30	32.02	.762	18.703	61.336	-106.23	143.64

Figure 6.5: t-test for background experiment.

the subjects were included in the analysis, they let us see which group did better. *Reporting a statistically significant difference without knowing in which direction the effect runs is a cardinal sin in statistics!* Here we see that the mean reaction time for the "yellow" group is 680 ms while the mean for the "cyan" group is 661 ms. If we find a statistically significant difference, the direction of the effect is that those tested with a cyan background performed better (faster reaction time). The sample standard deviation tells us about the variability of reaction times: if the reaction times are roughly Normal in distribution, then approximately 2/3 of the people when shown a yellow background score within 159 ms of the mean of 680 ms (i.e., between 521 and 839 ms), and approximately 95% of the people shown a yellow background score within 2*159=318 ms of 680 ms. Other than some uncertainty in the sample mean and standard deviation, this conclusion is unaffected by changing the size of the sample.

The means from "group statistics" show the direction of the effect and the standard deviations tell us about the inherent variability of what we are measuring.

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). I am using "pin down the population mean" as a way of saying that we want to quantify in a probabilistic sense in what possible interval our evidence places the population mean and how confident we are that it really falls into that interval. In other words we want to construct **confidence intervals** for the group population means.

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 1.96 (or its rounded value, 2) 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 value (quantile) of a t-distribution that defines a central probability of 95% (or some other value we choose). This value is calculated by the computer (or read off of a table of the t-distribution), but it does depend 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: if we are more uncertain about the true value of σ, we need to make a less well defined (wider) claim about where μ is.

So practically we interpret the SEM this way: we are roughly 95% certain that the true mean (μ) is within about 2 SEM of the sample mean (unless the sample size is small).

The mean and standard error of the mean from "group statistics" tell us about how well we have "pinned down" the population mean based on the inherent variability of the measure and the sample size.

The "Independent Samples Test" box shows the actual t-test results under the row labeled "Equal variances assumed". The columns labeled "Levene's Test for Equality of Variances" are *not* part of the t-test; they are part of a supplementary test of the assumption of equality of variances for the two groups. If the Levene's Test p-value (labeled "Sig" , for "significance", in SPSS output) is less than or equal to 0.05 then we would begin to worry that the equal variance assumption is violated, thus casting doubt on the validity of the t-test's p-value. For our example, the Levene's test p-value of 0.272 suggests that there is no need for worry about

that particular assumption.

The seven columns under "t-test for Equality of Means" are the actual t-test results. The t-statistic is given as 0.30. It is negative when the mean of the second group entered is larger than that of the first. The degrees of freedom are given under "df". The p-value is given under "Sig. (2-tailed)". The actual difference of the means is given next. The standard error of that difference is given next. Note that the t-statistic is computed from the difference of means and the SE of that difference as difference/(SE of difference). Finally a 95% confidence interval is given for the difference of means. (You can use the Options button to compute a different sized confidence interval.)

SPSS (but not many other programs) automatically gives a second line labeled "Equal variances not assumed". This is from one of the adjusted formulas to correct for unequal group variances. The computation of a p-value in the unequal variance case is quite an unsettled and contentious problem (called the Behrens-Fisher problem) and the answer given by SPSS is reasonably good, but not generally agreed upon. So if the p-value of the Levene's test is less than or equal to 0.05, many people would use the second line to compute an adjusted p-value ("Sig. (2-tailed)"), SEM, and CI based on a different null sampling distribution for the t-statistic in which the df are adjusted an appropriate amount downward. If there is no evidence of unequal variances, the second line is just ignored.

For model assumption checking, figure 6.6 shows separate histograms of the residuals for the two groups with overlaid Normal pdfs. With such a small sample size, we cannot expect perfectly shaped Normal distributions, even if the Normal error model is perfectly true. The histograms of the residuals in this figure look reasonably consistent with Normal distributions with fairly equal standard deviation, 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 evidence of the kind of gross violation of these assumptions that would make us conclude that the p-value is likely to be highly misleading. In more complex models, we will usually substitute a "residual vs. fit" plot and a quantile-normal plot of the residuals for these assumption checking plots.

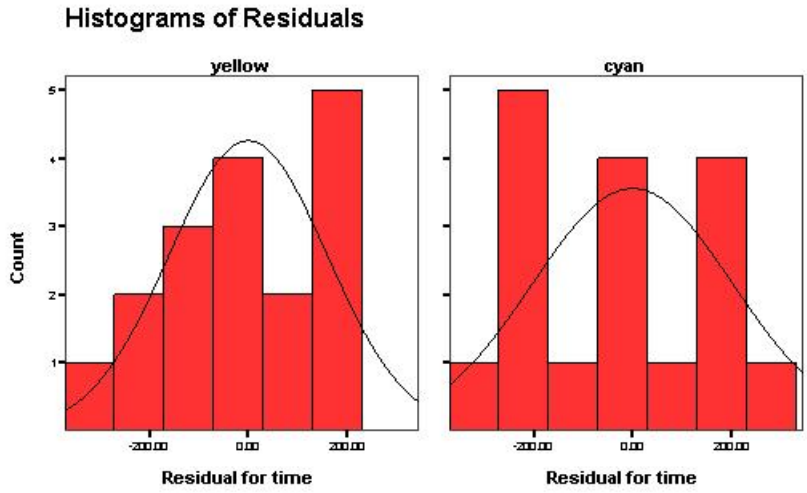


Figure 6.6: Histograms of residuals.

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.

Chapter 7

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 6 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. 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 Chapter 14. 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.

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.

7.1 Moral Sentiment Example

As an example of application of one-way ANOVA 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 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 8.1 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

Typical exploratory data analyses include a tabulation of the frequencies of the levels of a categorical explanatory variable like "emotion". Here we see 39 controls, 42 guilt subjects, and 45 shame subjects. Some sample statistics of cooperation broken down by each level of induced emotion are shown in table 7.1, and side-by-

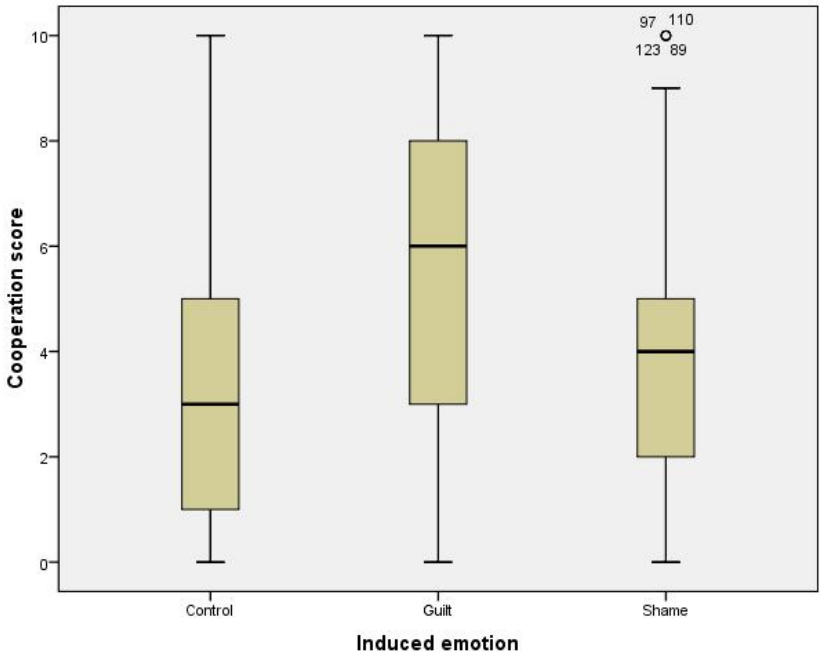


Figure 7.1: Boxplots of cooperation by induced emotion.

side boxplots shown in figure 7.1.

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 pre-checking model assumptions, 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.

	Induced emotion			Statistic	Std.Error
Cooperation score	Control	Mean		3.49	0.50
		95% Confidence Interval for Mean	Lower Bound	2.48	
			Upper Bound	4.50	
		Median		3.00	
		Std. Deviation		3.11	
		Minimum		0	
		Maximum		10	
		Skewness		0.57	0.38
		Kurtosis		-0.81	0.74
	Guilt	Mean		5.38	0.50
		95% Confidence Interval for Mean	Lower Bound	4.37	
			Upper Bound	6.39	
		Median		6.00	
		Std. Deviation		3.25	
		Minimum		0	
		Maximum		10	
		Skewness		-0.19	0.36
		Kurtosis		-1.17	0.72
	Shame	Mean		3.78	0.44
		95% Confidence Interval for Mean	Lower Bound	2.89	
			Upper Bound	4.66	
		Median		4.00	
		Std. Deviation		2.95	
		Minimum		0	
		Maximum		10	
		Skewness		0.71	0.35
		Kurtosis		-0.20	0.70

Table 7.1: Group statistics for the moral sentiment experiment.

7.2 How one-way ANOVA works

7.2.1 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 all 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 $N(\mu_i, \sigma^2)$.

Our 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. Because it is common that subjects who have a high outcome when given one treatment tend to have a high outcome when given another treatment, using the same subject twice would violate the independence assumption.

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.

(In case you have forgotten, the Greek capital sigma (Σ) stands for summation, i.e., adding. In this case, the notation says that we should consider all values of n_i where i is set to 1, 2, ..., k, and then add them all up. For example, if we have $k = 3$ levels of treatment, and the group samples sizes are 12, 11, and 14 respectively, then $n_1 = 12$, $n_2 = 11$, $n_3 = 14$ and $N = \sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 = 12 + 11 + 14 = 37$.)

Because of the random treatment assignment, the sample mean for any treatment group is representative of the population mean for assignment to that group for the entire population.

Technically, the sample group means are unbiased estimators of the population group means when treatment is randomly assigned. 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 term **assumption** in statistics refers to any specific part of a statistical model. For one-way ANOVA, the assumptions are normality, equal variance, and independence of errors. Correct assignment of individuals to groups is sometimes considered to be an implicit assumption.

The null hypothesis is a point hypothesis stating that "nothing interesting is happening." For one-way ANOVA, we use $H_0 : \mu_1 = \cdots = \mu_k$, 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 null hypothesis is called the "overall" null hypothesis and is the hypothesis tested by ANOVA, per se. If we have only two levels of our categorical explanatory

variable, then retaining or rejecting the overall null hypothesis, is all that needs to be done in terms of hypothesis testing. But if we have 3 or more levels ($k \geq 3$), then we usually need to followup on rejection of the overall null hypothesis with more specific hypotheses to determine for which population group means we have evidence of a difference. This is called contrast testing and discussion of it will be delayed until chapter 13.

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".

7.2.2 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. For ANOVA, we use the "F-statistic". The single formula for the F-statistic that is shown in most textbooks is quite complex and hard to understand. But we can build it up in small understandable steps.

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 69). In ANOVA we work with variances and also "variance-like quantities" which are not really the variance of anything, but are still calculated as SS/df. We will call all of these quantities **mean squares** or MS. i.e., $MS = SS/df$, which is a key formula that you should memorize. Note that these are not really means, because the denominator is the df, not n.

For one-way ANOVA we will work with two different MS values called "mean square within-groups", $\text{MS}_{\text{within}}$, and "mean square between-groups", $\text{MS}_{\text{between}}$. We know the general formula for any MS, so we really just need to find the formulas for $\text{SS}_{\text{within}}$ and $\text{SS}_{\text{between}}$, and their corresponding df.

The F statistic denominator: $\text{MS}_{\text{within}}$

$\text{MS}_{\text{within}}$ is a "pure" estimate of σ^2 that is unaffected by whether the null or alternative hypothesis is true. Consider figure 7.2 which represents 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. The extension to more groups and/or different numbers of subjects is straightforward.

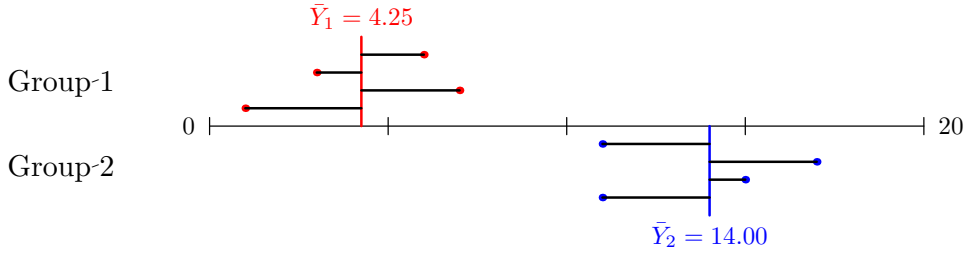


Figure 7.2: Deviations for within-group sum of squares

The deviation for subject j of group i in figure 7.2 is mathematically equal to $Y_{ij} - \bar{Y}_i$ where Y_{ij} is the observed value for subject j of group i and $\bar{Y}_i$ is the sample mean for group i.

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. The variation has standard deviation σ, so that, e.g., about 2/3 of the times the deviation lines are shorter than σ. Regardless of the truth of the null hypothesis, for each individual group, $\text{MS}_i = \text{SS}_i/\text{df}_i$ is a good estimate of σ^2. The value of $\text{MS}_{\text{within}}$ comes from a statistically appropriate formula for 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.

For an individual group, i, $\text{SS}_i = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ and $\text{df}_i = n_i - 1$. We can use some statistical theory beyond the scope of this course to show that in general, $\text{MS}_{\text{within}}$ is a good (unbiased) estimate of σ^2 if it is defined as

$$\text{MS}_{\text{within}} = \text{SS}_{\text{within}}/\text{df}_{\text{within}}$$

where $\text{SS}_{\text{within}} = \sum_{i=1}^{k} SS_i$, and $\text{df}_{\text{within}} = \sum_{i=1}^{k} \text{df}_i = \sum_{i=1}^{k}(n_i - 1) = N - k$.

$\text{MS}_{\text{within}}$ is a good estimate of σ^2 (from our model) regardless of the truth of H_0. This is due to the way $\text{SS}_{\text{within}}$ is defined. $\text{SS}_{\text{within}}$ (and therefore $\text{MS}_{\text{within}}$) has N-k degrees of freedom with $n_i - 1$ coming from each of the k groups.

The F statistic numerator: $\text{MS}_{\text{between}}$

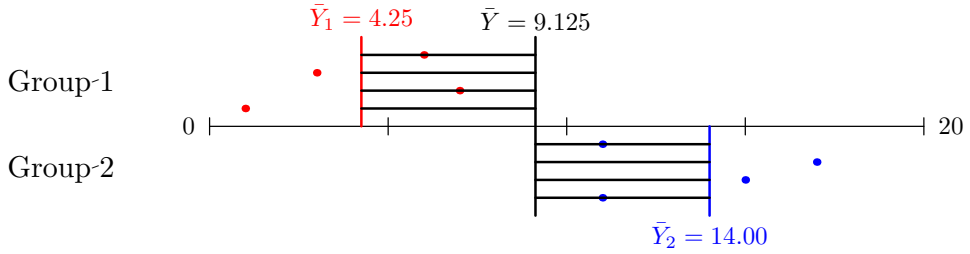


Figure 7.3: Deviations for between-group sum of squares

Now consider figure 7.3 which represents the between-group deviations used in the calculation of $\text{MS}_{\text{between}}$ for the same little 2-group 8-subject experiment as shown in figure 7.2. The single vertical black line is the average of all of the outcomes values in all of the treatment groups, usually called either the overall mean or the **grand mean**. 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 distance (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".)

$\text{SS}_{\text{between}}$ is the sum of the N squared between-group deviations, where the deviation is the same for all subjects in the same group. The formula is

$$\text{SS}_{\text{between}} = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{\bar{Y}})^2$$

where $\bar{\bar{Y}}$ is the grand mean. Because the k unique deviations add up to zero, we are free to choose only $k-1$ of them, and then the last one is fully determined by the others, which is why $\text{df}_{\text{between}} = k-1$ for one-way ANOVA.

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.

The F statistic ratio

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.

7.2.3 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.

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 it to its null sampling distribution which is the specific F-distribution which has degrees of freedom matching the numerator and denominator of the F-statistic.

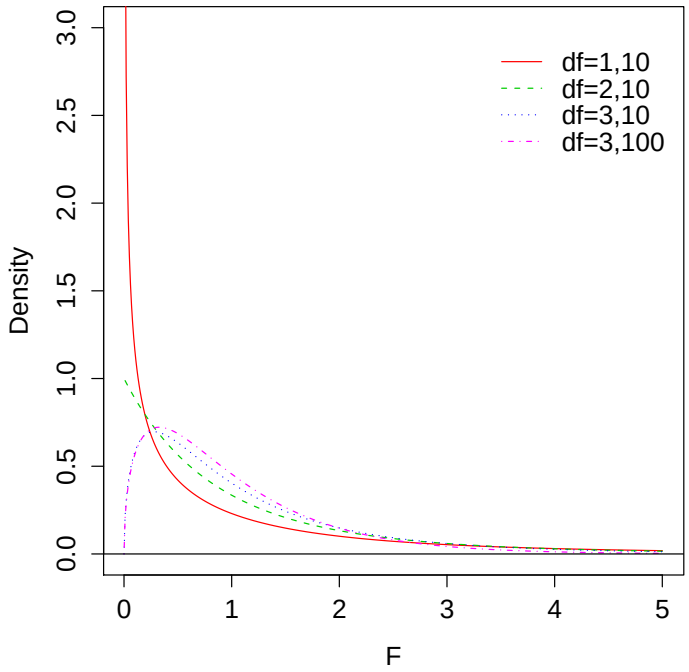


Figure 7.4: A variety of F-distribution pdfs.

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 7.4. 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

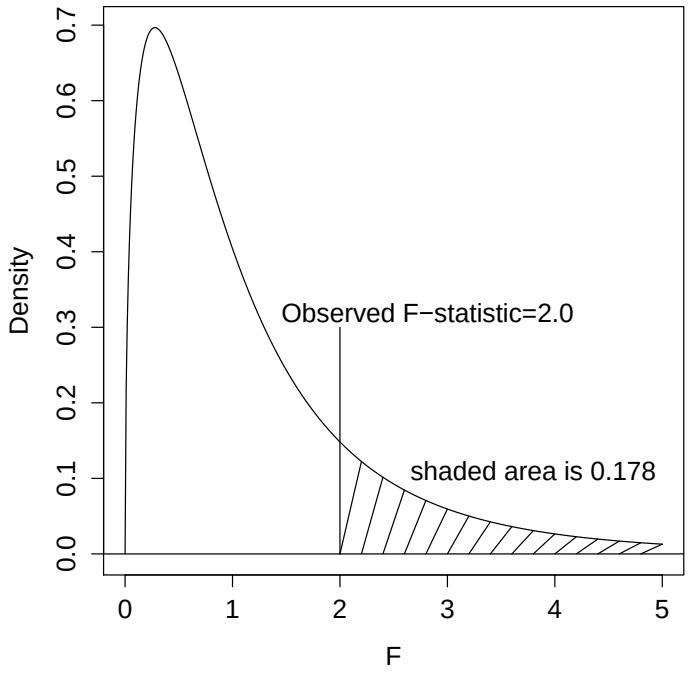


Figure 7.5: The F(3,10) pdf and the p-value for F=2.0.

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.

7.2.4 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).

A close up of the F-distribution with 3 and 10 degrees of freedom is shown in figure 7.5. 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

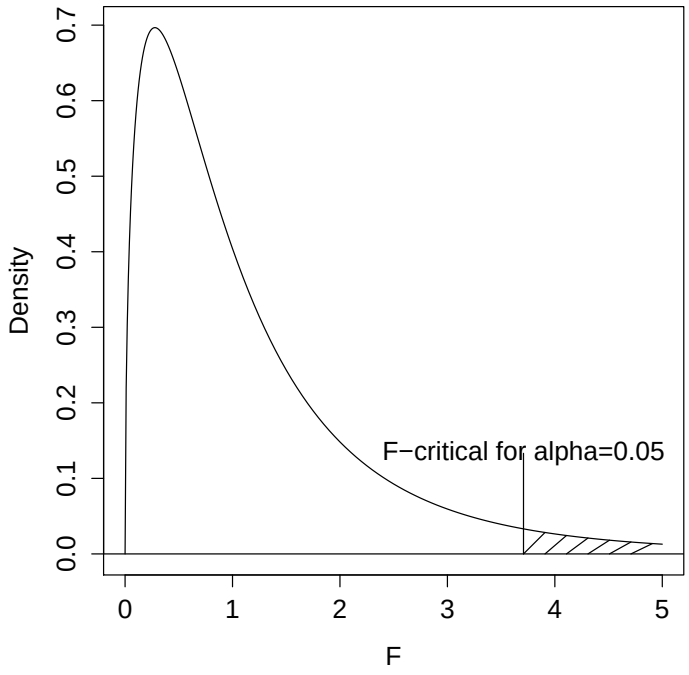


Figure 7.6: The F(3,10) pdf and its alpha=0.05 critical value.

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 7.6 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.

7.2.5 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 13.

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 6.2.7, 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 $\text{MS}_{\text{between}}/\text{MS}_{\text{within}}$. 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 $p \leq \alpha$.

7.3 Do it in SPSS

To run a one-way ANOVA in SPSS, use the Analyze menu, select Compare Means, then One-Way ANOVA. Add the quantitative outcome variable to the "Dependent List", and the categorical explanatory variable to the "Factor" box. Click OK to get the output. The dialog box for One-Way ANOVA is shown in figure 7.7.

You can also use the Options button to perform descriptive statistics by group, perform a variance homogeneity test, or make a means plot.

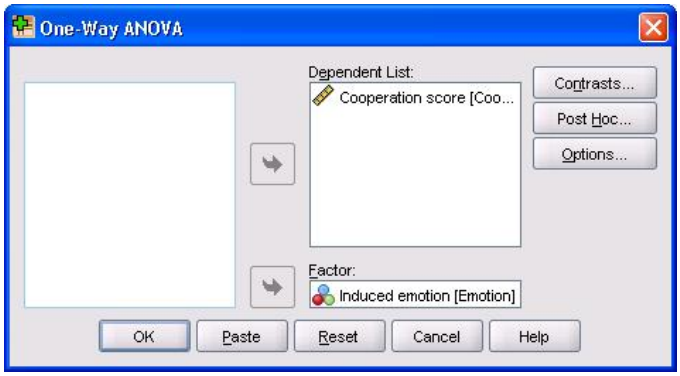


Figure 7.7: One-Way ANOVA dialog box.

You can use the Contrasts button to specify particular planned contrasts among the levels or you can use the Post-Hoc button to make unplanned contrasts (corrected for multiple comparisons), usually using the Tukey procedure for all pairs or the Dunnett procedure when comparing each level to a control level. See chapter 13 for more information.

7.4 Reading the ANOVA table

The **ANOVA table** is the main output of an ANOVA analysis. It always has the "source of variation" labels in the first column, plus additional columns for "sum of squares", "degrees of freedom", "means square", F, and the p-value (labeled "Sig." in SPSS).

For one-way ANOVA, there are always rows for "Between Groups" variation and "Within Groups" variation, and often a row for "Total" variation. In one-way ANOVA there is only a single F statistic ($\text{MS}_{\text{between}}/\text{MS}_{\text{within}}$), and this is shown on the "Between Groups" row. There is also only one p-value, because there is only one (overall) null hypothesis, namely $H_0 : \mu_1 = \cdots = \mu_k$, and because the p-value comes from comparing the (single) F value to its null sampling distribution. The calculation of MS for the total row is optional.

Table 7.2 shows the results for the moral sentiment experiment. There are several important aspects to this table that you should understand. First, as discussed above, the "Between Groups" lines refer to the variation of the group means around the grand mean, and the "Within Groups" line refers to the variation

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	86.35	2	43.18	4.50	0.013
Within Groups	1181.43	123	9.60		
Total	1267.78	125			

Table 7.2: ANOVA for the moral sentiment experiment.

of the subjects around their group means. The "Total" line refers to variation of the individual subjects around the grand mean. The Mean Square for the Total line is exactly the same as the variance of all of the data, ignoring the group assignments.

In any ANOVA table, the df column refers to the number of degrees of freedom in the particular SS defined on the same line. The MS on any line is always equal to the SS/df for that line. F-statistics are given on the line that has the MS that is the numerator of the F-statistic (ratio). The denominator comes from the MS of the "Within Groups" line for one-way ANOVA, but this is not always true for other types of ANOVA. It is always true that there is a p-value for each F-statistic, and that the p-value is the area under the null sampling distribution of that F-statistic that is above the (observed) F value shown in the table. Also, we can always tell which F-distribution is the appropriate null sampling distribution for any F-statistic, by finding the numerator and denominator df in the table.

An ANOVA is a breakdown of the total variation of the data, in the form of SS and df, into smaller independent components. For the 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, and then deviations of the individuals from their group means. 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 $\text{df}_{\text{Total}} = \text{df}_{\text{Between}} + \text{df}_{\text{Within}}$. This fact can be used to reduce the amount of calculation, or just to check that the calculation were done and recorded correctly.

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 full 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, 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 $\sqrt{9.60} = 3.10$, 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.

7.5 Assumption checking

Except for the skewness of the shame group, the skewness and kurtosis statistics for all three groups are within 2SE of zero (see Table 7.1), and that one skewness is only slightly beyond 2SE from zero. This suggests that there is no evidence against the Normality assumption. The close similarity of the three group standard deviations suggests that the equal variance assumption is OK. And hopefully the subjects are totally unrelated, so the independent errors assumption is OK. Therefore we can accept that the F-distribution used to calculate the p-value from the F-statistic is the correct one, and we "believe" the p-value.

7.6 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. Again, we defer looking at *which* groups differ to chapter 13.

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

The F-statistic of one-way ANOVA is easily calculated by a computer. The p-value is calculated from the F null sampling distribution with matching degrees of freedom. But only if we believe that the assumptions of the model are (approximately) correct should we believe that the p-value was calculated from the correct sampling distribution, and it is then valid.

Chapter 8

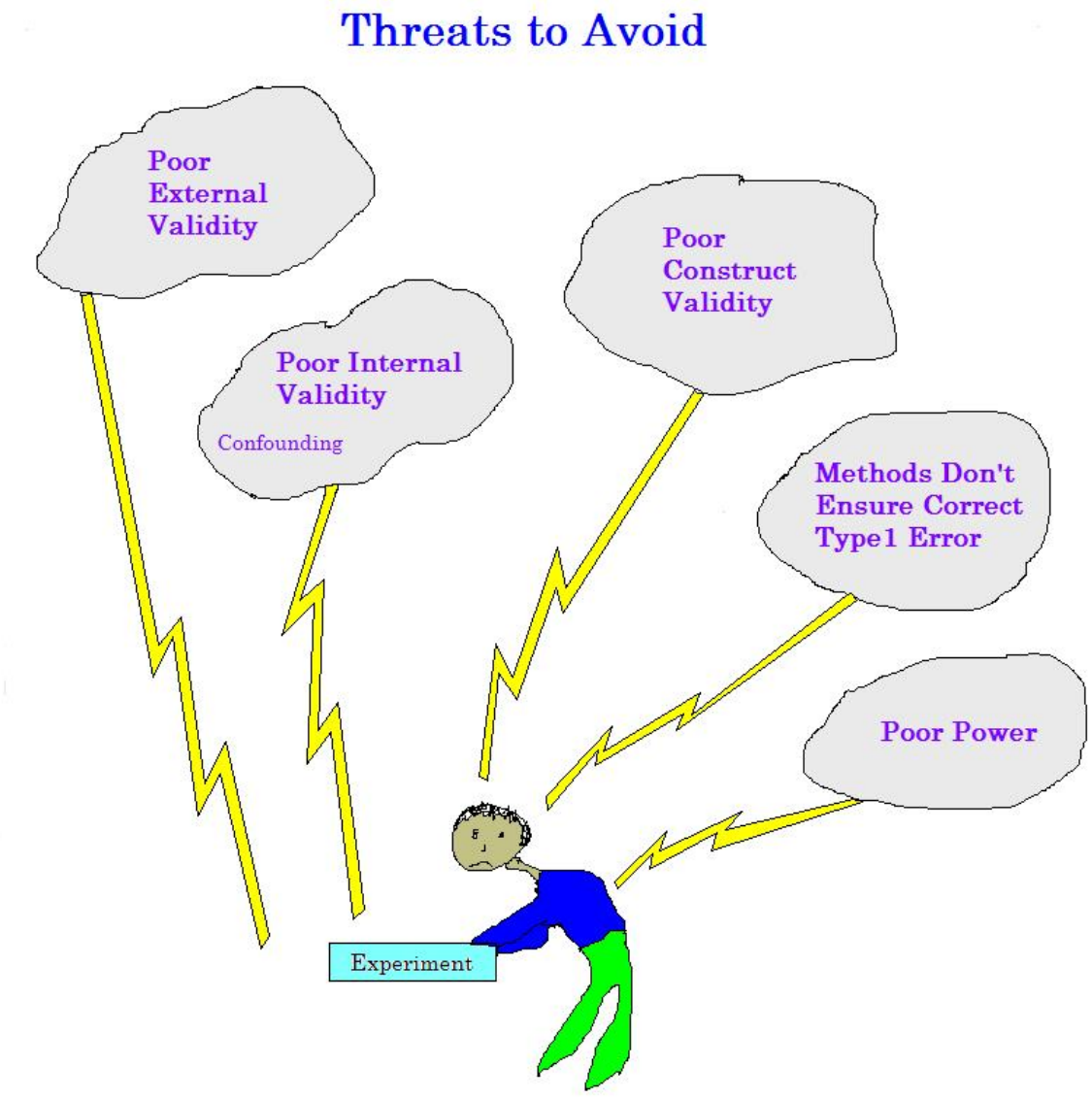
Threats to Your Experiment

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.



8.1 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.

If variables X and Y (e.g., the number of televisions (X) in various countries and the infant mortality rate (Y) of those countries) are found to be associated, then there are three basic possibilities. First X could be causing Y (televisions lead to more health awareness, which leads to better prenatal care) or Y could be causing X (high infant mortality leads to attraction of funds from richer countries, which leads to more televisions) or unknown factor Z could be causing both X and Y (higher wealth in a country leads to more televisions and more prenatal care clinics). It is worth memorizing these three cases, because they should always be considered when association is found in an observational study as opposed to a randomized experiment. (It is also possible that X and Y are related in more complicated ways including in large networks of variables with feedback loops.)

Causation ("X causes Y") can be logically claimed if X and Y are associated, and X precedes Y, and no plausible alternative explanations can be found, particularly those of the form "X just happens to vary along with some real cause of changes in Y" (called confounding).

Returning to the advertisement example, one stupid thing to do is to place all of the high status pictures in only the wealthiest neighborhoods or the largest stores,

while the low status pictures are only shown in impoverished neighborhoods or those with smaller stores. In that case a higher average number of items purchased for the stores with high status ads may be either due to the effect of socio-economic status or store size or perceived status of the ad. When more than one thing is different on average between the groups to be compared, the problem is called **confounding** and confounding is a fatal threat to internal validity.

Notice that the definition of confounding mentions "different on average". This is because it is practically impossible to have no differences between the subjects in different groups (beyond the differences in treatment). So our realistic goal is to have no difference on average. For example if we are studying both males and females, we would like the gender ratio to be the same in each treatment group. For the store example, we want the average pre-treatment total sales to be the same in each treatment group. And we want the distance from competitors to be the same, and the socio-economic status (SES) of the neighborhood, and the racial makeup, and the age distribution of the neighborhood, etc., etc. Even worse, we want all of the unmeasured variables, both those that we thought of and those we didn't think of, to be similar in each treatment group.

The sine qua non of internal validity is **random assignment of treatment** to experimental units (different stores in our ad example). Random treatment assignment (also called randomization) is usually the best way to assure that all of the potential confounding variables are equal on average (also called balanced) among the treatment groups. Non-random assignment will usually lead to either consciously or unconsciously unbalanced groups. If one or a few variables, such as gender or SES, are known to be critical factors affecting outcome, a good alternative is **block randomization**, in which randomization among treatments is performed separately for each level of the critical (non-manipulated) explanatory factor. This helps to assure that the level of this explanatory factor is balanced (not confounded) across the levels of the treatment variable.

In current practice 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 also acceptable to have the computer individually generate a random number from 1 to k for each subject, but it must be assured that the subject and/or researcher cannot re-run the process if they don't like the assignment.

Confounding can occur because we purposefully, but stupidly, design our experiment such that two or more things differ at once, or because we assign treatments non-randomly, or because the randomization "failed". As an example of designed confounding, consider the treatments "drug plus psychotherapy" vs. "placebo" for treating depression. If a difference is found, then we will not know whether the success of the treatment is due to the drug, the psychotherapy or the 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 better to have a study with the three treatment arms of drug, psychotherapy, and combination (with or without the placebo), so that we could assess the specific important questions of whether drug adds a benefit to psychotherapy and vice versa. 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 that the mix itself is effective while the individual components are not, so this might be a good experiment.

In terms of non-random assignment of treatment, this should only be done when necessary, and it should be recognized that it strongly, often fatally, harms the internal validity of the experiment. If you assign treatment in some pseudo-random way, e.g. alternating treatment levels, you or the subjects may purposely or inadvertently introduce confounding factors into your experiment.

Finally, it must be stated that although randomization cannot perfectly balance all possible explanatory factors, it is the best way to attempt this, particularly for unmeasured or unimagined factors that might affect the outcome. Although there is always a small chance that important factors are out of balance after random treatment assignment (i.e., failed randomization), the degree of imbalance is generally small, and gets smaller as the sample size gets larger.

In experiments, as opposed to observational studies, the assignment of levels of the explanatory variable to study units is under the control of the experimenter.

Experiments differ from **observational studies** in that in an experiment at least the main explanatory variables of interest are applied to the units of observation (most commonly subjects) *under the control of the experimenter*. Do not be fooled into thinking that just because a lot of careful work has gone into a study, it must therefore be an experiment. In contrast to experiments, in observational studies the subjects choose which treatment they receive. For example, if we perform magnetic resonance imaging (MRI) to study the effects of string instrument playing on the size of Broca's area of the brain, this is an observational study because the natural proclivities of the subjects determine which "treatment" level (control or string player) each subject has. The experimenter did not control this variable. The main advantage of an experiment is that the experimenter can randomly assign treatment, thus removing nearly all of the confounding. In the absence of confounding, a statistically significant change in the outcome provides good evidence for a causal effect of the explanatory variable(s) on the outcome. Many people consider internal validity to be not applicable to observational studies, but I think that in light of the availability of techniques to adjust for some confounding factors in observational studies, it is reasonable to discuss the internal validity of observational studies.

Internal validity is the ability to make causal conclusions. The huge advantage of randomized experiments over observational studies, is that causal conclusions are a natural outcome of the former, but difficult or impossible to justify in the latter.

Observational studies are always open to the possibility that the effects seen are due to confounding factors, and therefore have low internal validity. (As mentioned above, there are a variety of statistical techniques, beyond the scope of this book, which provide methods that attempt to "correct for" some of the confounding in observational studies.) As another example consider the effects of vitamin C on the common cold. A study that compares people who choose to take vitamin C versus those who choose not to will have many confounders and low internal validity. A

study that randomly assigns vitamin C versus a placebo will have good internal validity, and in the presence of a statistically significant difference in the frequency of colds, a causal effect can be claimed.

Note that confounding is a very specific term relating to the presence of a difference in the average level of any explanatory variable across the treatment groups. It should not be used according to its general English meaning of "something confusing".

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 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).

Besides lack of randomization and lack of blinding, omission of a control group is a cause of poor internal validity. A **control group** is a treatment group that represents some appropriate baseline treatment. It is hard to describe exactly what "appropriate baseline treatment" means, and this often requires knowledge of the subject area and good judgment. As an example, consider an experiment designed to test the effects of "memory classes" on short-term memory performance. If we have two treatment groups and are comparing subjects receiving two vs. five classes, and we find a "statistically significant difference", then we only know that adding three classes causes a memory improvement, but not if two is better than none. In some contexts this might not be important, but in others our critics will claim that there are important unanswered causal questions that we foolishly did not attempt to answer. You should always think about using a good control group, although it is not strictly necessary to always use one.

In a nutshell: It is only in blinded, randomized experiments that we can assure that the treatment precedes the outcome, and that there is little chance of confounding which would allow alternative explanations. It is these two conditions, along with statistically significant association, which allow a claim of causality.

8.2 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.

Among other things, 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. If there is a "gold standard", then your measure should have a high correlation with that test, 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. What is needed for a valid definition? First it should be recognized that circular logic applies here: as long as a measure is in some form that we would recognize as relating to obesity (as opposed to, say, smoking), then if it is a good predictor of health outcomes we can conclude that it is a good measure of obesity by definition. 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 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. Also these cutoff values and the names and number of categories have changed with time. 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 (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.

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. In addition, "factor analysis" suggests a single underlying factor (called "G" for general intelligence). 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.

8.3 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 Humanities upperclassmen at top northeastern universities who like free food and don't mind participating in experiments. Next you will want to use your judgment (and powers of persuasion) to consider ever expanding "spheres" of subjects who might be similar to your subjects. For example, you could widen the population to all northeastern students, then to all US students, then to all US young adults, etc. Finally you need to use your background knowledge and judgment to make your best arguments whether or not (or to what degree) you expect your findings to apply to these larger populations. If you cannot justify enlarging your population, then your study is likely to have little impact on scientific knowledge. If you enlarge too much, you may be severely criticized for over-generalization.

Three special forms of non-generalizability (poor external validity) are 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 subject who start a study do not complete it. Dropout can affect both internal and external validity, but the simplest form affecting external validity is when subjects who are too busy or less committed drop out only because of the length or burden of the experiment rather than in some way related to response to treatment. This type of dropout reduces the population to which generalization can be made, and in experiments 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. Here the 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.

8.4 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 that claim is actually false, i.e. the null hypothesis really is 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. Equally obvious after a little thought is the idea that we cannot be making a Type 1 error when we decide to retain the null hypothesis.

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! Another example of a statistical statement is "we are 95% confident that the true difference in mean outcome between the placebo and active treatments is between 6.5 and 8.7 seconds". Again, the exact meaning of this statement is a bit tricky, but understanding that is not critical for the current discussion (but see 6.2.7 for more details).

Due to the inherent uncertainties of nature we can never make definite, unqualified claims from our experiments. The best we can do is set certain limits on how often we will make certain false claims (but see the next section, on power, too). The conventional (but not logically necessary) limit on the rate of false positive results *out of all experiments in which the null hypothesis really is true* is 5%. The terms Type 1 error, false positive rate, and "alpha" (α) are basically synonyms for this limit.

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 6.2.2).

A special case of not maintaining Type 1 error is "data snooping". E.g., if you perform many different analyses of your data, each with a nominal Type 1 error rate of 5%, and then report just the one(s) with p-values less than 0.05, you are only fooling yourself and others if you think you have appropriately analyzed your experiment. As seen in the Section 13.3, 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.

8.5 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 $1 - \beta$.

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 12 for details on how to calculate and increase the power of an experiment.

The power of a planned experiment is the chance of getting a statistically significant result when a particular real treatment effect exists. Studying sufficient numbers of subjects is the most well known way to assure sufficient power.

In addition to sample size, the main (partially) controllable experimental characteristic that affects power is variability. If you can reduce variability, you can increase power. Therefore it is worthwhile to have a mnemonic device for helping you categorize and think about the **sources of variation**. One reasonable categorization is this:

- Measurement
- Environmental
- Treatment application
- Subject-to-subject

(If you are a New York baseball fan, you can remember the acronym METS.) It is not at all important to "correctly categorize" a particular source of variation. What is important is to be able to generate a list of the sources of variation in your (or someone else's) experiment so that you can think about whether you are able (and willing) to reduce each source of variation in order to improve the power of your experiment.

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. The term **precision** applies here, though technically precision is 1/variance. So a high precision implies a low variance (and thus standard deviation). It is worth knowing that a simple and usually a 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).

The trade-off between power and external validity also applies to treatment application variation. While some people include this in environmental variation, I think it is worth separating out because otherwise many people forget that it

is something that can be controlled in their experiment. Treatment application variability is differences in the quality or quantity of treatment among subjects assigned to the same (nominal) treatment. A simple example is when one treatment group gets, say 100 mg 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 (reduce external validity).

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.

For subject-to-subject variation there is a special way to improve power without reducing generalizability. This is the use of a **within-subjects design**, in which each subject receives two or more treatments. This is often an excellent way to improve power, although it is not applicable in all cases. See chapter 14 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 improve power without sacrificing generalizability.

The strength of your treatments (actually the difference in true outcomes between treatments) strongly affects power. Be sure that you are not 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.

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 factors whose effects are not of primary importance, but whose levels define more homogeneous groups called "blocks". In an ANOVA, for example, the block will be an additional factor beyond the primary treatment of interest, and inclusion of the block factor tends to improve power if the blocks are markedly more homogeneous than the whole. If the variability of the outcome (for each treatment group) is smaller than the variability ignoring the factor, then a good blocking factor was chosen. But because a wide variety of subjects with various levels of the blocking variable are all included in the study, generalizability is not sacrificed.

Examples of blocking factors include field in an agricultural experiment, age in many performance studies, and disease severity in medical studies. Blocking usually is performed when it is assumed that there is no differential effect of treatment across the blocks, i.e., no interaction (see Section 10.2). Ignoring an interaction when one is present tends to lead to misleading results, due to an incorrect structural model. Also, if there is an interaction between treatment and blocks, that usually becomes of primary interest.

A natural extension of blocking is some form of more complicated model with multiple **control variables** explicitly included in an appropriate mathematical form in the structural model. Continuous control variables are also called **covariates**.

	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 8.1: Simpson's paradox in medicine

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

8.6 Missing explanatory variables

Another threat to your experiment is not including important explanatory variables. For example, if the effect of a treatment is to raise the mean outcome in males and lower it in females, then not including gender as an explanatory variable (including its interaction with treatment) will give misleading results. (See chapters 10 and 11 for more on interaction.) In other cases, where there is no interaction, ignoring important explanatory variables decreases power rather than directly causing misleading results.

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 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 8.1.

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.

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 "variance vs.bias trade-off". The former artificially raises our Type 1 error if uncorrected, or lowers our power if corrected. The latter, in this context, can be considered to lower power when too many relatively unimportant explanatory variables are included.

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

8.7 Practicality and cost

Many attempts to improve an experiment are limited by cost and practicality. Finding ways to reduce threats to your experiment that are practical and cost-effective is an important part of experimental design. In addition, experimental science is usually guided by the KISS principle, which stands for Keep It Simple, Stupid. Many an experiment has been ruined because it was too complex to be carried out without confusion and mistakes.

8.8 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). You should consider all of these threats before performing your experiment and make appropriate adjustments as needed. 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 threat to your experiment, you will be in a better position to make choices that balance competing risk, and you will design a better experiment.

Chapter 9

Simple Linear Regression

An analysis appropriate for a quantitative outcome and a single quantitative explanatory variable.

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 variables are also possible values of the explanatory variables. We postulate 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.

So the structural model says that for each value of x the population mean of Y (over all of the subjects who have that particular value "x" for their explanatory variable) can be calculated using the simple linear expression $\beta_0 + \beta_1 x$. Of course we cannot make the calculation exactly, in practice, because the two parameters are unknown "secrets of nature". In practice, we make estimates of the parameters and substitute the estimates into the equation.

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. Technically it only says that the beta's must not be in a transformed form. It is OK to transform x or Y, and that allows many non-linear relationships to be represented on a new scale that makes the relationship linear.

The structural model 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). Of course, the value of σ^2 is an unknown parameter, and we can make an estimate of it from the data. The error model described so far includes not only the assumptions of "Normality" and "equal variance", but also the assumption of "fixed-x". 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, but usually 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 (usually a subject) 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.

The worst case of breaking the independent errors assumption in regression is when the observations are repeated measurement on the same experimental unit (subject).

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. (It is equivalent to say that all of the errors are Normally distributed.) Implicitly this indicates that the outcome should be a continuous quantitative variable. Practically speaking, real measurements are rounded and therefore some of their continuous nature is not available to us. If we round too much, the variable is essentially discrete and, with too much rounding, can no longer be approximated by the smooth Gaussian curve. Fortunately regression and ANOVA are both quite robust to deviations from the Normality assumption, and it is OK to use discrete or continuous outcomes that have at least a moderate number of different values, e.g., 10 or more. It can even be reasonable in some circumstances 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. Now ANOVA and regression give different answers because ANOVA makes no assumptions about the relationships of the three population means, but regression assumes a linear relationship. If the truth is linearity, the regression will have a bit more power

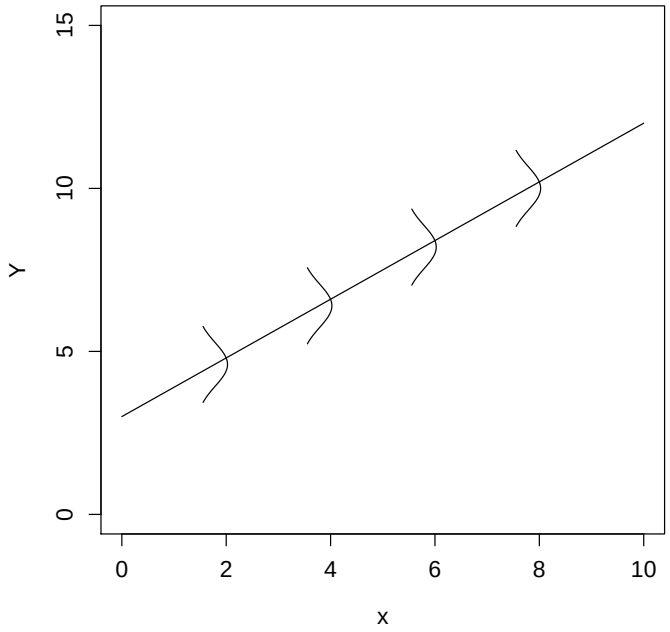


Figure 9.1: Mnemonic for the simple regression model.

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

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 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 effect on 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. There are data that span $x = 0$, or at least there are data points near $x = 0$. 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. See the section on interpretation below for more information.

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

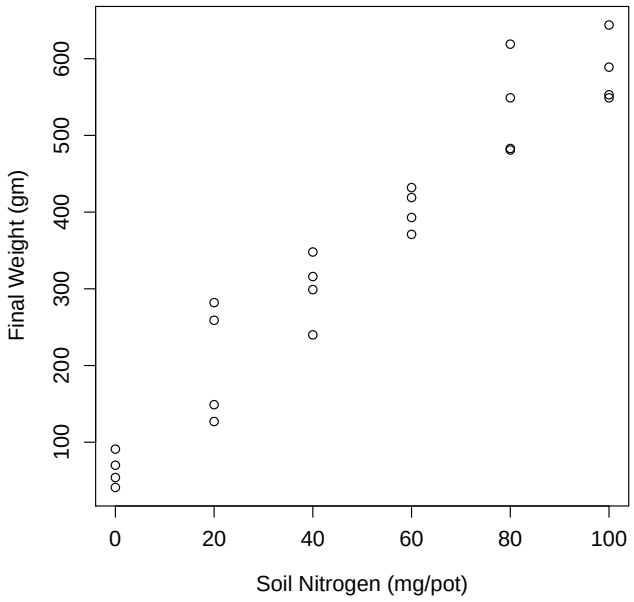


Figure 9.2: Scatterplot of corn data.

mg.

EDA, in the form of a scatterplot is shown in figure 9.2.

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 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 that linear regression is not a stupid 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. These estimates can be derived from either 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, that 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.

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 b_0 for the intercept and b_1 for the slope, the equation of the line is $y = b_0 + b_1 x$.

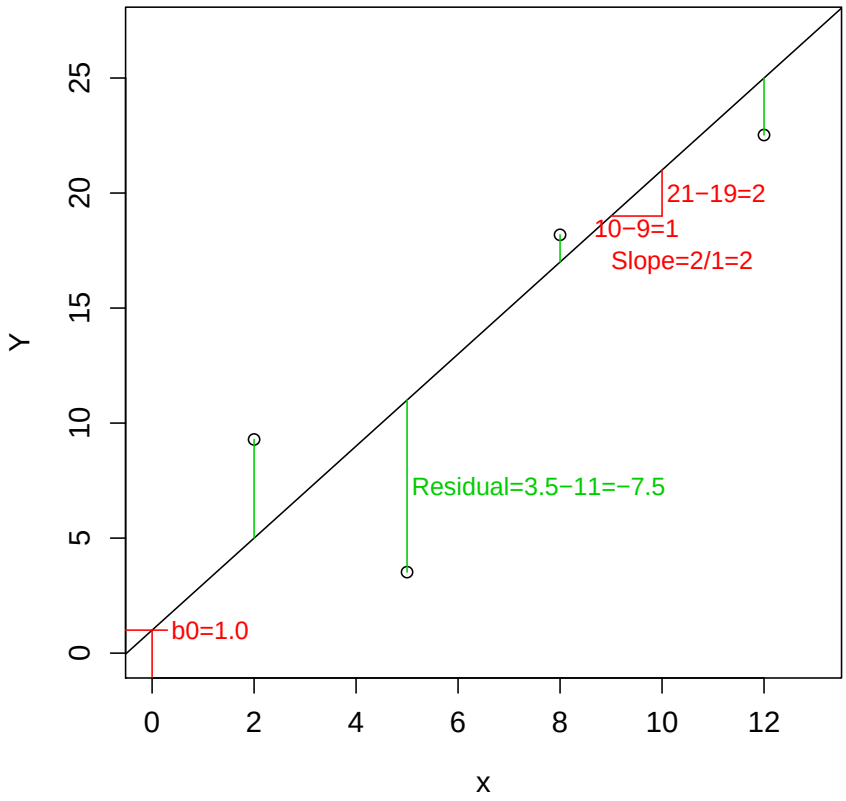


Figure 9.3: Least square principle.

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 b_0 and b_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 hat sub i", where "sub" means subscript, to indicate the fitted or predicted value of outcome y for subject i. (Some people also use the y_i' "y-prime sub i".) For subject i, who has explanatory variable x_i, the prediction is $\hat{y}_i = b_0 + b_1 x_i$ 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. It is interesting to note that 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. Instead we use calculus to find the values of b_0 and b_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.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

Also, the best estimate of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

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. Sometimes the symbols $\hat{\beta}_0$ and $\hat{\beta}_1$ are used instead of b_0 and b_1. Even though these symbols have Greek letters in them, the "hat" over the beta tells us that we are dealing with statistics, not parameters.

Here are the derivations of the coefficient estimates. SSR indicates sum of squared residuals, the quantity to minimize.

$$SSR = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \tag{9.1}$$
$$= \sum_{i=1}^{n}\left(y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2\right) \tag{9.2}$$

$$\begin{aligned}
\frac{\partial SSR}{\partial \beta_0} &= \sum_{i=1}^{n} (-2y_i + 2\beta_0 + 2\beta_1 x_i) && (9.3)\\
0 &= \sum_{i=1}^{n} \left(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i\right) && (9.4)\\
0 &= -n\bar{y} + n\hat{\beta}_0 + \hat{\beta}_1 n\bar{x} && (9.5)\\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} && (9.6)\\
\frac{\partial SSR}{\partial \beta_1} &= \sum_{i=1}^{n} \left(-2x_i y_i + 2\beta_0 x_i + 2\beta_1 x_i^2\right) && (9.7)\\
0 &= -\sum_{i=1}^{n} x_i y_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 && (9.8)\\
0 &= -\sum_{i=1}^{n} x_i y_i + (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 && (9.9)\\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} && (9.10)
\end{aligned}$$

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.

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

$$SE(b_0) = s\sqrt{\frac{\sum x^2}{n\sum(x^2) - (\sum x)^2}}$$

and

$$SE(b_1) = s\sqrt{\frac{n}{n\sum(x^2) - (\sum x)^2}}.$$

The basic regression output is shown in table 9.1 in a form similar to that produced by SPSS, but somewhat abbreviated. Specifically, "standardized coefficients" are not included.

In this table we see the number 94.821 to the right of the "(Constant)" label and under the labels "Unstandardized Coefficients" and "B". This is called the intercept estimate, estimated intercept coefficient, or estimated constant, and can

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
(Constant)	94.821	18.116	4.682	.000	47.251	122.391
Nitrogen added	5.269	.299	17.610	.000	4.684	5.889

Table 9.1: Regression results for the corn experiment.

be written as b_0, $\hat{\beta}_0$ or rarely B_0, but β_0 is incorrect, because the parameter value β_0 is a fixed, unknown "secret of nature". (Usually we should just say that b_0 equals 94.8 because the original data and most experimental data has at most 3 significant figures.)

The number 5.269 is the slope estimate, estimated slope coefficient, slope estimate for nitrogen added, or coefficient estimate for nitrogen added, and can be written as b_1, $\hat{\beta}_1$ or rarely B_1, but β_1 is incorrect. Sometimes symbols such as β_{nitrogen} or β_N for the parameter and b_{nitrogen} or b_N for the estimates will be used as better, more meaningful names, especially when dealing with multiple explanatory variables in multiple (as opposed to simple) regression.

To the right of the intercept and slope coefficients you will find their standard errors. As usual, standard errors are estimated standard deviations of the corresponding sampling distributions. For example, the SE of 0.299 for B_N gives an idea of the scale of the variability of the estimate B_N, which is 5.269 here but will vary with a standard deviation of approximately 0.299 around the true, unknown value of β_N if we repeat the whole experiment many times. The two t-statistics are calculated by all computer programs using the default null hypotheses of $H_0 : \beta_j = 0$ according to the general t-statistic formula

$$t_j = \frac{b_j - \text{hypothesized value of } \beta_j}{\text{SE}(b_j)}.$$

Then the computer uses the null sampling distributions of the t-statistics, i.e., the t-distribution with $n-2$ df, to compute the 2-sided p-values as the areas under the null sampling distribution more extreme (farther from zero) than the coefficient estimates for this experiment. SPSS reports this as "Sig.", and as usual gives the misleading output ".000" when the p-value is really "< 0.0005".

In simple regression the p-value for the null hypothesis $H_0 : \beta_1 = 0$ comes from the t-test for b_1. If applicable, a similar test is made for β_0.

SPSS also gives **Standardized Coefficients** (not shown here). These are the coefficient estimates obtained when both the explanatory and outcome variables are converted to so-called **Z-scores** by subtracting their means then dividing by their standard deviations. Under these conditions the intercept estimate is zero, so it is not shown. The main use of standardized coefficients is to allow comparison of the importance of different explanatory variables in multiple regression by showing the comparative effects of changing the explanatory variables by one standard deviation instead of by one unit of measurement. I rarely use standardized coefficients.

The output above also shows the "95% Confidence Interval for B" which is generated in SPSS by clicking "Confidence Intervals" under the "Statistics" button. In the given example we can say "we are 95% confident that β_N is between 4.68 and 5.89." More exactly, we know that using the method of construction of coefficient estimates and confidence intervals detailed above, and if the assumptions of regression are met, then each time we perform an experiment in this setting we will get a different confidence interval (center and width), and out of many confidence intervals 95% of them will contain β_N and 5% of them will not.

The confidence interval for 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. This also applies to β_0.

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 b_0 represents our estimate of the mean outcome when $x = 0$. Before making an interpretation of b_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 not be interpreted 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 b_0 (94.8) 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. So often it is better to use the CI for b_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. (It would be quite misleading to report the mean no-nitrogen plant weight as 94.821 gm because it gives a false impression of high precision.)

After interpreting the *estimate* of b_0 and it's 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. For 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 we might be unsure whether the rats were initially at a stable weight, it might make sense to test $H_0 : \beta_0 = 0$. If the null hypothesis is rejected then we conclude that it is not true that the weight gain is zero when the dose is zero (control group), so the initial weight was not a stable baseline weight.

Interpret the estimate, b_0, only if there are data near zero and setting the explanatory variable to zero makes scientific sense. The meaning of b_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 causa-

tion. Be sure to substitute association if you are looking at an observational study. 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 b_1. If it is positive then b_1 represents the increase in outcome caused by each one-unit increase in the explanatory variable. If b_1 is negative, then each one-unit increase in the explanatory variable is associated with a *fall* in outcome of magnitude equal to the absolute value of b_1.

A significant p-value indicates that we should reject the null hypothesis that $\beta_1 = 0$. 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. Particularly in the case of when we retain the null hypothesis, the interpretation of the CI for β_1 is better than simply relying on the general meaning of retain.

The interpretation of b_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 for a randomized experiment) a change in y. Rejection indicates that changes in x *cause* changes in y (assuming a randomized experiment).

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 (signed) difference $y_i - \hat{y}_i$. A positive 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.

A plot of all residuals on the y-axis vs. the predicted values on the x-axis, called a **residual vs. 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. As mentioned above, the fixed-x assumption cannot be checked with residual analysis (or any other data analysis). Serial correlation can be checked with special residual analyses, but is not visible on the two standard residual plots. The other types of correlated errors are not detected by standard residual analyses.

To analyze a residual vs. 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. (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 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. 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

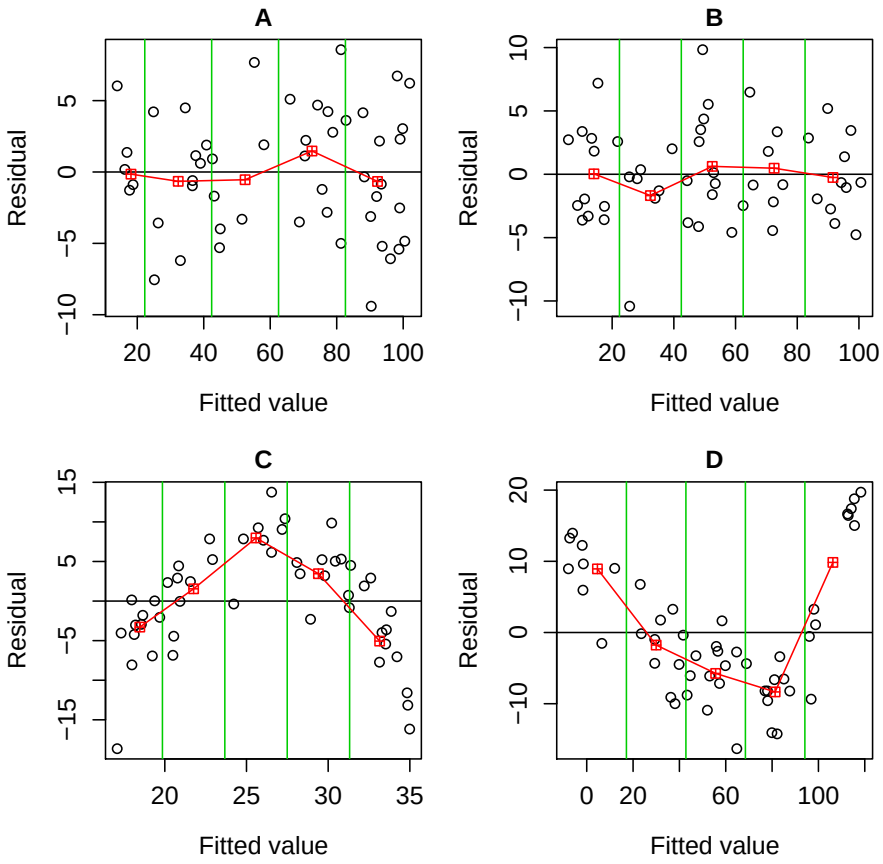


Figure 9.4: Sample residual vs. fit plots for testing linearity.

(vertical) mean of the residuals for each vertical stripe, then actually 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 that 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

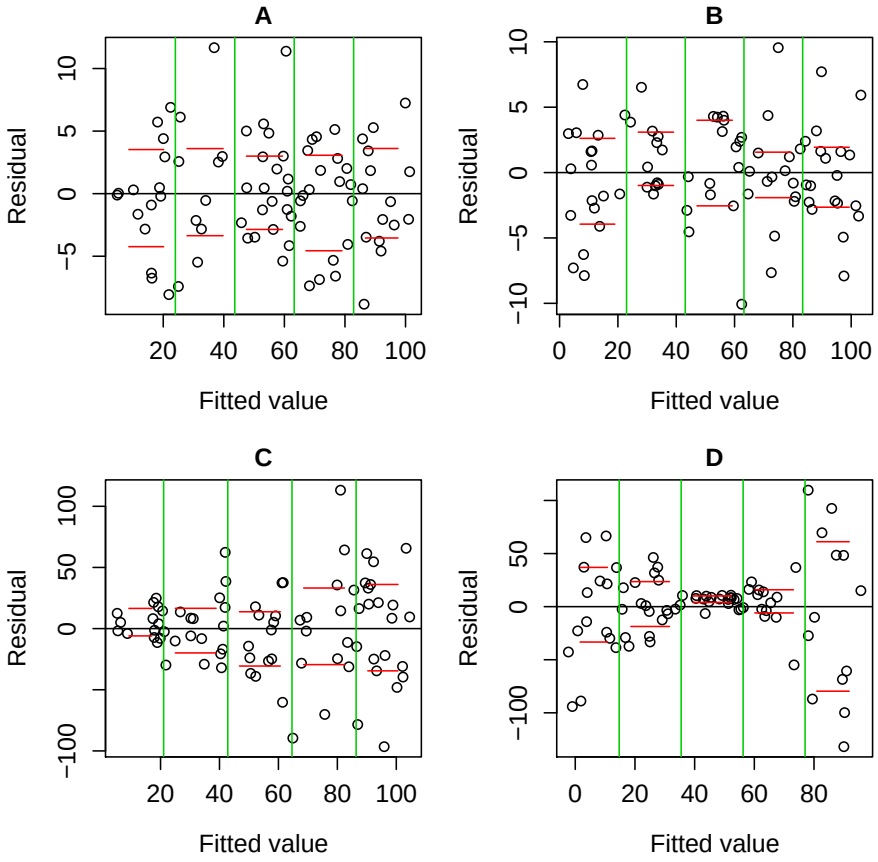


Figure 9.5: Sample residual vs. fit plots for testing equal variance.

data that produced plots C and D. With practice you will get better at reading 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 vs. 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 show 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 vs. 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 4.3.4). 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.

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.

If the linearity assumption is violated more than a fairly small amount, the regression loses its meaning. The most obvious way this happens is in the interpretation of b_1. We interpret b_1 as the change in the mean of Y for a one-unit

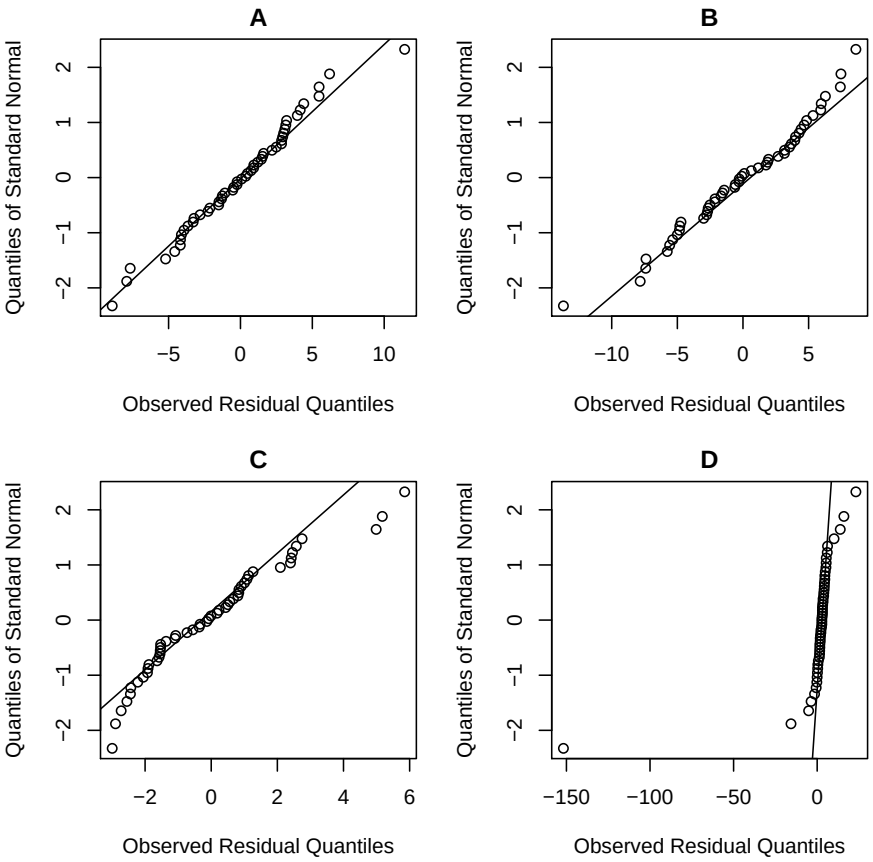


Figure 9.6: Sample QN plots of regression residuals.

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 fix it by transforming the x and/or y variables. Common transformations are log and square root. Alternatively it is common to *add* additional new explanatory variables in the form of a square, cube, etc. of the original x variable one at a time until the residual vs. fit plot shows linearity of the residuals. 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". The original way the data is measured usually has some degree of arbitrariness. Also, common measurements like pH for acidity, decibels for sound, and the Richter earthquake scale are all log scales. Often transformed values are transformed back to the original scale when results are reported (but the fact that the analysis was on a transformed scale must also be reported).

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. 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 the x measurement is of similar or larger magnitude to the variability of the 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.

Additional regression output is shown in table 9.2 which has what SPSS labels "Residual Statistics" on top and what it labels "Model Summary" on the bottom. The Residual Statistics summarize the predicted (fit) and residual values, as well as "standardized" values of these. The standardized values are transformed to Z-scores. You can use this table to detect possible outliers. If you know a lot about the outcome variable, use the unstandardized residual information to see if the minimum, maximum or standard deviation of the residuals is more extreme than you expected. If you are less familiar, standardized residuals bigger than about 3 in absolute value suggest that those points may be outliers.

The "Standard Error of the Estimate", s, is the best estimate of σ from our model (on the standard deviation scale). So it represents how far data will fall from the regression predictions on the scale of the outcome measurements. For the corn analysis, only about 5% of the data falls more than 2(49)=98 gm away from

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	84.8	611.7	348.2	183.8	24
Residual	-63.2	112.7	0.0	49.0	24
Std. Predicted Value	-1.43	1.43	0.00	1.00	24
Std. Residual	-1.26	2.25	0.00	0.978	24

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.966	0.934	0.931	50.061

Table 9.2: Additional regression results for the corn experiment.

the prediction line. Some programs report the mean squared error (MSE), which is the estimate of σ^2.

The R^2 value or **multiple correlation coefficient** is equal to the square of the simple correlation of x and y in simple regression, but not in multiple regression. In either case, R^2 can be interpreted as the fraction (or percent if multiplied by 100) of the total variation in the outcome that is "accounted for" by regressing the outcome on the explanatory variable.

A little math helps here. 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 "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

$$R^2 = \frac{\text{explained variance}}{\text{total variance}} = \frac{\text{var}(Y) - \text{MSE}}{\text{var}(Y)}.$$

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 the 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 $b_0 = 2$ and $b_1 = 1.5$, then we can say the best estimate of the log of the task time when no drug is given is 2 or that the 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 get multiplied by 4.48.

If the explanatory variable is common log of dose and the outcome is blood sugar level, and $b_0 = 85$ and $b_1 = 18$ 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.

Transformations of x or y to a different scale are very useful for fixing broken assumptions.

9.10 How to perform simple linear regression in SPSS

To perform simple linear regression in SPSS, select Analyze/Regression/Linear... from the menu. You will see the "Linear Regression" dialog box as shown in figure 9.7. Put the outcome in the "Dependent" box and the explanatory variable in the "Independent(s)" box. I recommend checking the "Confidence intervals" box for "Regression Coefficients" under the "Statistics..." button. Also click the "Plots..." button to get the "Linear Regression: Plots" dialog box shown in figure 9.8. From here under "Scatter" put "*ZRESID" into the "Y" box and "*ZPRED" into the "X" box to produce the residual vs. fit plot. Also check the "Normal probability plot" box.

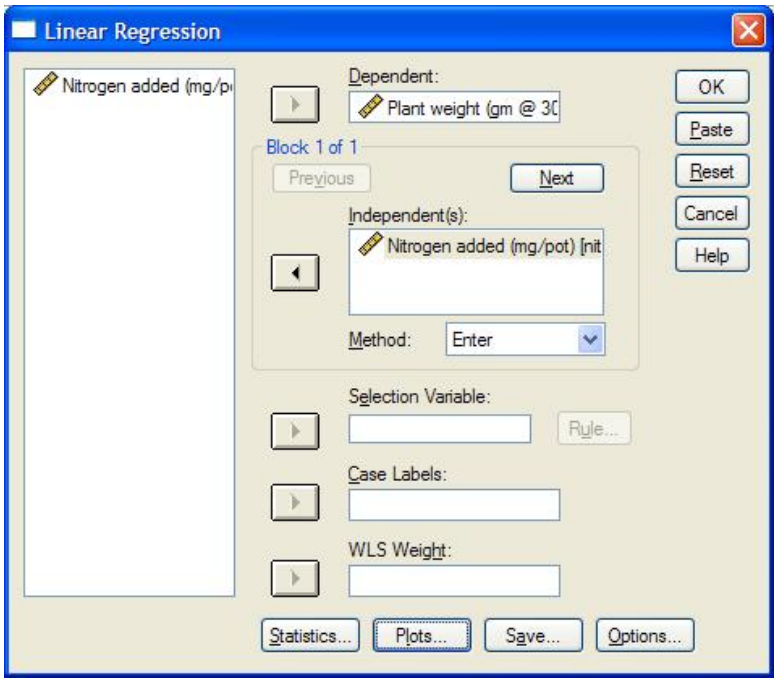


Figure 9.7: Linear regression dialog box.

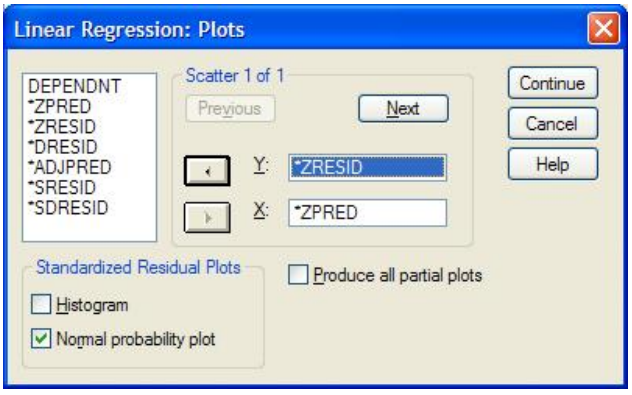


Figure 9.8: Linear regression plots dialog box.

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, b_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 vs. 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 some other continuous explanatory variable also has an effect on the outcome.

This chapter introduces several new important concepts including multiple regression, interaction, and use of indicator variables, then uses them to present a model appropriate for the setting of a quantitative outcome, and two explanatory variables, one categorical and one quantitative. 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 the 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 one β parameter corresponding to 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 fixed-x assumption which indicates that each of the quantitative explanatory variables is measured with little or no imprecision. All of the error model assumptions also apply. These assumptions state that for all subjects that have the same levels of all explanatory variables the outcome is Normally distributed around the true mean (or that the errors are Normally distributed with mean zero), and that the variance, σ^2, of the outcome around the true mean (or of the errors) is the same for every other set of values of the explanatory variables. And we assume that the errors are independent of each other.

Let's examine what the (no-interaction) 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 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.

From simplifying the structural model to specific cases we learn that the no-interaction 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, it also implies that 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. These relationships must be plausible in any given situation for the no-interaction multiple regression model to be considered. Some of these restrictions can be relaxed by including interactions (see below).

It is important to notice that the concept of changing the value of one explanatory variable while holding the others constant is meaningful in experiments, but generally not meaningful in observational studies. Therefore, interpretation of the slope coefficients in observational studies is fraught with difficulties and the potential for misrepresentation.

Multiple regression can occur in the experimental setting with two or more continuous explanatory variables, but it is perhaps more common to see one manipulated explanatory variable and one or more observed control variables. In that setting, inclusion of the control variables increases power, while the primary interpretation is focused on the experimental treatment variable. Control variables function in the same way as blocking variables (see 8.5) in that they affect the outcome but are not of primary interest, and for any specific value of the control variable, the variability in outcome associated with each value of the main experimental explanatory variable is reduced. Examples of control variables for many

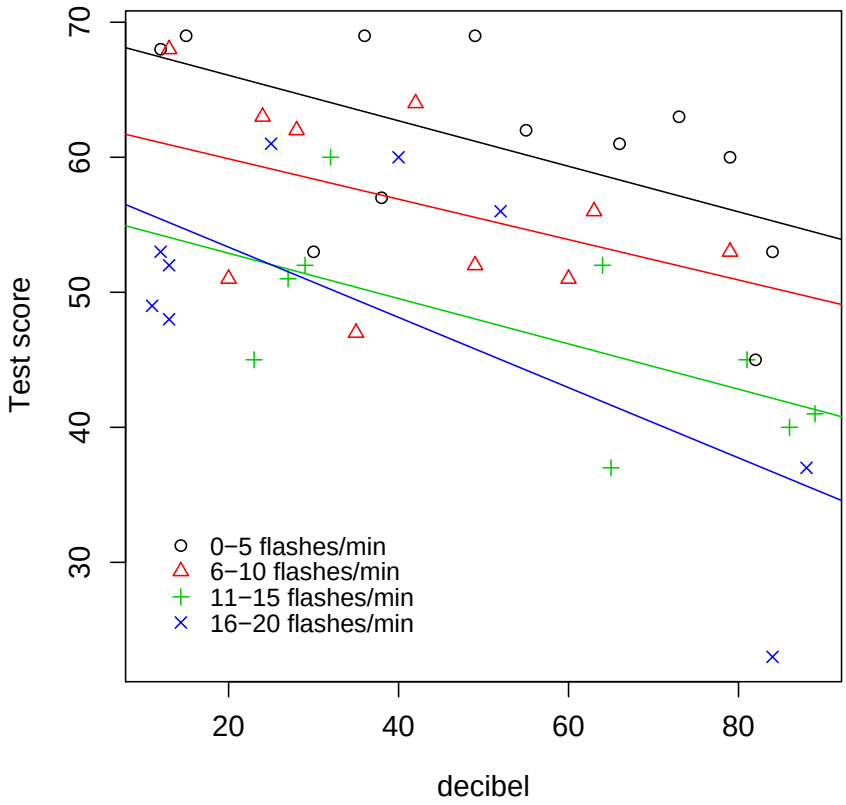


Figure 10.1: EDA for the distraction example.

psychological studies include things like ability (as determined by some auxiliary information) and age.

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.

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
(Constant)	74.688	3.260	22.910	<0.0005	68.083	81.294
db	-0.200	0.043	-4.695	<0.0005	-0.286	-0.114
freq	-1.118	0.208	-5.38	<0.0005	-1.539	-0.697

Table 10.1: Regression results for distraction experiment.

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.744	0.553	0.529	6.939

Table 10.2: Distraction experiment model summary.

Exploratory data analysis is difficult in the multiple regression setting because we need more than a two dimensional graph. For two explanatory variables and one outcome variable, programs like SPSS have a 3-dimensional plot (in SPSS try Graphs/ChartBuilder and choose the "Simple 3-D Scatter" template in the Scatter/Dot gallery; double click on the resulting plot and click the "Rotating 3-D Plot" toolbar button to make it "live" which allows you to rotate the plot so as to view it at different angles). For more than two explanatory variables, things get even more difficult. One approach that can help, but has some limitations, is to plot the outcome separately against each explanatory variable. For two explanatory variables, one variable can be temporarily demoted to categories (e.g., using the visual bander in SPSS), and then a plot like figure 10.1 is produced. Simple regression fit lines are added for each category. 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 (see below for details).

The multiple regression results are shown in tables 10.1 10.2, and 10.3.

	Sum of Squares	df	Mean Square	F	Sig.
Regression	22202.3	2	1101.1	22.9	<0.0005
Residual	1781.6	37	48.152		
Total	3983.9	39			

Table 10.3: Distraction experiment ANOVA.

Really important fact: There is an one-to-one relationship between the coefficients in the multiple regression output and the model equation for the mean of Y given the x's. There is exactly one term in the equation for each line in the coefficients table.

Here is an interpretation of the analysis of this experiment. (Computer reported numbers are rounded to a smaller, more reasonable number of decimal places – usually 3 significant figures.) 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 ($p<0.0005$, 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 ($p<0.0005$, 95%CI=[0.70,1.54]) at any fixed auditory distraction value. About 53% of the variability in test scores is accounted for by taking the values of the two distractions into account. (This comes from adjusted R^2.) The estimate of the standard deviation of test scores for any fixed combination of sound and light distraction is 6.9 points.

The validity of these conclusions is confirmed by the following assumption checks. The quantile-normal plot of the residuals confirms Normality of errors, the residual vs. fit plot confirms linearity and equal variance. (Subject 32 is a mild outlier with standardized residual of -2.3). The fixed-x assumption is met because the values of the distractions are precisely set by the experimenter. The independent errors assumption is met because separate subjects are used for each test, and the subjects were not allowed to collaborate.

It is also a good idea to further confirm linearity for each explanatory variable

with plots of each explanatory variable vs. the residuals. Those plots also look OK here.

One additional test should be performed before accepting the model and analysis discussed above for these data. We should test the "additivity" assumption which 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). The violation of this assumption usually takes the form of "interaction" which is the topic of the next section. The test needed is the p-value for the interaction term of a separate multiple regression model run with an interaction term.

One new interpretation is for the p-value of <0.0005 for the F statistic of 22.9 in the ANOVA table for the multiple regression. The 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 $H_0 : \beta_V = \beta_A = 0$ (or better yet, $H_0 : \beta_{visual} = \beta_{auditory} = 0$

Multiple regression is a direct extension of simple regression to multiple explanatory variables. Each new explanatory variable adds one term to the structural model.

10.2 Interaction

Interaction is a major concept in statistics that applies whenever there are two or more explanatory variables. Interaction is said to 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

Setting	x_S	x_L	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 10.4: Demonstration of the additivity of $E(Y) = 100 - 5x_S - 3x_L$.

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_S and x_L 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 $E(Y) = 100 - 5x_S - 3x_L$. Sample calculations are shown in Table 10.4. Line 1 shows the arbitrary starting values $x_S = 2$, $x_L = 4$. 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 $x_S = 2$, but change x_L 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_S from 2 to 3 and x_L 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" effects 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. Some other programs and textbooks use "A:B". Some computer programs can automatically create interaction variables, and some require you to create them. (You can always create them yourself, even if the program has a mechanism for automatic creation.) Peculiarly, SPSS has the automatic mechanism for some types of analyses but not others.

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 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

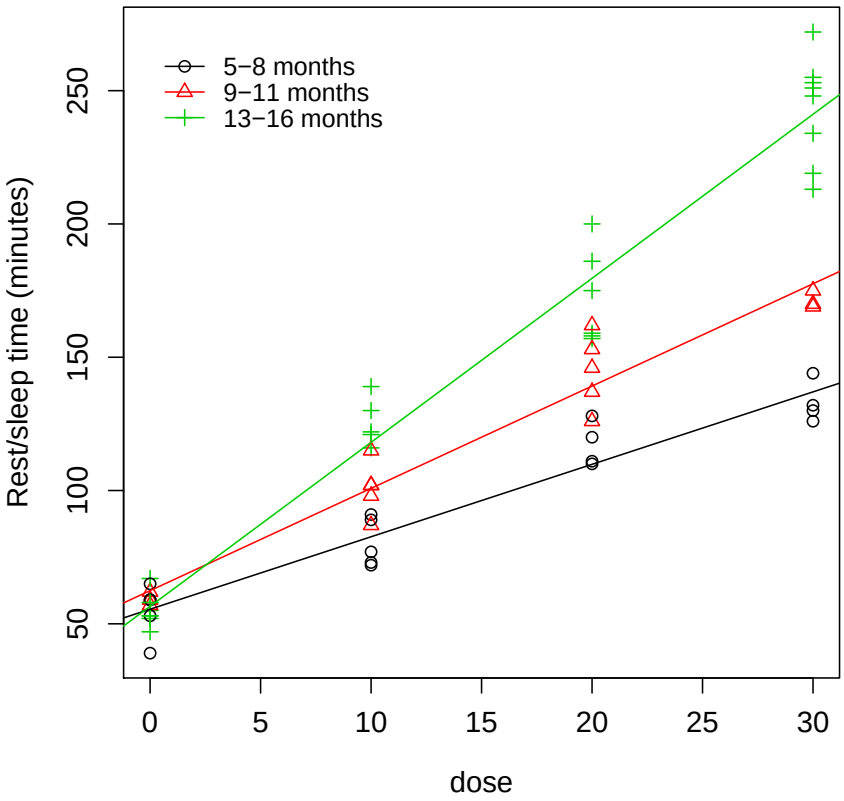


Figure 10.2: EDA for the lethargy example.

variable, 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 a 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 then making it reflect specific cases. Here, we use more meaningful variable names and parameter subscripts. Specifically, $\beta_{\text{d*a}}$ is the symbol for the single interaction parameter.

$$\begin{aligned} E(Y|\text{dose}, \text{age}) &= \beta_0 + \beta_{\text{dose}}\text{dose} + \beta_{\text{age}}\text{age} + \beta_{\text{d*a}}\text{dose} \cdot \text{age} \\ E(Y|\text{dose}, \text{age} = a) &= \beta_0 + \beta_{\text{dose}}\text{dose} + a\beta_{\text{age}} + a\beta_{\text{d*a}} \cdot \text{dose} \\ E(Y|\text{dose}, \text{age} = a) &= (\beta_0 + a\beta_{\text{age}}) + (\beta_{\text{dose}} + a\beta_{\text{d*a}})\text{dose} \end{aligned}$$

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. The conclusion is that the interaction model is one where the effects of any one-unit change in one explanatory variable while holding the other(s) constant is a change in the mean outcome, but the *size* (and maybe direction) of that change *depends* on the value(s) that the other explanatory variable(s) is/are set to.

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.

The multiple regression results are shown in tables 10.5 10.6, and 10.7.

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
(Constant)	48.995	5.493	8.919	<0.0005	37.991	59.999
Drug dose	0.398	0.282	1.410	0.164	-0.167	0.962
Rat age	0.759	0.500	1.517	0.135	-0.243	1.761
DoseAge IA	0.396	0.025	15.865	<0.0005	0.346	0.446

Table 10.5: Regression results for lethargy experiment.

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.992	0.985	0.984	7.883

Table 10.6: Lethargy experiment model summary.

	Sum of Squares	df	Mean Square	F	Sig.
Regression	222249	3	1101.1	22.868	<0.0005
Residual	3480	56	48.152		
Total	225729	59			

Table 10.7: Lethargy experiment ANOVA.

Here is an interpretation of the analysis of this experiment written in language suitable for an exam answer. 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<0.0005) 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. The estimate of the standard deviation of resting/sleeping time for any fixed combination of dose and age is 7.9 minutes.

The validity of these conclusions is confirmed by the following assumption checks. The quantile-normal plot of the residuals confirms Normality of errors, the residual vs. fit plot confirms linearity and equal variance. The fixed-x assumption is met because the dose is precisely set by the experimenter and age is precisely observed. The independent errors assumption is met because separate subjects are used for each test, and the subjects were not allowed to collaborate. Linearity is further confirmed by plots of each explanatory variable vs. the residuals.

Note that the p-value for the interaction line of the regression results (coefficient) table 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 must affect the outcome, so (except in certain special circumstances) *you should not interpret the p-values of the main effects if the interaction has a significant p-value.* 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.

If we want to write prediction equations with numbers instead of symbols, we should use Y' or $\hat{Y}$ on the left side, to indicate a "best estimate" rather than the

true but unknowable values represented by $E(Y)$ 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

$$\hat{Y} = 49.0 + 0.398(\text{dose}) + 0.76(12) + 0.396(\text{dose} \cdot 12)$$

which is $\hat{Y} = 58.1 + 5.15(\text{dose})$.

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 explained 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 to correspond to their "1" level.

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.)

This scheme for constructing new variables insures appropriate multiple regression analysis of categorical explanatory variables. As mentioned above, sometimes you need to create these variables explicitly, and sometime a statistical program will create them for you, either explicitly or silently.

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 know one or more explanatory variables are suspected to both affect the outcome and to define groups of subjects that are more homogeneous in terms of their outcomes for any treatment, then we know that we can use the blocking principle to increase power. 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 MSwithin 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. We do need to create the interaction variable in SPSS. The interaction p-value is 0.375 (not shown), so there is no evidence of a significant interaction (different slopes).

The results of the additive model (excluding the interaction) are shown in tables 10.8 10.9, and 10.10.

Of primary interest is the estimate of the benefit of using program B over program A, which is 10 points (t=2.40, p=0.020) with a 95% confidence interval of 2 to 18 points. Somewhat surprisingly the estimate of σ, which now refers to

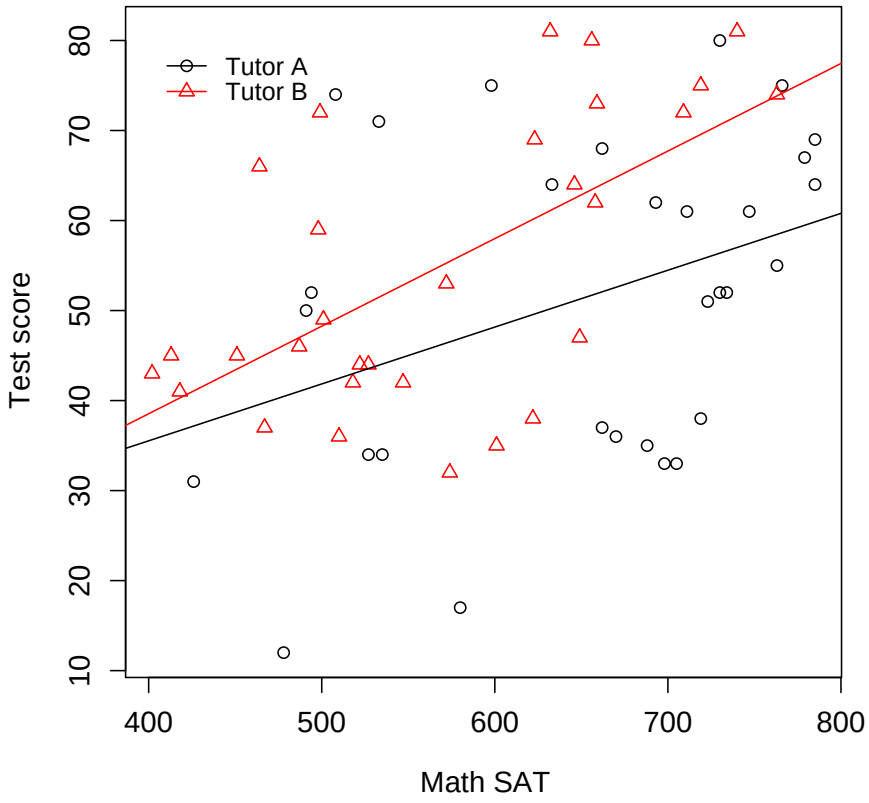


Figure 10.3: EDA for the math test / CAI example.

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
(Constant)	-0.270	12.698	-0.021	0.983	-25.696	25.157
ProgB	10.093	4.206	2.400	0.020	1.671	18.515
Math SAT	0.079	0.019	4.171	<0.0005	0.041	0.117

Table 10.8: Regression results for CAI experiment.

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.492	0.242	0.215	15.082

Table 10.9: CAI experiment model summary.

	Sum of Squares	df	Mean Square	F	Sig.
Regression	4138	2	2069.0	0.095	<0.0005
Residual	12966	57	227.5		
Total	17104	59			

Table 10.10: CAI experiment ANOVA.

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-squared = 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<0.0005). Every 10 additional math SAT points is associated with a 0.4 to 1.2 point rise in test score.

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.

The simplified model equations are shown here.

$$\begin{aligned} E(Y|\text{ProgB}, MSAT) &= \beta_0 + \beta_{\text{ProgB}}\text{ProgB} + \beta_{\text{MSAT}}\text{MSAT} \\ \text{Program A: } E(Y|\text{ProgB} = 0, MSAT) &= \beta_0 + \beta_{\text{MSAT}}\text{MSAT} \\ \text{Program B: } E(Y|\text{ProgB} = 1, MSAT) &= (\beta_0 + \beta_{\text{ProgB}}) + \beta_{\text{MSAT}}\text{MSAT} \end{aligned}$$

To be perfectly explicit, β_{MSAT} is the slope parameter for MSAT and β_{ProgB} is the parameter for the indicator variable ProgB. This parameter is technically a "slope", but really determines a difference in intercept for program A vs. program B.

For the analysis of the data shown here, the predictions are:

$$\begin{aligned} \hat{Y}(\text{ProgB}, MSAT) &= -0.27 + 10.09\text{ProgB} + 0.08\text{MSAT} \\ \text{Program A: } \hat{Y}(\text{ProgB} = 0, MSAT) &= -0.27 + 0.08\text{MSAT} \\ \text{Program B: } \hat{Y}(\text{ProgB} = 1, MSAT) &= 9.82 + 0.08\text{MSAT} \end{aligned}$$

Note that although 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 with no interaction 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. Often when we do an ANCOVA, we are "hoping" that there is no interaction because that indicates a more complicated reality, which is harder to explain. On the other hand sometimes a more complicated view of the world is just more interesting!

The multiple regression results shown in tables 10.11 and 10.12 refer to an experiment testing the effect of three different treatments (A, B and C) on a quantitative outcome, performance, which can range from 0 to 200 points, while controlling for skill variable S, which can range from 0 to 100 points. The data are available at Performance.dat. EDA showing the relationship between skill and

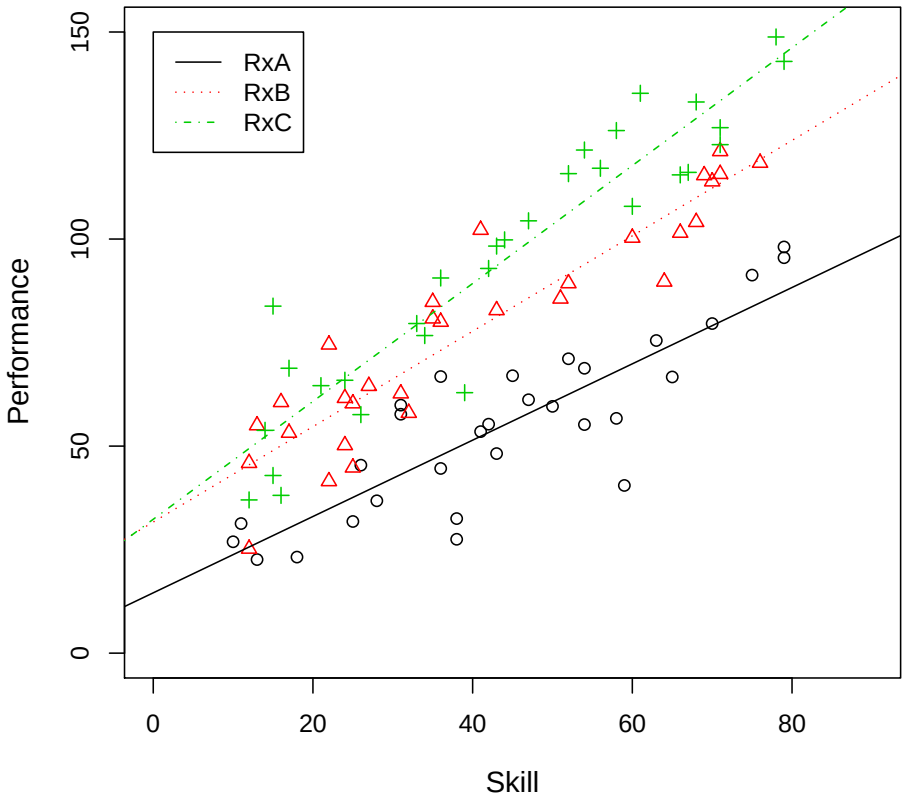


Figure 10.4: EDA for the performance ANCOVA example.

performance separately for each 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. Two interaction variables were created by multiplying S by RxB and S by RxC to create the single, two column interaction of Rx and S. Because it is logical and customary to consider the interaction between a continuous explanatory variable and a k level categorical explanatory variable, where $k > 2$, as a *single* interaction with $k-1$ degrees of freedom and $k-1$ lines in a coefficient table, we use a special procedure in SPSS (or other similar programs) to find a single p-value for the null hypothesis that model is additive vs. the alternative that there is an interaction. The SPSS procedure using the Linear Regression module is to use two "blocks" of independent variables, placing the main effects (here RxB, RxC, and Skill) into block 1, and the going to the "Next" block and placing the two interaction variables (here, RxB*S and RxC*S) into block 2. The optional statistic "R Squared Change" must also be selected.

The output that is labeled "Model Summary" (Table 10.11) and that is produced with the "R Squared Change" option is explained here. Lines are shown for two models. The first model is for the explanatory variables in block 1 only, i.e., the main effects, so it is for the additive ANCOVA model. The table shows that this model has an adjusted R^2 value of 0.863, and an estimate of 11.61 for the standard error of the estimate (σ). The second model adds the single 2 df interaction to produce the full interaction ANCOVA model with separate slopes for each treatment. The adjusted R^2 is larger suggesting that this is the better model. One good formal test of the necessity of using the more complex interaction model over just the additive model is the "F Change" test. Here the test has an F statistic of 6.36 with 2 and 84 df and a p-value of 0.003, so we reject the null hypothesis that the additive model is sufficient, and work only with the interaction model (model 2) for further interpretations. (The Model-1 "F Change test" is for the necessity of the additive model over an intercept-only model that predicts the intercept for all subjects.)

Using mnemonic labels for the parameters, the structural model that goes with this analysis (Model 2, with interaction) is

$$E(Y|\mathrm{Rx},\mathrm{S}) = \beta_0 + \beta_{\mathrm{RxB}}\mathrm{RxB} + \beta_{\mathrm{RxC}}\mathrm{RxC} + \beta_{\mathrm{S}}\mathrm{S} + \beta_{\mathrm{RxB*S}}\mathrm{RxB}\cdot\mathrm{S} + \beta_{\mathrm{RxC*S}}\mathrm{RxC}\cdot\mathrm{S}$$

You should be able to construct this equation directly from the names of the explanatory variables in Table 10.12.

Using Table 10.12, the parameter estimates are $\beta_0 = 14.56$, $\beta_{\mathrm{RxB}} = 17.10$, $\beta_{\mathrm{RxC}} = 17.77$, $\beta_{\mathrm{S}} = 0.92$, $\beta_{\mathrm{RxB*S}} = 0.23$, and $\beta_{\mathrm{RxC*S}} = 0.50$.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.931	0.867	0.863	11.61
2	0.941	0.885	0.878	10.95

	Change Statistics				
Model	R Square Change	F Change	df1	df2	Sig. F Change
1	0.867	187.57	3	86	<0.0005
2	0.017	6.36	2	84	0.003

Table 10.11: Model summary results for generic experiment.

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	3.22	3.39	0.95	0.344
	RxB	27.30	3.01	9.08	<0.0005
	RxC	39.81	3.00	13.28	<0.0005
	S	1.18	0.06	19.60	<0.0005
2	(Constant)	14.56	5.00	2.91	0.005
	RxB	17.10	6.63	2.58	0.012
	RxC	17.77	6.83	2.60	0.011
	S	0.92	0.10	8.82	<0.0005
	RxB*S	0.23	0.14	1.16	0.108
	RxC*S	0.50	0.14	3.55	0.001

Table 10.12: Regression results for generic experiment.

To understand this complicated model, we need to write simplified equations:

$$\begin{aligned}
\text{RxA: } E(Y|\text{Rx=A, S}) &= \beta_0 + \beta_{\text{S}}\text{S} \\
\text{RxB: } E(Y|\text{Rx=B, S}) &= (\beta_0 + \beta_{\text{RxB}}) + (\beta_{\text{S}} + \beta_{\text{RxB*S}})\text{S} \\
\text{RxC: } E(Y|\text{Rx=C, S}) &= (\beta_0 + \beta_{\text{RxC}}) + (\beta_{\text{S}} + \beta_{\text{RxC*S}})\text{S}
\end{aligned}$$

Remember that 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. (It is often 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 the interpretation of 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. Notice that the interpretation can be thought of a description of the EDA plot which uses ANCOVA results to specify which observations one might make about the plot that are statistically verifiable.

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.

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 14.6 points (a 95% CI would be more meaningful). If it were scientifically interesting, we could also say that this value of 14.6 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, df=84, p=0.012, 0.011). The estimates are 17.1 and 17.8 points higher for B and C respectively compared to A (and again, CIs would be useful here).

We can also say that there is a statistically significant effect of skill on performance for subjects given treatment A (t=8.82, p< 0.0005). The best estimate is 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 for 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.

Additional testing, using methods we have not learned, can be performed to show that performance is better for treatments B and C than treatment A at all observed levels of skill.

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).

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.

10.5 Do it in SPSS

To create $k-1$ indicator variables from a k-level categorical variable in SPSS, run Transform/RecodeIntoDifferentVariables, as shown in figure 5.16, $k-1$ times. Each new variable name should match one of the non-baseline levels of the categorical variable. Each time you will set the old and new values (figure 5.17) to convert the named value to 1 and "all other values" to 0.

To create $k-1$ interaction variables for the interaction between a k-level categorical variable and a quantitative variable, use Transform/Compute $k-1$ times. Each new variable name should specify what two variables are being multiplied. A label with a "*", ":" or the word "interaction" or abbreviation "I/A" along with the categorical level and quantitative name is a really good idea. The "Numeric Expression" (see figure 5.15) is just the product of the two variables, where "*" means multiply.

To perform multiple regression in any form, use the Analyze/Regression/Linear menu item (see figure 9.7), and put the outcome in the Dependent box. Then put all of the main effect explanatory variables in the Independent(s) box. Do **not** use the original categorical variable – use only the $k-1$ corresponding indicator variables. If you want to model non-parallel lines, add the interaction variables as a second block of independent variables, and turn on the "R Square Change" option under "Statistics". As in simple regression, add the option for CI's for the estimates, and graphs of the normal probability plot and residual vs. fit plot. Generally, if the "F change test" for the interaction is greater than 0.05, use "Model 1", the additive model, for interpretations. If it is ≤ 0.05, use "Model 2", the interaction model.

Chapter 11

Two-Way ANOVA

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

If an experiment has a quantitative outcome and two categorical explanatory variables that are defined in such a way that 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, then the most common analysis method is **two-way ANOVA**. Because there are two different explanatory variables the effects on the outcome of a change in one variable may either not depend on the level of the other variable (additive model) or it may depend on the level of the other variable (interaction model). 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 x m ANOVA". ANOVA with more that two explanatory variables is often called **multi-way ANOVA**. If a quantitative explanatory variable is also included, that variable is usually called a **covariate**.

In two-way ANOVA, the error model is the usual one of Normal distribution with equal variance for all subjects that share levels of both (all) of the explanatory variables. Again, we will call that common variance σ^2. And 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.

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, and a plot of those μ's could show any arbitrary pattern.

In contrast, the no-interaction (additive) model does have a restriction on the population means of the outcomes. For the no-interaction model we can think of the mean restrictions as saying that the effect on the outcome of any specific level change for one explanatory variable is the same for every fixed setting of the other explanatory variable. This is called an **additive model**. Using the notation of the previous paragraph, the mathematical form of the additive model is $\mu_{ac} - \mu_{bc} = \mu_{ad} - \mu_{bd}$ for any valid levels a, b, c, and d. (Also, $\mu_{ab} - \mu_{ac} = \mu_{db} - \mu_{dc}$.)

A more intuitive presentation of the additive model is a plot of the population means as shown in figure 11.1. The same information is shown in both panels. In each the 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 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.

Note that the concept of interaction vs. an additive model is the same for ANCOVA or a two-way ANOVA. In the additive model the effects of a change in

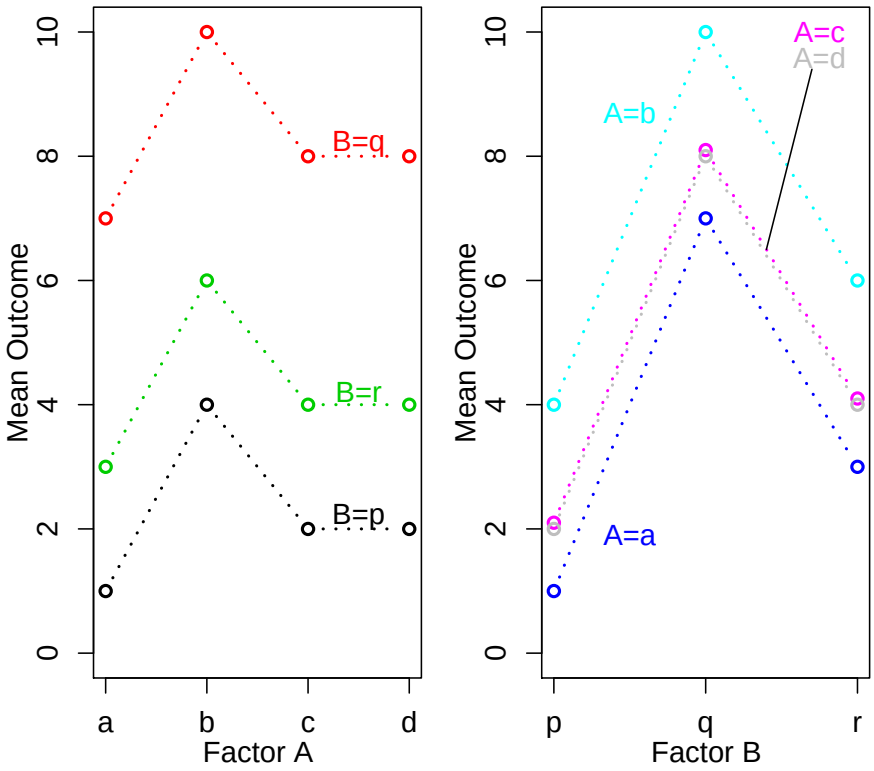


Figure 11.1: Population means for a no-interaction two-way ANOVA example.

one explanatory variable on the outcome does *not* depend on the value or level of the other explanatory variable, and the effect of a change in an explanatory variable can be described while not stating the (fixed) level of the other explanatory variable. And for the models underlying both analyses, if an interaction is present, the effects on the outcome of changing one explanatory variable *depends* on the specific value or level of the other explanatory variable. Also, the lines representing the mean of y at all values of quantitative variable x (in some practical interval) for each particular level of the categorical variable are all parallel (additive model) or not all parallel (interaction) in ANCOVA. In two-way ANOVA the order of the levels of the categorical variable represented on the x-axis is arbitrary and there is nothing between the levels, but nevertheless, if lines are drawn to aid the eye, these lines are all parallel if there is no interaction, and not all parallel if there is an interaction.

The two possible means models for two-way ANOVA are the additive model and the interaction model. The additive model assumes that the effects on the outcome of a particular level change for one explanatory variable does not depend on the level of the other explanatory variable. If an interaction model is needed, then the effects of a particular level change for 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 11.1, 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.

11.1 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 best be considered to be 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.

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 testing or in possible "drift" in the accuracy of the noise measurement over the duration of th 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

		TYPE		
		Standard	Octel	Total
SIZE	small	6	6	12
	medium	6	6	12
	large	6	6	12
Total		18	18	36

Table 11.1: Cross-tabulation for car noise example.

interaction term is not statistically significant.

Some useful EDA is shown in table 11.1 and figures 11.2 and 11.3. The cross-tabulation lets us see that each **cell** of the experiment, i.e., each set of outcomes that correspond to a given set of levels of the explanatory variables, has six subjects (cars tested). This situation where there are the same number of subjects in all cells is called a **balanced design**. One of the key features of this experiment which tells us that it is OK to use the assumption of independent errors is that a different subject (car) is used for each test (row in the data). 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.

The boxplots show that the small and medium sized cars have more noise than the large cars (although this may not be a good generalization, assuming that only one car model was testing in each size class). 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.

The error bar plot shows mean plus or minus 2 SE. A good alternative, which looks very similar, is to show the 95% CI around each mean. For this plot, the standard deviations and sample sizes for each of the six groups are separately used to construct the error bars, but this is less than ideal if the equal variance assumption is met, in which case a pooled standard deviation is better. In this example, the best approach would be to use one pooled standard deviation for

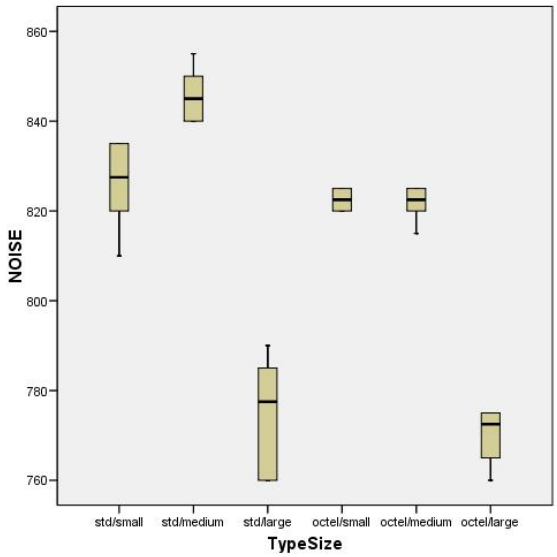


Figure 11.2: Side-by-side boxplots for car noise example.

each filter type.

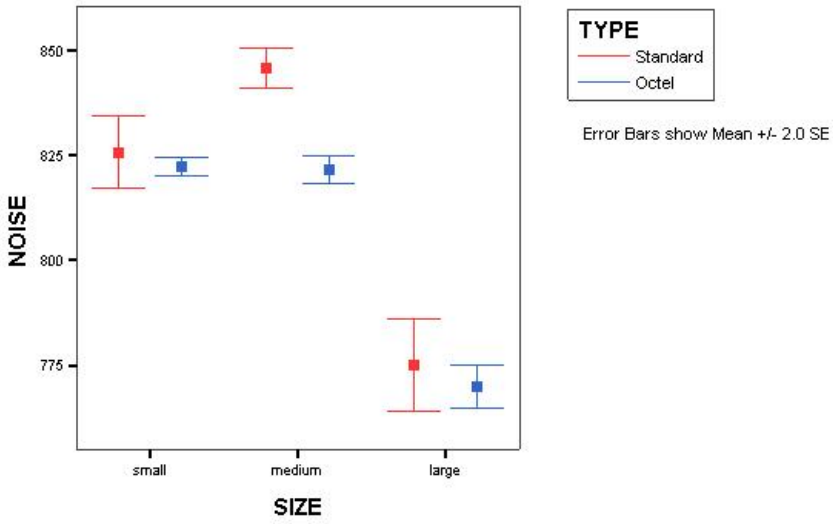


Figure 11.3: Error bar plot for car noise example.

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	27912	5	5582	85.3	<0.0005
SIZE	26051	2	13026	199.1	<0.0005
TYPE	1056	1	1056	16.1	<0.0005
SIZE*TYPE	804	2	402	6.1	<0.0005
Error	1962	30	65		
Corrected Total	29874	35			

Table 11.2: ANOVA for the car noise experiment.

11.2 Interpreting the two-way ANOVA results

The results of a two-way ANOVA of the car noise example are shown in tables 11.2 and 11.3. The ANOVA table is structured just like the one-way ANOVA table. The SS column represents the sum of squared deviations for each of several different ways of choosing which deviations to look at, and these are labeled "Source (of Variation)" for reasons that are discussed more fully below. Each SS has a corresponding df (degrees of freedom) which is a measure of the number of independent pieces of information present in the deviations that are used to compute the corresponding SS (see section 4.6). And each MS is the SS divided by the df for that line. Each MS is a variance estimate or a variance-like quantity, and as such its units are the squares of the outcome units.

Each F-statistic is the ratio of two MS values. For the between-groups ANOVA discussed in this chapter, the denominators are all MS_{error} (MSE) which 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 the null hypothesis is true. Just like in one-way ANOVA, a component of SS_{error} is computed for each treatment cell as deviations of individual subject outcomes from the sample mean of all subjects in that cell; the component df for each cell is $n_{ij} - 1$ (where n_{ij} is the number of subjects exposed to level i of one explanatory variable and level j of the other); and the SS and df are computed by summing over all cells.

Each F-statistic is compared against it's 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.

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.

The ANOVA table analyzes the total variation of the outcome in the experiment by decomposing the SS (and df) into components that add to the total (which only works because the components are what is called orthogonal). One decomposition visible in the ANOVA table is that the SS and df add up for "Corrected model" + "Error" = "Corrected Total". When interaction is included in the model, this decomposition is equivalent to a one-way ANOVA where all of the ab cells in a table with a levels of one factor and b levels of the other factor are treated as ab levels of a single factor. In that case the values for "Corrected Model" correspond to the "between-group" values of a one-way ANOVA, and the values for "Error" correspond to the "within-group" values. The null hypothesis for the "Corrected Model" F-statistic is that all ab population cell means are equal, and the deviations involved in the sum of squares are the deviations of the cell sample means from the overall mean. Note that this has $ab - 1$ df. The "Error" deviations are deviations of the individual subject outcome values from the group means. This has $N - ab$ df. In our car noise example $a = 2$ filter types, $b = 3$ sizes, and $N = 36$ total noise tests run.

SPSS gives two useless lines in the ANOVA table, which are not shown in figure 11.2. These are "Intercept" and "Total". Note that most computer programs report what SPSS calls the "Corrected Total" as the "Total".

The rest of the ANOVA table is a decomposition of the "Corrected Model" into main effects for size and type, as well as the interaction of size and type (size*type). You can see that the SS and df add up such that "Corrected Model" = "size" + "type" + "size*type". This decomposition can be thought of as saying that the deviation of the cell means from the overall mean is equal to the size deviations plus the type deviations plus any deviations from the additive model in the form of interaction.

In the presence of an interaction, the p-value for the interaction is most im-

portant and the main effects p-values are generally ignored if the interaction is significant. This is mainly because if the interaction is significant, then some changes in *both* explanatory variables must have an effect on the outcome, regardless of the main effect p-values. The null hypothesis for the interaction F-statistic is that there is an additive relationship between the two explanatory variables in their effects on the outcome. 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 11.4.

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. More detailed explanations require "simple effects testing", see chapter 13.

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.

Examination of the plots or the Marginal Means table suggests (but does not prove) that the important difference is that the noise level is higher for the standard

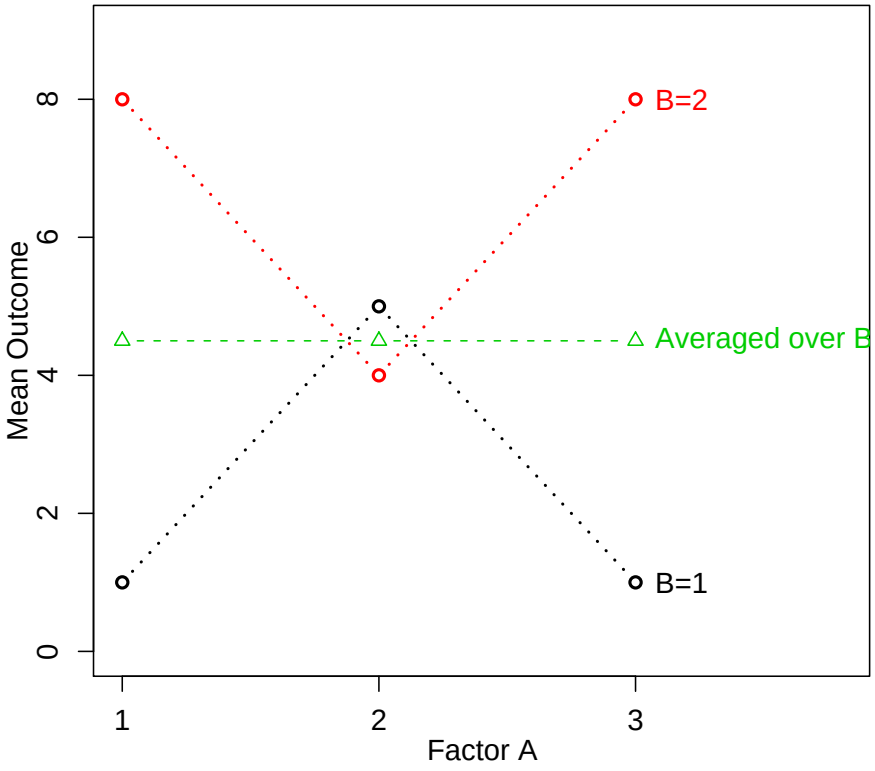


Figure 11.4: Significant interaction with misleading non-significant main effect of factor A.

				95% Confidence Interval	
SIZE	TYPE	Mean	Std. Error	Lower Bound	Upper Bound
small	Standard	825.83	3.30	819.09	832.58
	Octel	822.50	3.30	815.76	829.24
medium	Standard	845.83	3.30	839.09	852.58
	Octel	821.67	3.30	814.92	828.41
large	Standard	775.00	3.30	768.26	781.74
	Octel	770.00	3.30	763.26	776.74

Table 11.3: Estimated Marginal Means for the car noise experiment.

filter than the Octel filter for the medium sized car, but the filters have equivalent effects for the small and large cars.

If the interaction p-value is not statistically significant, then in most situations most analysts would re-run the ANOVA without the interaction, i.e., as a main effects only, additive model. The interpretation of main effects F-statistics in a non-interaction two-way ANOVA is easy. 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.) If there are only two levels, then we can and should immediately report which one is "better" by looking at the sample means. If there are more than two levels, we can only say that there are some differences in mean outcome among the levels, but we need to do additional analysis in the form of "contrast testing" as shown in chapter 13 to determine which levels are statistically significantly different.

Inference for the two-way ANOVA table 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 the same at every fixed level of the other factor.

It is worth noting that a transformation, such as a log transformation of the outcome, would not correct the unequal variance of the outcome across the groups defined by treatment combinations for this example (see figure 11.2). A log transformation corrects unequal variance only in the case where the variance is larger for groups with larger outcome means, which is not the case here. Therefore,

other than using much more complicated analysis methods which flexibly model changes in variance, the best solution to the problem of unequal variance in this example, is to use the "Keppel" correction which roughly corrects for moderate degrees if violation of the equal variance assumption by substituting $\alpha/2$ for α. For this problem, we still reject the null hypothesis of an additive model when we compare the p-value to 0.025 instead of 0.05, so the correction does not change our conclusion.

Figure 11.5 shows the 3 by 3 residual plot produced in SPSS by checking the Option "Residual plot". The middle panel of the bottom row shows the usual residual vs. 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. Verifying that the means for all of the vertical bands are at zero is a check that the mean model is OK. For two-way ANOVA this comes down to checking that dropping the interaction term was a reasonable thing to do. In other words, if a no-interaction model shows a pattern to the means, the interaction is probably needed. This default plot is poorly designed, and does not allow checking Normality. I prefer the somewhat more tedious approach of using the Save feature in SPSS to save predicted and residual values, then using these to make the usual full size residual vs. fit plot, plus a QN plot of the residuals to check for Normality.

Residual checking for two-way ANOVA is very similar to regression and one-way ANOVA.

11.3 Math and gender example

The data in mathGender.dat are from an observational study carried out to investigate the relationship between the ACT Math Usage Test and the explanatory variables gender (1=female, 2=male) and level of mathematics coursework taken (1=algebra only, 2=algebra+geometry, 3=through calculus) 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. An analysis of these data of the type discussed in this chapter can be called a 3x2 ("three by two") ANOVA because those are the numbers of levels of the two categorical explanatory variables.

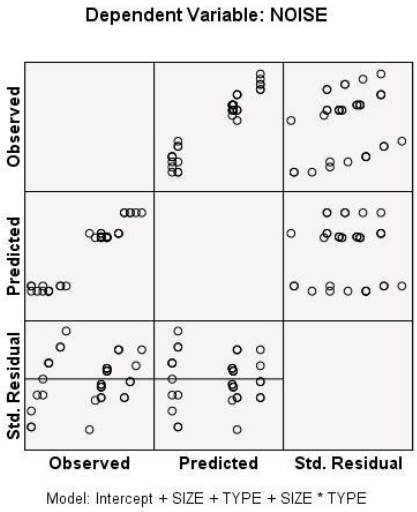


Figure 11.5: Residual plots for car noise example.

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, which takes the clustering into schools into account. For the analysis below, we assume that student are randomly sampled throughout the country so that including two students from the same school would only be a rare coincidence.

This is an observational study, so our conclusions will be described in terms of association, not causation. Neither gender nor coursework was randomized to different students.

The cross-tabulation of the explanatory variables is shown in table 11.4. As opposed to the previous example, this is not a balanced ANOVA, because it has unequal cell sizes.

Further EDA shows that each of the six cells has roughly the same variance for the test scores, and none of the cells shows test score skewness or kurtosis suggestive of non-Normality.

		Gender		
		Female	Male	Total
Coursework	algebra	82	48	130
	to geometry	387	223	610
	to calculus	54	67	121
Total		523	338	861

Table 11.4: Cross-tabulation for the math and gender example.

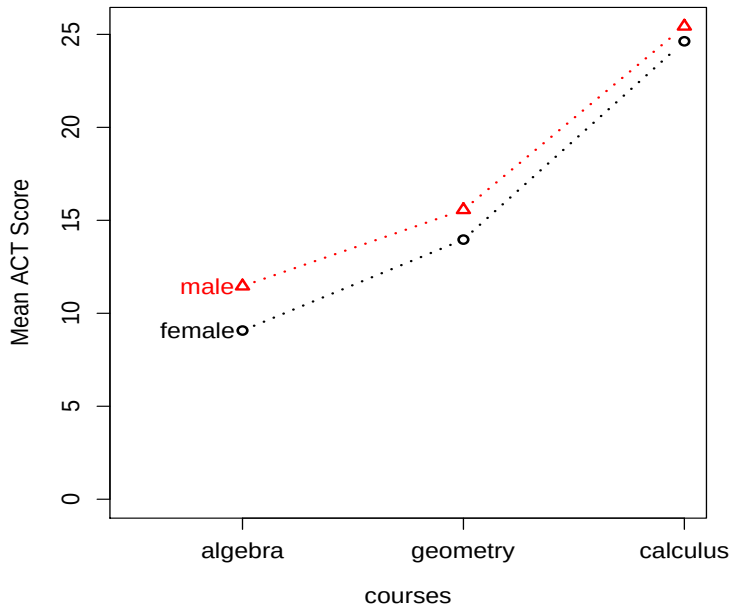


Figure 11.6: Cell means for the math and gender example.

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16172.8	5	3234.6	132.5	<0.0005
courses	14479.5	2	7239.8	296.5	<0.0005
gender	311.9	1	311.9	12.8	<0.0005
courses*gender	37.6	2	18.8	0.8	0.463
Error	20876.8	855	24.4		
Corrected Total	37049.7	860	43.1		

Table 11.5: ANOVA with interaction for the math and gender example.

A profile plot of the cell means is shown in figure 11.6. The first impression is that students who take more courses have higher scores, males have slightly higher scores than females, and perhaps the gender difference is smaller for students who take more courses.

The two-way ANOVA with interaction is shown in table 11.5.

The deviations used in the sums of squared deviations (SS) in a two-way ANOVA with interaction are just a bit more complicated than in one-way ANOVA. The main effects deviations are calculated as in one-way interaction, just ignoring the other factor. Then the interaction SS is calculated by using the main effects to construct the best "parallel pattern" means and then looking at the deviations of the actual cell means from the best "parallel pattern means".

The interaction line of the table (courses*gender) 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 gender. The formula for interaction df is $(k-1)(m-1)$ for any k by m ANOVA.

As a minor point, note that the MS is given for the "Corrected Total" line. Some programs give this value, which equals the variance of all of the outcomes ignoring the explanatory variables. The "Corrected Total" line adds up for both the SS and df columns but not for the MS column, to either "Corrected Model" + "Error" or to all of the main effects plus interactions plus the Error.

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16135.2	3	5378.4	220.4	<0.0005
courses	14704.7	2	7352.3	301.3	<0.0005
gender	516.6	1	516.6	21.2	<0.0005
Error	20914.5	857	24.4		
Corrected Total	37049.7	860			

Table 11.6: ANOVA without interaction for the math and gender example.

The main point of this ANOVA table is that the interaction between the explanatory variables gender and courses is not significant (F=0.8, 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 genders, and gender 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.

The ANOVA table for a two-way ANOVA without interaction is shown in table 11.6.

Our conclusion, using a significance level of $\alpha = 0.05$ is that both courses and gender affect test score. Specifically, because gender has only two levels (1 df), we can directly check the Estimated Means table (table 11.7) to see that males have a higher mean. Then we can conclude based on the small p-value that being male is associated with a higher math ACT score compared to females, for each level of courses. This is not in conflict with the observation that some females are better than most males, because it is only a statement about means. In fact the estimated means table tells us that the mean difference is 2.6 while the ANOVA table tells us that the standard deviation in any group is approximately 5 (square root of 24.4), so the overlap between males and females is quite large. Also, this kind of study certainly cannot distinguish differences due to biological factors from those due to social or other factors.

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 males and females 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 13.

courses	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
algebra	10.16	0.44	9.31	11.02
to geometry	14.76	0.20	14.36	15.17
to calculus	14.99	0.45	24.11	25.87

gender	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
female	14.84	0.26	15.32	16.36
male	17.44	0.30	16.86	18.02

Table 11.7: Estimated means for the math and gender example.

We can also note that the residual (within-group) variance is 24.4, so our estimate of the population standard deviation for each group is $\sqrt{24.4} = 4.9$. Therefore about 95% of test scores for any gender and level of coursework are within 9.8 points of that group's mean score.

11.4 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 11.7, 11.8, 11.9, 11.10, and 11.11 which include plus or minus 2 SE error bars (roughly 95% CI for the population means).

Table 11.8 shows the p-values from two-way ANOVA's of these five cases.

In case A you can see that it takes very little "wiggle", certainly less than the size of the error bars, to get the lines to be parallel, so an additive model should be OK, and indeed the interaction p-value is 0.802. We should re-fit a model without an interaction term. We see that as we change sound levels (move left or right), the mean outcome (y-axis value) does not change much, so sound level does not affect the outcome and we get a non-significant p-value (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.

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 11.8: P-values for various light/sound experiment cases.

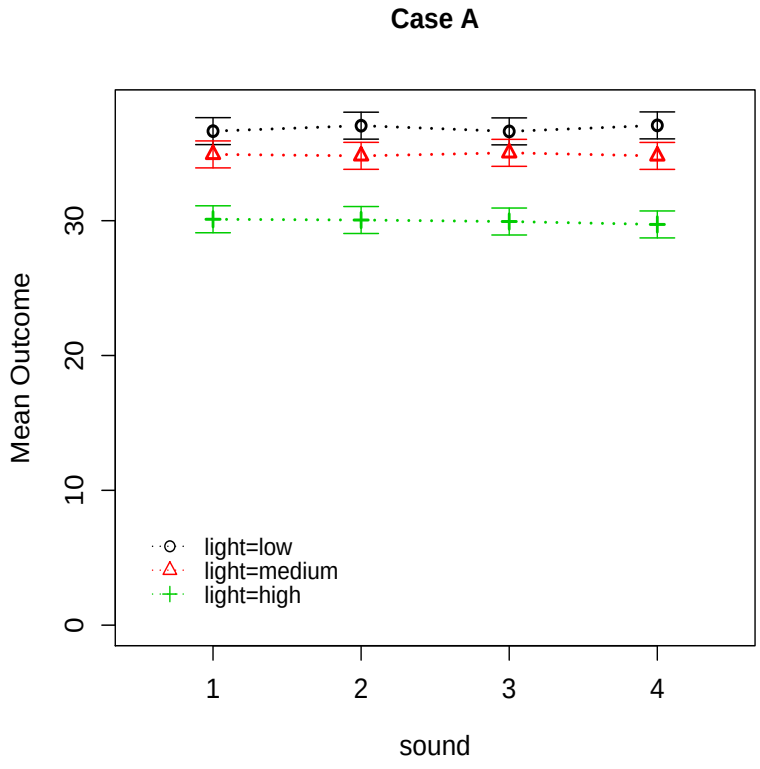


Figure 11.7: Case A for light/sound experiment.

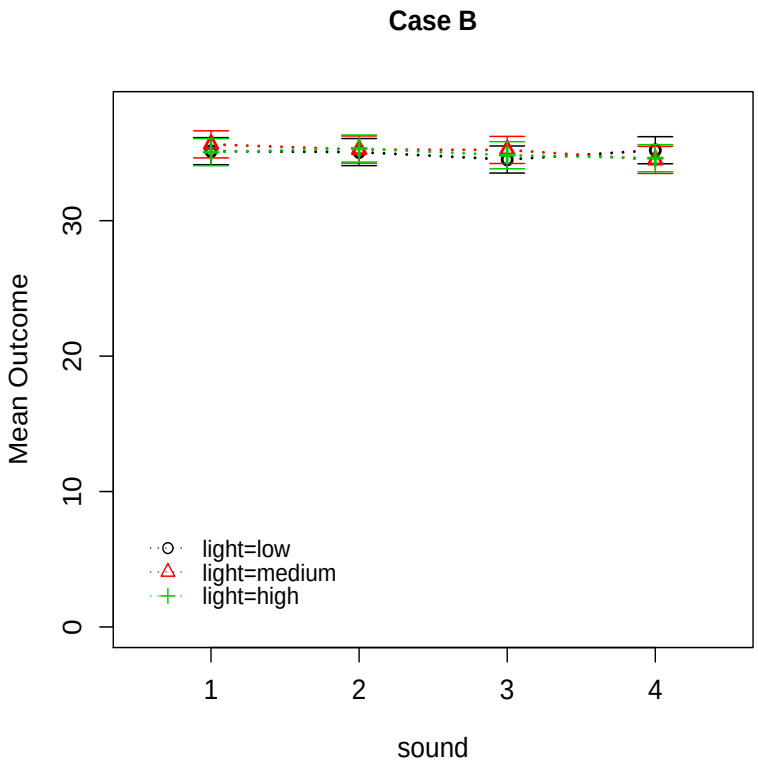


Figure 11.8: Case B for light/sound experiment.

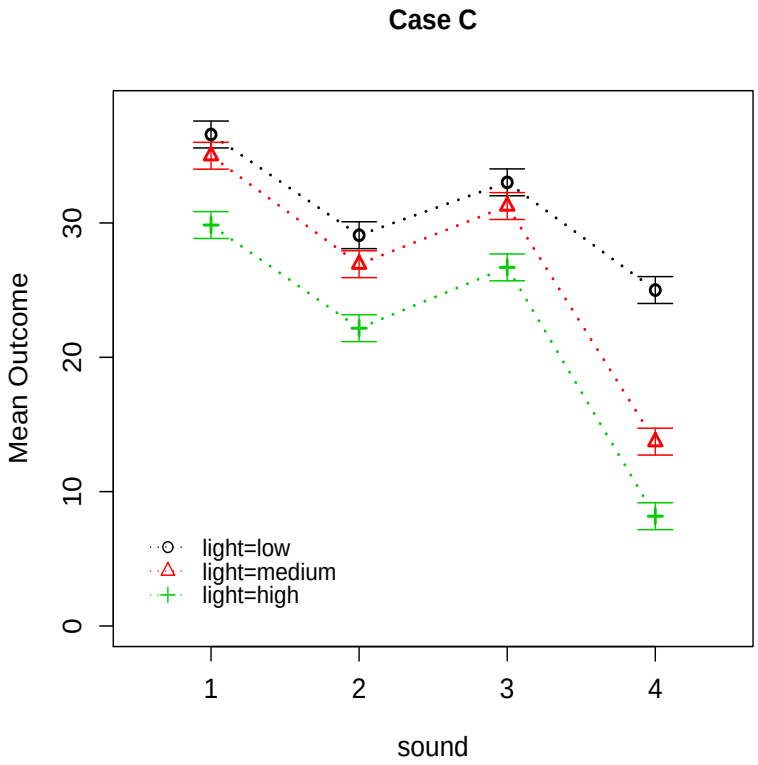


Figure 11.9: Case C for light/sound experiment.

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 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 now have no real interest. We know that both light and sound affect the outcome because the interaction p-value is significant. E.g., although we need contrast testing to be sure, it is quite obvious

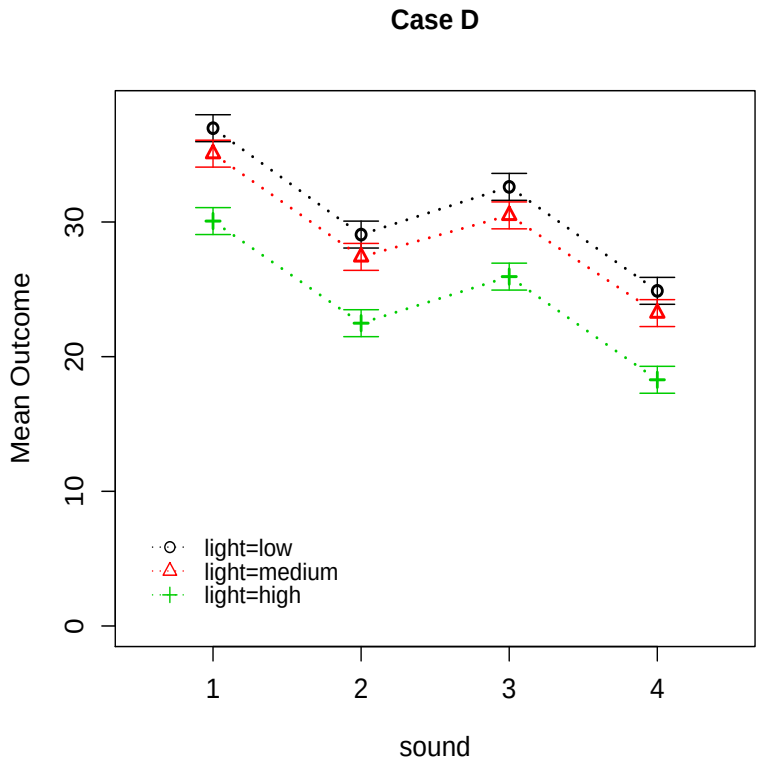


Figure 11.10: Case D for light/sound experiment.

that changing from low to high light level for any sound 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 for either factor, at at least one level of the other factor there are two levels of the first factor for which the outcome differs.

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) does not change the outcome. But the sound factor is statistically significant because changing between at least some pairs of sound levels for any light level does affect the outcome.

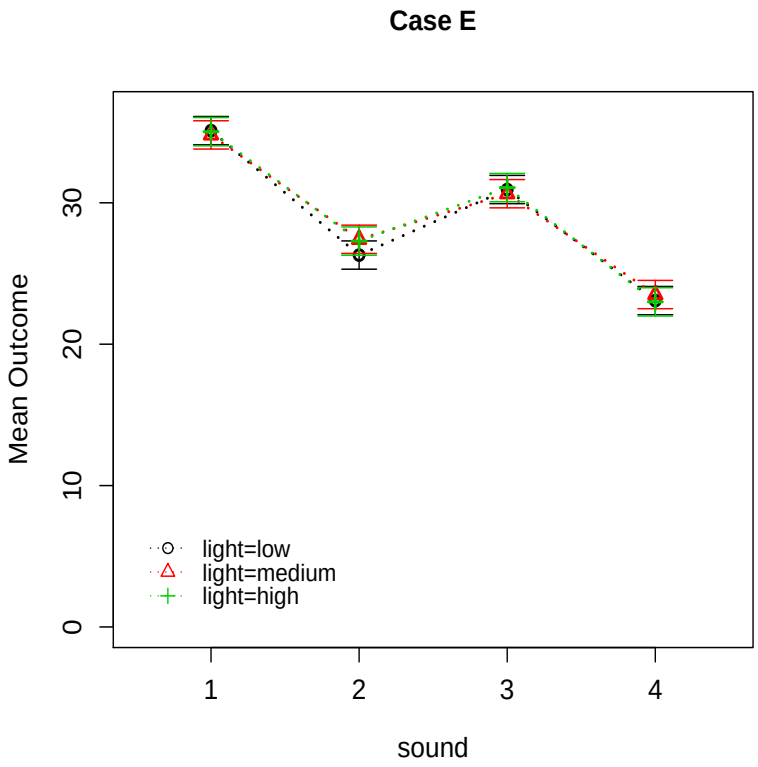


Figure 11.11: Case E for light/sound experiment.

Taking error into account, in most cases you can get a good idea which p-values will be significant just by looking at a (no-interaction) profile plot.

11.5 Do it in SPSS

To perform two-way ANOVA in SPSS use Analyze/GeneralLinearModel/Univariate from the menus. The "univariate" part means that there is only one kind of outcome measured for each subject. In this part of SPSS, you do *not* need to manually code indicator variables for categorical variables, or manually code interactions.

The Univariate dialog box is shown in figure 11.12. Enter the quantitative outcome in the Dependent Variable box. Enter the categorical explanatory variables in the Fixed Factors box. This will fit a model *with* an interaction.

Figure 11.12: SPSS Univariate dialog box.

To fit a model without an interaction, click the Model button to open the Univariate:Model dialog box, shown in figure 11.13. From here, choose "Custom"

instead of "Full Factorial", then do whatever it takes (there are several ways to do this) to get both factors, but not the interaction into the "Model" box, then click Continue.

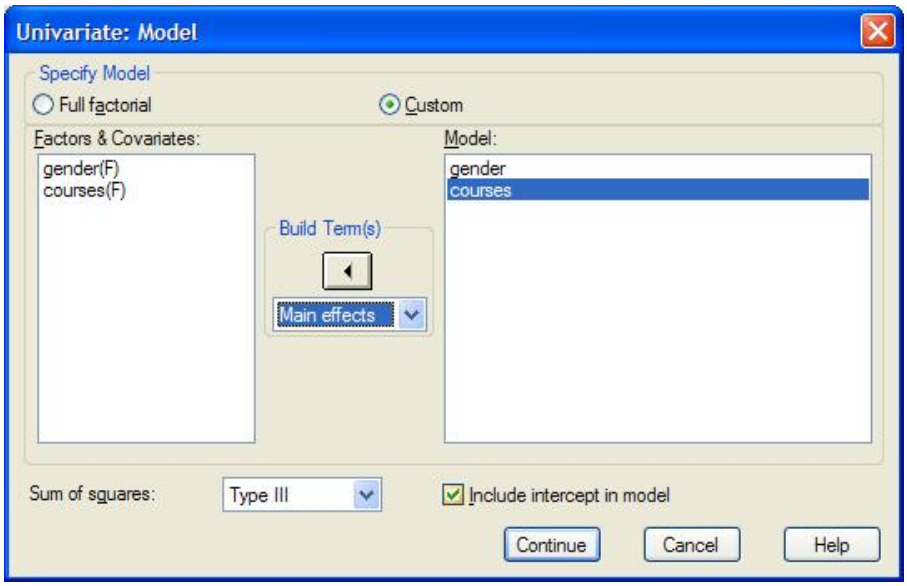


Figure 11.13: SPSS Univariate:Model dialog box.

For either model, it is a good idea to go to Options and turn on "Descriptive statistics", and "Residual plot". The latter is the 3 by 3 plot in which the usual residual vs. fit plot is in the center of the bottom row. Also place the individual factors in the "Display Means for" box if you are fitting a no-interaction model, or place the interaction of the factors in the box if you are fitting a model with an interaction.

If you use the Save button to save predicted and residual values (either standardized or unstandardized), this will create new columns in you data sheet; then a scatter plot with predicted on the x-axis and residual on the y-axis gives a residual vs. fit plot, while a quantile-normal plot of the residual column allows you to check the Normality assumption.

Under the Plots button, put one factor (usually the one with more levels) in the "Horizontal Axis" box, and the other factor in the "Separate Lines" box, then click Add to make an entry in the Plots box, and click Continue.

Finally, click OK in the main Univariate dialog box to perform the analysis.

Chapter 12

Statistical Power

12.1 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 analysis of 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 6.2.5).

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 $F_{\text{experiment}} \geq F_{\text{critical}}$. If the assumptions are met (I won't keep repeating this) then 5% of the time when experiments are run on equivalent treatments, (i.e. $\mu_1 = \cdots = \mu_k$), we will falsely reject H_0 because our experiment's F-value happens to fall above F-critical. This is the so-called Type 1 error (see Section 8.4). 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.

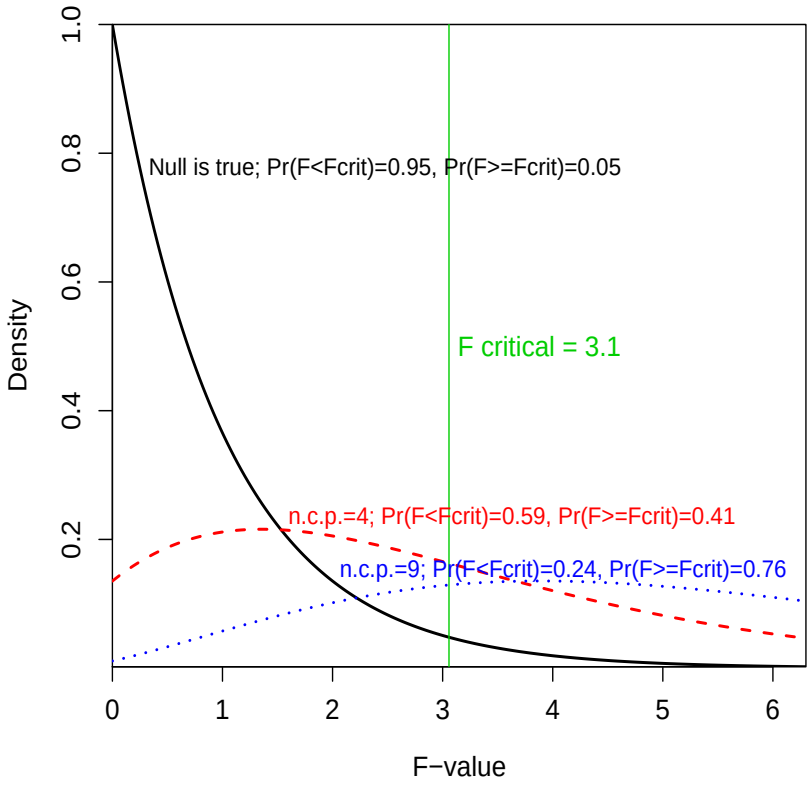


Figure 12.1: Null and alternative F sampling distributions.

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. An alternative sampling distribution represents how likely different values of a statistic such as F would be if we repeat an experiment many times when a particular alternative hypothesis is true. You can think of this as the histogram that results from running the experiment many times when the particular alternative is true and the F-statistic is calculated for each experiment.

As an example, figure 12.1 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 ($MS_{between}$) will have $k-1 = 2$ df, and the denominator(MS_{within}) will have $k(n-1) = 147$ df. Therefore the null sampling distribution 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.

Now consider what happens if the null hypothesis is not true (but the error model assumptions hold). There are many ways that the null hypothesis can be false, so 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). The number of treatments is also implicitly included on this list. 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).

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 $\sigma^2 = 100$ 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. (Note that if this alternative scenario is correct, it is impossible to make a Type 1 error; such an error can only be made when the truth is that the population means do *not* differ.)

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, 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.

You are probably not surprised to learn that for any given experiment and inference method (statistical test), the power to correctly reject a given alternative hypothesis lies somewhere between 5% and (almost) 100%. 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. The corresponding chance of obtaining $p \leq 0.05$ when a particular alternative scenario is true, called the power of the experiment, ranges from as low as 5% to near 100%.

12.2 Improving power

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). We can equivalently analyze data from an experiment like this using either the independent samples t-test or one-way ANOVA.

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, $|\mu_V - \mu_P|$.

This is demonstrated in figure 12.2. 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 can be thought of as the histogram of the actual click outcomes for one font for all news websites on the World Wide Web. 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

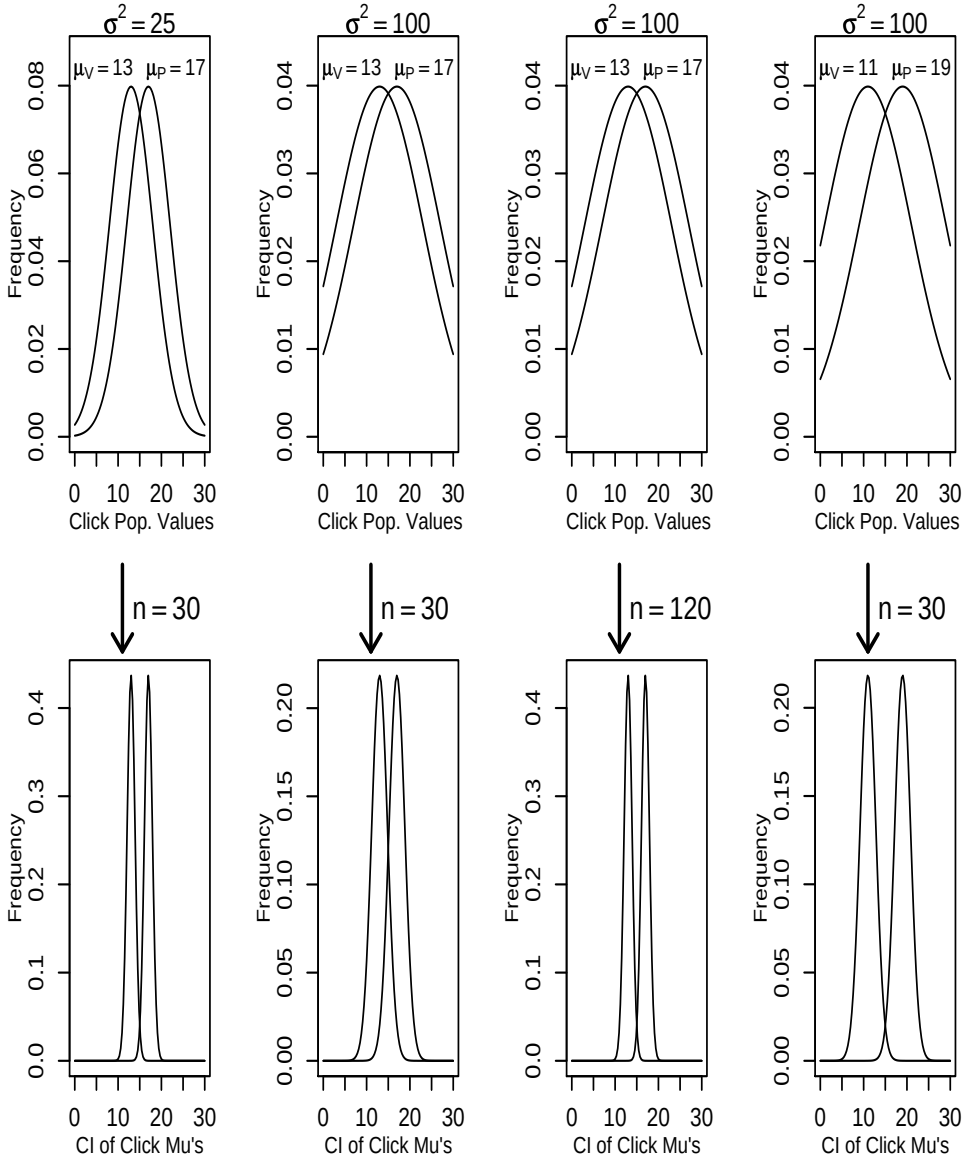


Figure 12.2: Effects of changing variance, sample size, and mean difference on power. Top row: population distributions of the outcome. Bottom row: sampling distributions of the sample mean for the given sample size.

markedly reduces the overlap, so the power is improved. How did we reduce the common variance? Either by reducing some of the four sources of variation or by using a within-subjects design, or by using a blocking variable or quantitative control variable. Specific examples for reducing the sources of variation include using only television-related websites, controlling the position of the ad on the website, and using only one font size for the ad. (Presumably for this experiment there is no measurement error.) 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 (in a 2-way ANOVA). We would guess that for each level of this second variable the variance of the outcome for either treatment would be smaller than if we had ignored the television-relatedness factor. Finally using a quantitative variable like site volume (hit count) as an additional explanatory variable in an ANCOVA setting would similarly reduce variability (i.e., σ^2) at each hit count value.

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. In this example, it is hard to see how the difference between the two fonts can be made larger, but in other experiments it is possible to make the treatments more different (i.e., make the active treatment, but not the control, "stronger") to increase 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 8.5 reduction in all sources of variation except measurement variability tends to also 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.

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.

12.3 Specific researchers' lifetime experiences

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 and claim an interesting finding. "Negative" means finding $p > \alpha$, which is the situation for which we retain H_0 and therefore don't have enough evidence to claim an interesting finding. "True" means correct (i.e. reject H_0 when H_0 is false or retain H_0 when H_0 is true), and "false" mean incorrect. These terms are commonly put together, e.g., a false positive refers to the case where $p \leq 0.05$, 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. 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.

It is worth knowing something about the relationship of power to confidence intervals. Roughly, wide confidence intervals correspond to experiments with low power, and narrow confidence intervals correspond to experiments with good power.

The error rates that experimenters are really interested in, i.e., the probability that I am making an error for my current experiment, are not knowable. These error rates differ from both α and β=1-power.

12.4 Expected Mean Square

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 repeated experiments.

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 not a variance, because it is calculated from fixed parameters rather than from random quantities, but it obviously is a "variance-like" quantity. 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 (including subject-to-subject, treatment application, environmental, and measurement variability). 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 12.1 and 12.2.

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	$\sigma_e^2 + n\sigma_A^2$
Error (residual)	MS_{error}	σ_e^2

Table 12.1: Expected mean squares for a one-way ANOVA.

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 12.2: 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.

12.5 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 of 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 a reasonably high power.

It is worth noting that you will not be able to calculate the "true" power of your experiment. Rather you will use a combination of mathematics and judgement to make a useful estimation of the power.

There are an infinite number of alternative hypothesis. 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):

$$\text{Expected Value of F} = \text{Expected value of } \frac{MS_A}{MS_{\text{error}}} \approx \frac{\sigma_e^2 + n\sigma_A^2}{\sigma_e^2}$$

Obviously 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 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 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.

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.

12.6 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. 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. From prior subject matter knowledge, he estimate 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.

12.7 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
```

In SPSS, put the value of 1-α (here, 1-0.10=0.90) in a spreadsheet cell, e.g., in a column named "P". The use Transform/Compute to create a variable called, say, "Fcrit", using the formula "IDF.F(P,3,40)". This will give 2.23. The use Transform/Compute to create a variable called, say, "power", using the formula "1-NCDF.F(Fcrit,3,40,6.11)". This will give 0.62.

12.8 A power applet

The Russ Lenth power applet is very nice way to calculate power. It is available on the web at http://www.cs.uiowa.edu/~rlenth/Power. If you are using it more that occasionally you should copy the applet to your website. Here I will cover ANOVA and regression. Additional topic are in future chapters.

12.8.1 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.

12.8.2 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".

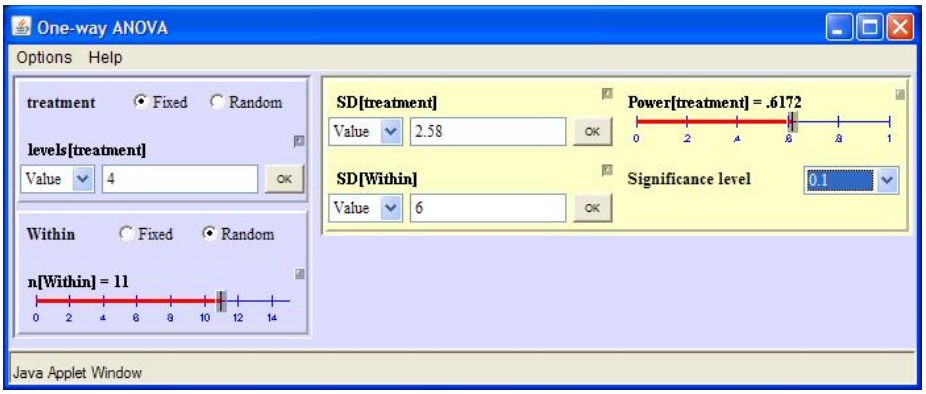


Figure 12.3: One-way ANOVA with Lenth power applet.

Enter "n" under "Observations per factor combination", and click to study the power of "F tests". A window opens that looks like figure 12.3.

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 12.4.

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 12.7 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 12.7

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 12.4. The lower curve is for the smaller value of SD[treatment].

12.8.3 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]

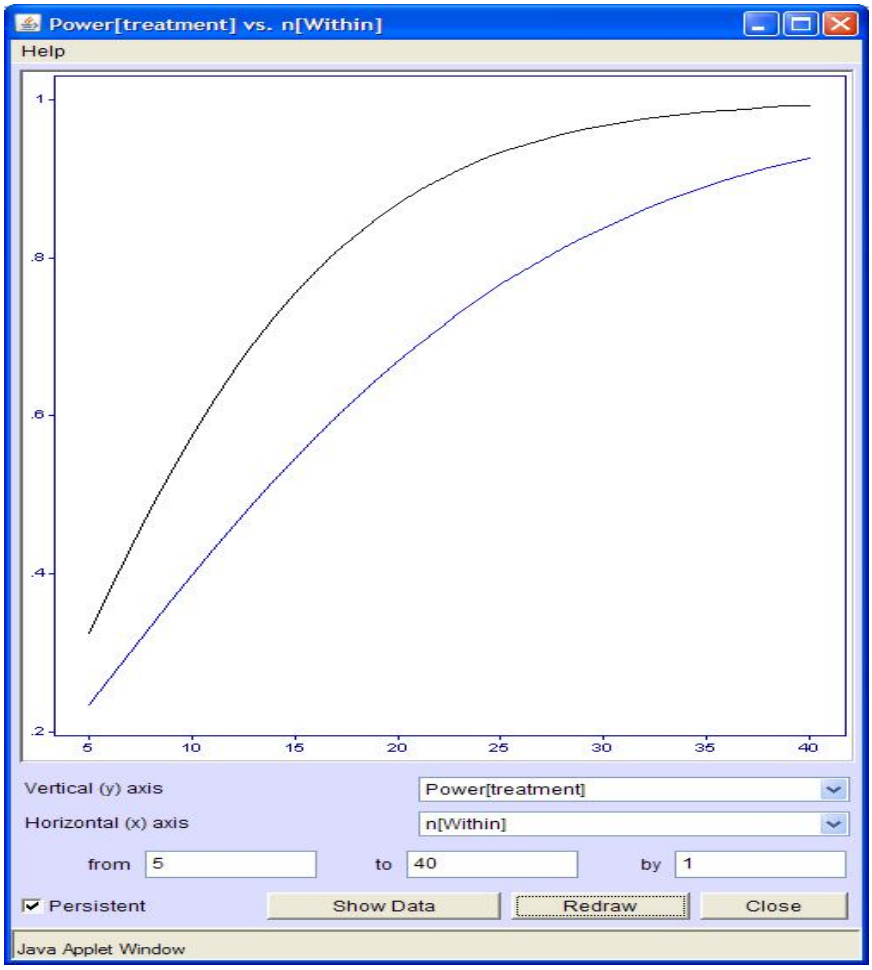


Figure 12.4: One-way ANOVA power plot from Lenth power applet.

(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}.

12.8.4 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 12.5.

12.8.5 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 12.6 appears. Leave "No. of predictors" (number of explanatory variables) at 1, and set "Alpha", "Error SD" (estimate of σ), and "(Total) Sample

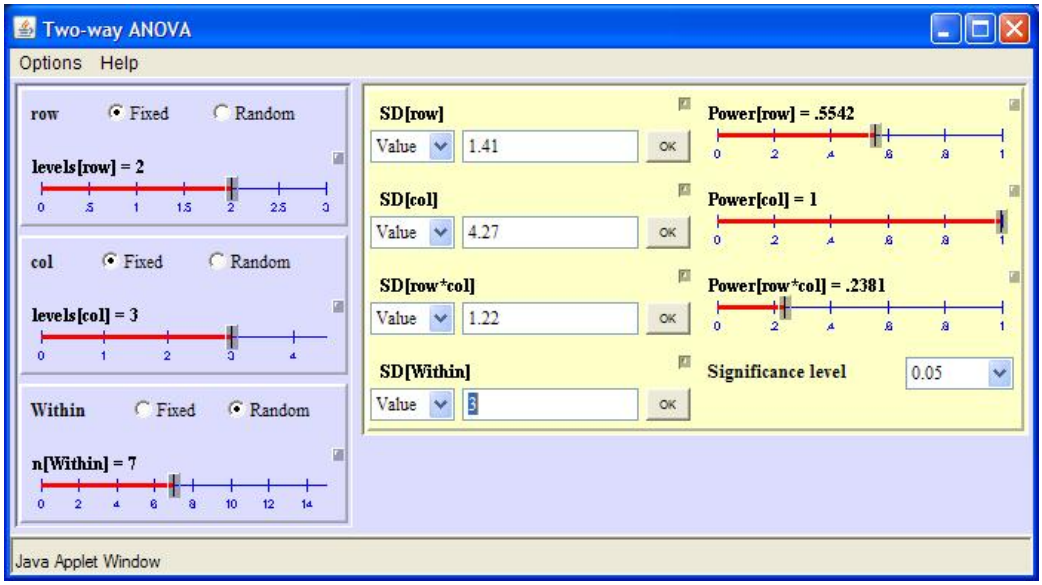


Figure 12.5: Two-way ANOVA with Lenth power applet.

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.

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.

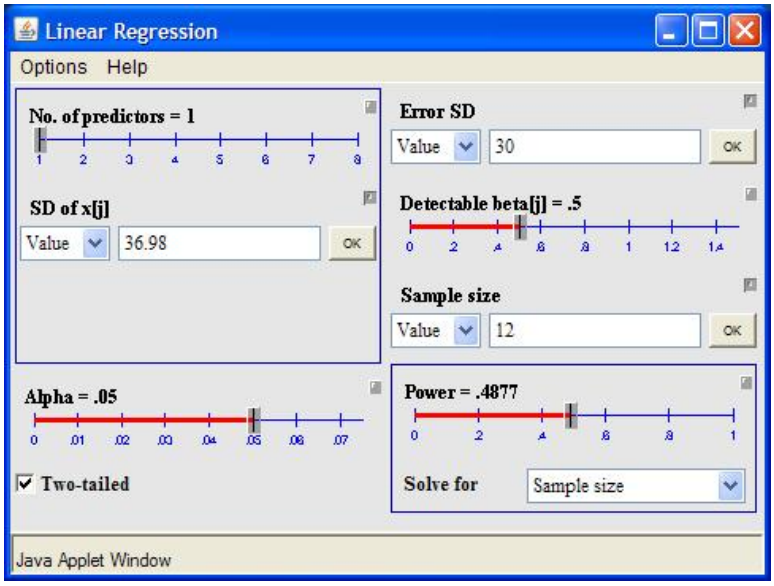


Figure 12.6: Linear regression with Lenth power applet.

Chapter 13

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. For some specific group population means we would like to know if we have sufficient evidence that they differ from certain other group population means. E.g., in a test of the effects of control and two active treatments to increase vocabulary, we might find that based on a the 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 hypothesis**.

Comparison and analytic comparison are other synonyms for contrast.

13.1 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: $H_1 : \mu_1 \neq \mu_5$.

A contrast null hypotheses 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 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, 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 hypothesis 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 OK to multiply the set of coefficients by any (non-zero) number. E.g., 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 in to account when you interpret confidence intervals. If you really want to get a confidence interval on the difference in average group population outcome means for the first two vs. the next three treatments, it will be directly interpretable only in the fraction form.

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

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.

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.

13.2 Planned comparisons

The ANOVA module of most statistical computer packages allow entry of custom hypotheses through contrast coefficients, but the p-values produced are only valid under stringent conditions called **planned comparisons** or planned contrasts or planned custom hypotheses. Without meeting these conditions, the p-values will be smaller than 0.05 more than 5% of the time, often far more, when the null hypothesis is true (i.e., when you are studying ineffectual treatments). In other words, these requirement are needed to maintain the Type 1 error rate across the entire experiment.

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 conditions needed for planned comparisons are:

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 ($\mu_1 = \cdots = \mu_k$) is not rejected in the ANOVA.
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. For the 36309 course, you can consider this paragraph optional. 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 reiterate the requirements of planned comparisons, let's consider the consequences of breaking each requirement. If you construct your contrasts after looking at your experimental results, you will naturally choose to compare the biggest and the smallest sample means, which suggests that you are implicitly comparing all of the sample means to find this interesting pair. Since each comparison has a 95% chance of correctly retaining the null hypothesis when it is true, after m independent tests you 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. Put another way, choosing which groups to compare after looking at results puts you at risk of making a false claim 14, 23 and 40% of the time respectively. (In reality the numbers are often slightly better because of lack of independence of the contrasts.)

The same kind of argument applies to looking at your planned comparisons without first "screening" with the overall p-value of the ANOVA. Screening protects your Type 1 experiment-wise error rate, while lack of screening raises it.

Using orthogonal contrasts is also required to maintain your Type 1 experiment-wise error rate. Correlated null hypotheses tend to make the chance of having several simultaneous rejected hypotheses happen more often than should occur when the null hypothesis is really true.

Finally, making more than $k-1$ planned contrasts (or $k-1$ and $m-1$ contrasts for a two-way $k \times m$ ANOVA without interaction) increases your Type 1 error

because each additional test is an additional chance to reject the null hypothesis incorrectly whenever the null hypothesis actually is true.

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.

In SPSS, anything entered as "Contrasts" (in menus) or "LMATRIX" (in Syntax, see Section 5.1) is tested as if it is a planned contrast.

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.

13.3 Unplanned or post-hoc contrasts

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 $p \leq \alpha/15 = 0.0033$.

The meaning of conservative is that this procedure is often more stringent than necessary, and using some other valid procedure might show a statistically significant result in some cases where the Bonferroni correction shows no statistical significance.

The Bonferroni procedure is completely general. 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 here that none of these procedures is needed for $k = 2$. If you try to apply them, you will either get some form of "not applicable" or you will get no penalty, i.e., the overall $\mu_1 = \mu_2$ hypothesis p-value is what is applicable.

Another post-hoc procedure is Dunnett's test. This makes the appropriate penalty correction for comparing one (control) group to all other groups.

The total number of available post-hoc procedures is huge. Whenever you see such an embarrassment of riches, you can correctly conclude that there is some lack of consensus on the matter, and that applies here. I recommend against using most of these, and certainly it is very bad practice to try as many 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 or from $(k-1) \times \text{IDF.F}(0.95, k-1, N-k)$ in SPSS.

Although Scheffé is a choice in the SPSS Post-Hoc dialog box, it doesn't make much sense to choose this because it only compares all possible pairs, but applies the penalty needed to allow all possible contrasts. 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). You can either use the menu or syntax in SPSS to compute the contrast estimate (g) and its standard error (SE(g)), or calculate these manually. 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.

13.4 Do it in SPSS

SPSS has a Contrast button that opens a dialog box for specifying planned contrasts and a PostHoc button that opens a dialog box for specifying various post-hoc procedures. In addition, planned comparisons can be specified by using the Paste button to examine and extend the Syntax (see Section 5.1) of a command to include one or more contrast calculations.

13.4.1 Contrasts in one-way ANOVA

Here we will examine planned and post-hoc contrast analyses for an experiment with three levels of an independent variable called "additive" (which is a chemical additive to a reaction, and has nothing to do with additive vs. interaction model types). The outcome is the number of hours until the reaction completes.

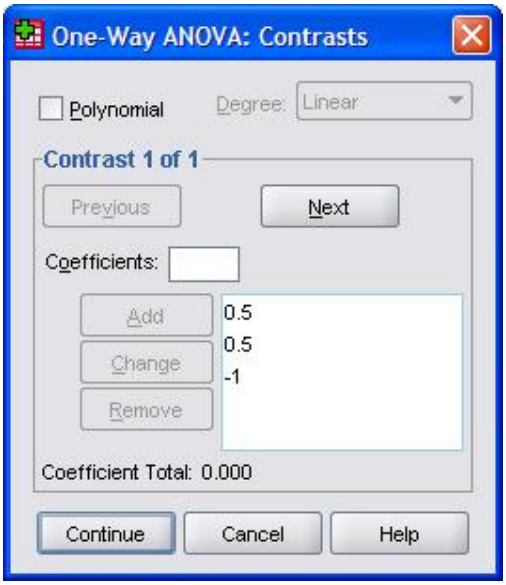


Figure 13.1: One-way ANOVA contrasts dialog box.

For a k-level one-way (between-subjects) ANOVA, accessed using Analyze/OneWayANOVA on the menus, the Contrasts button opens the "One-Way ANOVA: Contrasts" dialog box (see figure 13.1). From here you can enter the coefficients for each planned contrast. For a given contrast, enter the k coefficients that define any given contrast into the box labeled "Coefficients:" as a decimal number (no fractions allowed). Click the "Add" button after entering each of the coefficients. For a k-level ANOVA, you must enter all k coefficients, even if some are zero. Then you should check if the "Coefficient Total" equals 0.000. (Sometimes, due to rounding, this might be slightly above or below 0.000.) If you have any additional contrasts to add, click the Next button and repeat the process. Click the Continue button when you are finished. The figure shows a planned contrast for comparing the mean outcome (hours) for additives 1 and 2 to the mean outcome for additive 3.

When entering contrast coefficients in one-way ANOVA, SPSS will warn you and give no result if you enter more or less than the appropriate number of coefficients. It will not warn you if you enter more than $k - 1$ contrasts, if your coefficients do not add to 0.0, or if the contrasts are not orthogonal. Also, it will not prevent you from incorrectly analyzing post-hoc comparisons as planned comparisons.

The results for this example are given in Table 13.1. You should always look

Contrast Coefficients

	additive		
Contrast	1	2	3
1	0.5	0.5	-1
2	1	-1	0

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
hrs	Assume equal variance	1	-0.452	0.382	-1.18	47	0.243
		2	0.485	0.445	1.09	47	0.282
	Does not assume equal variance	1	-0.452	0.368	-1.23	35.58	0.228
		2	0.485	0.466	1.04	28.30	0.307

Table 13.1: Contrast results for one-way ANOVA.

at the Contrast Coefficients table to verify which contrasts you are testing. In this table, contrast 1, using coefficients (0.5, 0.5, -1) is testing $H_{01} : \frac{\mu_1+\mu_2}{2} - \mu_3 = 0$. Contrast 2 with coefficients (1, -1, 0) is testing $H_{02} : \mu_1 - \mu_2 = 0$.

The Contrast Tests table shows the results. Note that "hrs" is the name of the outcome variable. The "Value of the Contrast" entry is the best estimate of the contrast. For example, the best estimate of $\mu_1 - \mu_2$ is 0.485. The standard error of this estimate (based on the equal variance section) is 0.445 giving a t-statistic of 0.485/0.445=1.09, which corresponds to a p-value of 0.282 using the t-distribution with 47 df. So we retain the null hypothesis, and an approximate 95% CI for $\mu_1 - \mu_2$ is $0.485 \pm 2 \times 0.445 = [-0.405, 1.375]$. If you have evidence of unequal variance (violation of the equal variance assumption) you can use the lower section which is labeled "Does not assume equal variances."

In SPSS, the two post-hoc tests that make the most sense are Tukey HSD and Dunnett. Tukey should be used when the only post-hoc testing is among all pairs of population means. Dunnett should be used when the only post-hoc testing is between a control and all other population means. Only one of these applies to a given experiment. (Although the Scheffé test is useful for allowing post-hoc testing of all combinations of population means, turning that procedure on in SPSS does not make sense because it still only tests all pairs, in which case Tukey is more appropriate.)

Multiple Comparisons

hrs

Tukey HSD

(I) additive	(J) additive	Mean Difference (I-J)	Std.Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	0.485	0.445	0.526	-0.593	1.563
	3	-0.209	0.445	0.886	-1.287	0.869
2	1	-0.485	0.445	0.526	-1.563	0.593
	3	-0.694	0.439	0.263	-1.756	0.367
3	1	0.209	0.445	0.886	-0.869	1.287
	2	0.694	0.439	0.263	-0.367	1.756

Homogeneous Subsets

hrs

Tukey HSD

additive	N	Subset for alpha=0.05
		1
2	17	16.76
1	16	17.244
3	17	17.453
Sig.		0.270

Table 13.2: Tukey Multiple Comparison results for one-way ANOVA.

Table 13.2 shows the Tukey results for our example. Note the two columns labeled I and J. For each combination of levels I and J, the "Mean Difference (I-J)" column gives the mean difference subtracted in that order. For example, the first mean difference, 0.485, tells us that the sample mean for additive 1 is 0.485 higher than the sample mean for additive 2, because the subtraction is I (level 1) minus J (level 2). The standard error of each difference is given. This standard error is used in the Tukey procedure to calculate the corrected p-value that is appropriate for post-hoc testing. For any contrast that is (also) a planned contrast, you should ignore the information given in the Multiple Comparisons table, and instead use the information in the planned comparisons section of the output. (The p-value for a planned comparison is smaller than for the corresponding post-hoc test.)

The Tukey procedure output also gives a post-hoc 95% CI for each contrast. Note again that if a contrast is planned, we use the CI from the planned contrasts section and ignore what is in the multiple comparisons section. Contrasts that are made post-hoc (or analyzed using post-hoc procedures because they do not meet the four conditions for planned contrasts) have appropriately wider confidence intervals than they would have if they were treated as planned contrasts.

The Homogeneous Subsets table presents the Tukey procedure results in a different way. The levels of the factor are presented in rows ordered by the sample means of the outcome. There are one or more numbered columns that identify "homogeneous subsets." One way to read this table is to say that all pairs are significantly different except those that are in the same subset. In this example, with only one subset, no pairs have a significant difference.

You can alternately use the menu item Analyze/GeneralLinearModel/Univariate for one-way ANOVA. Then the Contrasts button does not allow setting arbitrary contrasts. Instead, there a fixed set of named planned contrasts. Figure 13.2 shows the "Univariate: Contrasts" dialog box. In this figure the contrast type has been changed from the default "None" to "Repeated". Note the word "Repeated" under Factors confirms that the change of contrast type has actually been registered by pressing the Change button. *Be sure to also click the Change button whenever you change the setting of the Contrast choice, or your choice will be ignored!* The pre-set contrast choices include "Repeated" which compares adjacent levels, "Simple" which compares either the first or last level to all other levels, polynomial which looks for increasing orders of polynomial trends, and a few other less useful ones. These are all intended as planned contrasts, to be chosen before running the experiment.

Figure 13.2: Univariate contrasts dialog box.

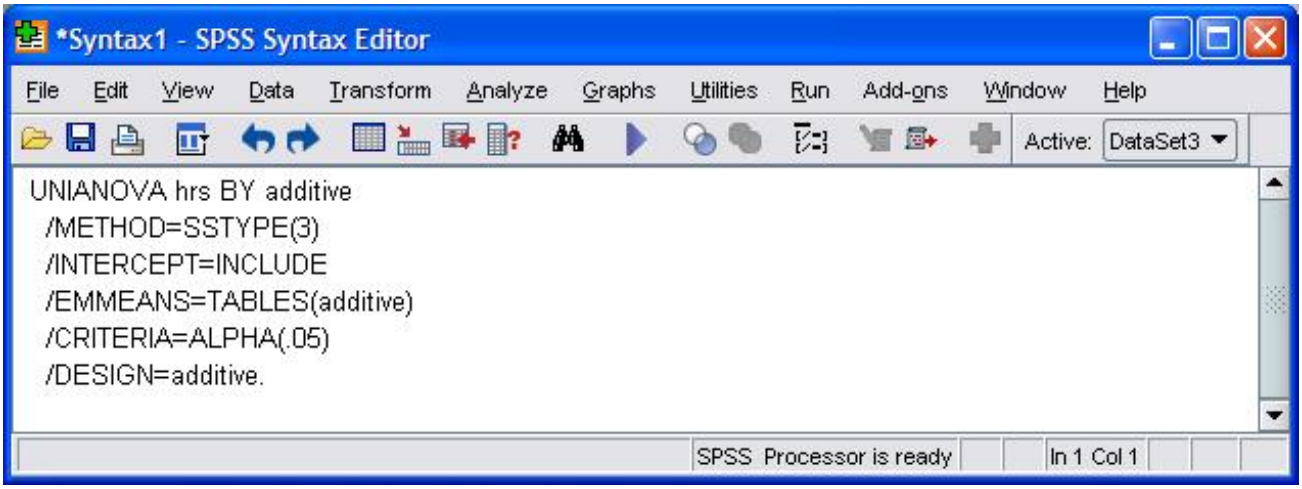


Figure 13.3: Univariate syntax window.

Custom Hypothesis Tests #1

Contrast Results (K Matrix)

Contrast			Dependent hrs
L1	Contrast Estimate		0.138
	Hypothesized Value		0
	Difference(Estimate-Hypothesized)		0.138
	Std. Error		0.338
	Sig.		0.724
	95% Confidence Interval for Difference	Lower Bound	-0.642
		Upper Bound	0.918

Based on user-specified contrast coefficients: first vs. second and third

Table 13.3: Planned contrast in one-way ANOVA using LMATRIX syntax.

.

To make a *custom* set of planned contrasts in the Univariate procedure, click the Paste button of the Univariate dialog box. This brings up a syntax window with the SPSS native commands that are equivalent to the menu choices you have made so far (see Figure 13.3). You can now insert some appropriate subcommands to test your custom hypotheses. You can insert the extra lines anywhere between the first line and the final period. The lines that you would add to the Univariate syntax to test $H_{01} : \mu_1 - \frac{\mu_2+\mu_3}{2} = 0$ and $H_{02} : \mu_2 - \mu_3 = 0$ are:

```
/LMATRIX = "first vs. second and third" additive 1 -1/2 -1/2
/LMATRIX = "second vs. third" additive 0 1 -1
```

Note that you can type any descriptive phrase inside the quotes, and SPSS will not (cannot) test if your phrase actually corresponds to the null hypothesis defined by your contrasts. Also note that fractions are allowed here. Finally, note that the name of the factor (additive) precedes its list of coefficients.

The output of the first of these LMATRIX subcommands is shown in Table 13.3. This gives the p-value and 95%CI appropriate for a planned contrast.

13.4.2 Contrasts for Two-way ANOVA

Contrasts in two-way (between-subjects) ANOVA *without* interaction work just like in one-way ANOVA, but with separate contrasts for each factor. Using the Univariate procedure on the Analyze/GeneralLinearModel menu, if one or both factors has more than two levels, then pre-defined planned contrasts are available with the Contrasts button, post-hoc comparisons are available with the Post-Hoc button, and arbitrary planned contrasts are available with Paste button and LMATRIX subcommands added to the Syntax.

For a $k \times m$ two-way ANOVA *with* interaction, two types of contrasts make sense. For planned comparisons, out of the km total treatment cells, you can test up to $(k-1)(m-1)$ pairs out of the $\binom{km}{2} = \frac{km(km-1)}{2}$ total pairs. With the LMATRIX subcommand you can only test a particular subset of these: comparisons between any two levels of one factor when the other factor is fixed at any particular level. To do this, you must first check the order of the two factors in the DESIGN line of the pasted syntax. If the factors are labeled A and B, the line will look either like

```
/DESIGN=A B A*B
```

or

```
/DESIGN=B A B*A
```

Let's assume that we have the "A*B" form with, say, 3 levels of factor A and 2 levels of factor B. Then a test of, say, level 1 vs. 3 of factor A when factor B is fixed at level 2 is performed as follows: Start the LMATRIX subcommand in the usual way:

```
/LMATRIX="compare A1B2 to A3B2"
```

Then add coefficients for the varying factor, which is A in this example:

```
/LMATRIX="compare A1B2 to A3B2"  A 1 0 -1
```

Finally add the "interaction coefficients". There are km of these and the rule is "the first factor varies slowest". This means that if the interaction is specified as

A*B in the DESIGN statement then the first set of coefficients corresponds to all levels of B when A is set to level 1, then the next set is all levels of B when A is set to level 2, etc. For our example with we need to set A1B2 to 1 and A3B2 to -1, while setting everything else to 0. The correct subcommand is:

```
/LMATRIX="compare A1B2 to A3B2"  A 1 0 -1   A*B 0 1   0 0   0 -1
```

It is helpful to space out the A*B coefficients in blocks to see what is going on better. The first block corresponds to level 1 of factor A, the second block to level 2, and the third block to level 3. Within each block the first number is for B=1 and the second number for B=2. It is in this sense that B is changing quickly and A slowly as we move across the coefficients. To reiterate, position 2 in the A*B list corresponds to A=1 and B=2, while position 6 corresponds to A=3 and B=2. These two have coefficients that match those of the A block (1 0 -1) and the desired contrast ($\mu_{A1B2} - \mu_{A3B2}$).

To test other types of planned pairs or to make post-hoc tests of all pairs, you can convert the analysis to a one-way ANOVA by combining the factors using a calculation such as 10*A+B to create a single factor that encodes the information from both factors and that has km different levels. Then just use one-way ANOVA with either the specific planned hypotheses or the with the Tukey post-hoc procedure.

The other kind of hypothesis testing that makes sense in two-way ANOVA with interaction is to test the interaction effects directly with questions such as "is the effect of changing from level 1 to level 3 of factor A when factor B=1 the same or different from the effect of changing from level 1 to level 3 of factor A when factor B=2?" This corresponds to the null hypothesis: $H_0 : (\mu_{A3B1} - \mu_{A1B1}) - (\mu_{A3B2} - \mu_{A1B2}) = 0$. This can be tested as a planned contrast within the context of the two-way ANOVA with interaction by using the following LMATRIX subcommand:

```
/LMATRIX="compare A1 to A3 for B1 vs. B2"  A*B -1 1   0 0   1 -1
```

First note that we *only* have the interaction coefficients in the LMATRIX subcommand for this type of contrast. Also note that because the order is A then B in A*B, the A levels move change slowly, so the order of effects is A1B1 A1B2 A2B1 A2B2 A3B1 A3B2. Now you can see that the above subcommand matches the above null hypothesis. For an example of interpretation, assume that for fixed levels of both B=1 and B=2, A3-A1 is positive. Then a positive Contrast Estimate

for this contrast would indicate that the outcome difference with B=1 is greater than the difference with B=2.

Chapter 14

Within-Subjects Designs

ANOVA must be modified to take correlated errors into account when multiple measurements are made for each subject.

14.1 Overview of within-subjects designs

Any categorical explanatory variable for which each subject experiences all of the levels is called a **within-subjects factor**. (Or sometimes a subject may experience several, but not all levels.) These levels could be different "treatments", or they may be different 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 the 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 14.5.

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 15.) All of the

experiments discussed in the preceding chapters are between-subjects designs.

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 between-subjects and within-subjects refer to experimental designs that either do not or do make multiple measurements on each subject.

When a within-subjects factor is used in an experiment, new methods are needed that do not make the assumption of no correlation (or, somewhat more strongly, independence) of errors for the multiple measurements made on the same subject. (See section 6.2.8 to review the independent errors assumption.)

Why would we want to make multiple measurements on the same subjects? 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. 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 from our investigation of the relative effects of different treatments. This reduced variability directly increases power, often dramatically. We may use this increased power directly, or we may use it indirectly to allow a reduction in the number of subjects studied.

These are very important advantages to using within-subjects designs, and such designs are widely used. The major reasons for not using within-subjects designs are when it is impossible to give multiple treatments to a single subject or because of concern about confounding. An example of a case where a within-subjects design is impossible is a study of surgery vs. drug treatment for a disease; subjects generally would receive one or the other treatment, not both.

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.)

You need to know how to distinguish within-subjects from between-subjects factors. Within-subjects designs have the advantages of more power and allow observation of change over time. The main disadvantage is possible confounding, which can often be overcome by using counterbalancing.

14.2 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 the (univariate) Normal distribution. The mean and variance, along with the standard bell-shape characterize the kinds of outcome values that we expect to see. Switching from the population to the sample, we can put the value of the outcome on the x-axis of a plot and the relative frequency of that

value on the y-axis to get a histogram that shows which values are most likely and from which we can visualize how likely a range of values is.

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 14.1 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.)

The idea of the bivariate distribution can easily be extended to more than two dimensions, but is of course much harder to visualize. A multivariate distribution with k-dimensions has a k-length vector (ordered set of numbers) representing its mean. It also has a $k \times k$ dimensional matrix (rectangular array of numbers) representing the variances of the individual variables, and all of the 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.,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix},$$

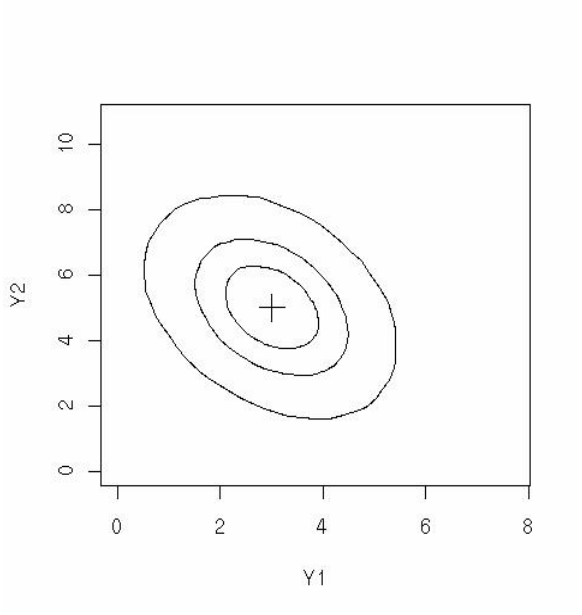


Figure 14.1: Contours enclosing 1/3, 2/3 and 95% of a bivariate Normal distribution with a negative covariance.

and a variance-covariance matrix, e.g.,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \gamma_{1,2} & \gamma_{1,3} \\ \gamma_{1,2} & \sigma_2^2 & \gamma_{2,3} \\ \gamma_{1,3} & \gamma_{2,3} & \sigma_3^2 \end{bmatrix}.$$

Here we are using $\gamma_{i,j}$ to represent the covariance of variable Y_i with Y_j.

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 $Y \sim N(\mu, \Sigma)$.

14.3 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.

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: e.g. call the difference of two of the measurements on each subject the response, and use standard techniques. 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 repetition of the same measurement 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 not met fairly frequently. 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 as oppposed to dropping all data from every subject who is missing one or more measurements), and it accommodates unequal and/or irregular spacing of repeated measurements. Mixed models can also be extended to non-normal outcomes. (See chapter 15.)

14.4 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 SPSS, the paired t-test requires the "wide" form of data in the spreadsheet rather than the "tall" form we have used up until now. 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).

The paired t-test uses a one-sample t-test on the single column of computed differences. Although we have not yet discussed the one-sample t-test, it is a straightforward extension of other t-tests like the independent-sample t-test of Chapter 6 or the one for regression coefficients in Chapter 9. We have an estimate of the difference in outcome between the two treatments in the form of the mean of the difference column. We can compute the standard error for that difference (which is the square root of the variance of the difference column divided by the number of subjects). Then we can construct the t-statistic as the estimate divided by the SE of the estimate, and under the null hypothesis that the population mean difference is zero, this will follow a t-distribution with $n - 1$ df, where n is the number of subjects.

The results from SPSS for comparing control to TENS ROM is shown in table 14.1. The table tells us that the best point estimate of the difference in population means for ROM between control and TENS is 17.70 with control being higher (because the direction of the subtraction is listed as control minus TENS). The uncertainty in this estimate due to random sampling variation is 7.256 on the standard deviation scale. (This was calculated based on the sample size of 10 and the observed standard deviation of 22.945 for the observed sample.) We are 95% confident that the true reduction in ROM caused by TENS relative to the control is between 1.3 and 34.1, so it may be very small or rather large. The t-statistic of 2.439 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, the incorrect, between-subjects one-way ANOVA analysis of these data gives a p-value of 0.123, leading to the (probably) incorrect conclusion that the two treatments both have the same population mean of ROM. For future discussion we note that the within-groups SS for this incorrect analysis is 10748.5 with 18 df.

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.

Paired Differences							
			95% Confidence Interval of the Difference				
Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
17.700	22.945	7.256	1.286	34.114	2.439	9	0.037

Table 14.1: Paired t-test for control-TENS ROM in the osteoarthritis experiment.

There is an important concept relating to the repeatability of levels of a factor. 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. (In SPSS it is equivalent to fit a model without an interaction.) Based on the formula for expected MS of an interaction (see section 12.4), 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.

The two-way ANOVA results are shown in table 14.2. Although we normally ignore the intercept, it is included here to demonstrate the idea that in within-subjects ANOVA (and other cases called nested ANOVA) the denominator of the F-statistic, which is labeled "error", can be different for different numerators (which

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	173166.05	1	173166.05	185.99	<0.0005
	Error	8379.45	9	931.05		
rx	Hypothesis	1566.45	1	1566.45	5.951	0.035
	Error	2369.05	9	263.23		
subject	Hypothesis	8379.45	9	931.05	3.537	0.037
	Error	2369.05	9	263.23		

Table 14.2: Two-way ANOVA results for the osteoarthritis experiment.

correspond to the different null hypotheses). The null hypothesis of main interest here is that the three treatment population means are equal, and that is tested and rejected on the line called "rx". The null hypothesis for the random subject effect is that the population variance of the subject-to-subject means (of all three treatments) is zero.

The key observation from this table is that the treatment (rx) SS and MS corresponds to the between-groups SS and MS in the incorrect one-way ANOVA, while the sum of the subject SS and error SS is 10748.5, which is the within-groups SS for the incorrect one-way ANOVA. This is a decomposition of the four sources of error (see Section 8.5) that contribute to σ^2, which is estimated by SS_{within} in the one-way ANOVA. In this two-way ANOVA the subject-to-subject variability is estimated to be 931.05, and the remaining three sources contribute 263.23 (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.

How do we know which error terms to use for which F-tests? That requires more mathematical statistics than we cover in this course, but SPSS will produce an EMS table, and it is easy to use that table to figure out which ratios are 1.0 when the null hypotheses are true.

It is worth mentioning that in SPSS a one-way within-subjects ANOVA can be analyzed either as a two-way ANOVA with subjects as a random factor (or even as a fixed factor if a no-interaction model is selected) or as a repeated measures analysis (see next section). The p-value for the overall null hypothesis, that the population outcome means are equal for all levels of the factor, is the same for

each analysis, although which auxiliary statistics are produced differs.

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).

14.5 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 tends to be used.

This narrow-sense repeated measures analysis is what you get if you choose "General Linear Model / Repeated Measures" in SPSS. It includes the second and third approaches of our list of approaches given in the introduction to this chapter. The various sections of the output are labeled univariate or multivariate to distinguish which type of analysis is shown.

This section discusses the k-level ($k \geq 2$) one-way within-subjects ANOVA using repeated measures in the narrow sense. The next section discusses the mixed within/between subjects two-way ANOVA.

First we need to look at 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. Within a subject the multiple measurements are assumed to be correlated. For the univariate analyses, the 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 variances are equal and all of the covariances (and correlations) are equal. 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.

In contrast, the multivariate portions of repeated measures analysis output are based on an unconstrained variance-covariance pattern. Essentially, all of the variances and covariances are estimated from the data, which allows accommodation of a wider variety of variance-covariance structures, but loses some power, particularly when the sample size is small, due to "using up" some of the data and degrees of freedom for estimating a more complex variance-covariance structure.

Because the univariate analysis requires the assumption of sphericity, it is customary to first examine the Mauchly's test of sphericity. 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, being often of low power and also overly sensitive to mild violations of the Normality assumption. 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 possible unequal to), and therefore Mauchly's test is inapplicable and not calculated when there are only two levels of treatment.

The basic overall univariate test of equality of population means for the within-subjects factor is labeled "Tests of Within-Subjects Effects" in SPSS and is shown in table 14.3. If we accept the sphericity assumption, e.g., because the test of sphericity is non-significant, then we use the first line of the treatment section and the first line of the error section. In this case F=MS_{between} divided by MS_{within}=1080.9/272.4=3.97. The p-value is based on the F-distribution with 2 and 18 df. (This F and p-value are exactly the same as the two-way ANOVA with subject as a random factor.)

If the sphericity assumption is violated, then one of the other, corrected lines of the Tests of Within-Subjects Effects table is used. There is some controversy about when to use which correction, but generally it is safe to go with the Huynh-Feldt correction.

The alternative, multivariate analysis, labeled "Multivariate Tests" in SPSS is shown in table 14.4. The multivariate tests are tests of the same overall null hypothesis (that all of the treatment population means are equal) as was used for

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
rx	Sphericity Assumed	2161.8	2	1080.9	3.967	.037
	Greenhouse-Geisser	2161.8	1.848	1169.7	3.967	.042
	Huynh-Feldt	2161.8	2.000	1080.9	3.967	.042
	Lower-bound	2161.8	1.000	1169.7	3.967	.042
Error(rx)	Sphericity Assumed	4904.2	18	272.4		
	Greenhouse-Geisser	4904.2	16.633	294.8		
	Huynh-Feldt	4904.2	18,000	272.4		
	Lower-bound	4904.2	9.000	544.9		

Table 14.3: Tests of Within-Subjects Effects for the osteoarthritis experiment.

the univariate analysis.

The approach for the multivariate analysis is to first construct a set of $k-1$ orthogonal contrasts. (The main effect and interaction p-values are the same for every set of orthogonal contrasts.) Then SS are computed for each contrast in the usual way, and also "sum of cross-products" are also formed for pairs of contrasts. These numbers are put into a $k-1$ by $k-1$ matrix called the SSCP (sums of squares and cross products) matrix. In addition to the (within-subjects) treatment SSCP matrix, an error SSCP matrix is constructed analogous to computation of error SS. The ratio of these matrices is a matrix with F-values on the diagonal and ratios of treatment to error cross-products off the diagonal. We need to make a single F statistic from this matrix to get a p-value to test the overall null hypothesis. Four methods are provided for reducing the ratio matrix to a single F value. These are called Pillai's Trace, Wilk's Lambda, Hotelling's Trace, and Roy's Largest Root. There is a fairly extensive, difficult-to-understand literature comparing these methods, but it most cases they give similar p-values.

The decision to reject or retain the overall null hypothesis of equal population outcome means for all levels of the within-subjects factor is made by looking at

Effect		Value	F	Hypothesis df	Error df	Sig.
modality	Pillai's Trace	0.549	4.878	2	8	0.041
	Wilk's Lambda	0.451	4.878	2	8	0.041
	Hotelling's Trace	1.220	4.878	2	8	0.041
	Roy's Largest Root	1.220	4.878	2	8	0.041

Table 14.4: Multivariate Tests for the osteoarthritis experiment.

the p-value for one of the four F-values computed by SPSS. I recommend that you use "Pillai's trace". The thing you should not do is pick the line that gives the answer you want! In a one-way within-subjects ANOVA, the four F-values will always agree, while in more complex designs they will disagree to some extent.

Which approach should we use, univariate or multivariate? Luckily, they agree most of the time. When they disagree, it could be because the univariate approach is somewhat more powerful, particularly for small studies, and is thus preferred. Or it could be that the correction is insufficient in the case of far deviation from sphericity, in which case the multivariate test is preferred as more robust. In general, you should at least look for outliers or mistakes if there is a disagreement.

An additional section of the repeated measures analysis shows the planned contrasts and is labeled "Tests of Within-Subjects Contrasts". This section is the same for both the univariate and multivariate approaches. It gives a p-value for each planned contrast. The default contrast set is "polynomial" which is generally only appropriate for a moderately large number of levels of a factor representing repeated measures of the same measurement over time. In most circumstances, you will want to change the contrast type to simple (baseline against each other level) or repeated (comparing adjacent levels).

It is worth noting that post-hoc comparisons are available for the within-subjects factor under Options by selecting the factor in the Estimated Marginal Means box and then by checking the "compare main effects" box and choosing Bonferroni as the method.

14.6 Mixed between/within-subjects designs

One of the most common designs used in psychology experiments is a two-factor ANOVA, where one factor is varied between subjects and the other within subjects. The analysis of this type of experiment is a straightforward combination of the analysis of two-way between subjects ANOVA and the concepts of within-subject analysis from the previous section.

The interaction between a within- and a between-subjects factor shows up in the within-subjects section of the repeated measures analysis. As usual, the interaction should be examined first. If the interaction is significant, then (changes in) both factors affect the outcome, regardless of the p-values for the main effects. Simple effects contrasts in a mixed design are not straightforward, and are not available in SPSS. A profile plot is a good summary of the results. Alternatively, it is common to run separate one-way ANOVA analyses for each level of one factor, possibly using planned and/or post-hoc testing. In this case we test the simple effects hypotheses about the effects of differences in level of one factor at fixed levels of the other factor, as is appropriate in the case of interaction. Note that, depending on which factor is restricted to a single level for these analyses, the appropriate ANOVA could be either within-subjects or between-subjects.

If an interaction is not significant, we usually remove it from the model, but that is not possible in repeated measures ANOVA. You should ignore it and interpret both "main effects" overall null hypothesis as equal means for all levels of one factor averaging over (or ignoring) the other factor. For the within-subjects factor, either the univariate or multivariate tests can be used.

There is a separate results section for the overall null hypothesis for the between subjects factor. Because this section compares means between levels of the between-subjects factor, and those means are reductions of the various levels of the within-subjects factor to a single number, there is no concern about correlated errors, and there is only a single univariate test of the overall null hypothesis.

For each factor you may select a set of planned contrasts (assuming that there are more than two levels and that the overall null hypothesis is rejected). Finally, post-hoc tests are available for the between-subjects factor, and either the Tukey or Dunnett test is usually appropriate (where Dunnett is used only if there is no interest in comparisons other than to the control level). For the within-subjects factor the Bonferroni test is available with Estimated Marginal Means.

Repeated measures analysis is appropriate when one (or more) factors is a within-subjects factor. Usually univariate and multivariate tests agree for the overall null hypothesis for the within-subjects factor or any interaction involving a within-subjects factor. Planned (main effects) contrasts are appropriate for both factors if there is no significant interaction. Post-hoc comparisons can also be performed.

14.6.1 Repeated Measures in SPSS

To perform a repeated measures analysis in SPSS, use the menu item "Analyze / General Linear Model / Repeated Measures." The example uses the data in circleWide.sav. This is in the "wide" format with a separate column for each level of the repeated factor.

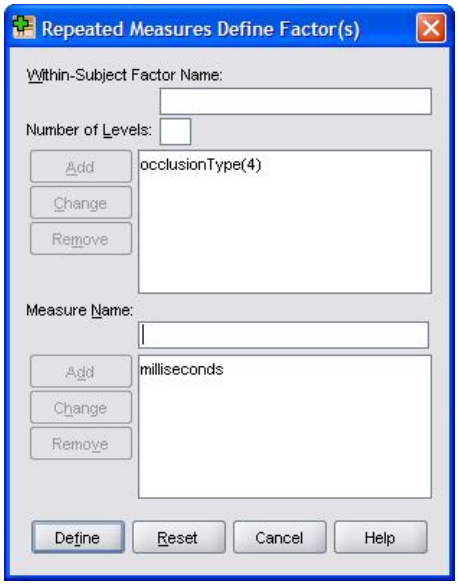


Figure 14.2: SPSS Repeated Measures Define Factor(s) dialog box.

Unlike other analyses in SPSS, there is a dialog box that you must fill out before seeing the main analysis dialog box. This is called the "Repeated Measures Define Factor(s)" dialog box as shown in Figure 14.2. Under "Within-Subject Factor

Name" you should enter a (new) name that describes what is different among the levels of your within-subjects factor. Then enter the "Number of Levels", and click Add. In a more complex design you need to do this for each within-subject factor. Then, although not required, it is a very good idea to enter a "Measure Name", which should describe what is measured at each level of the within-subject factor. Either a term like "time" or units like "milliseconds" is appropriate for this box. Click the "Define" button to continue.

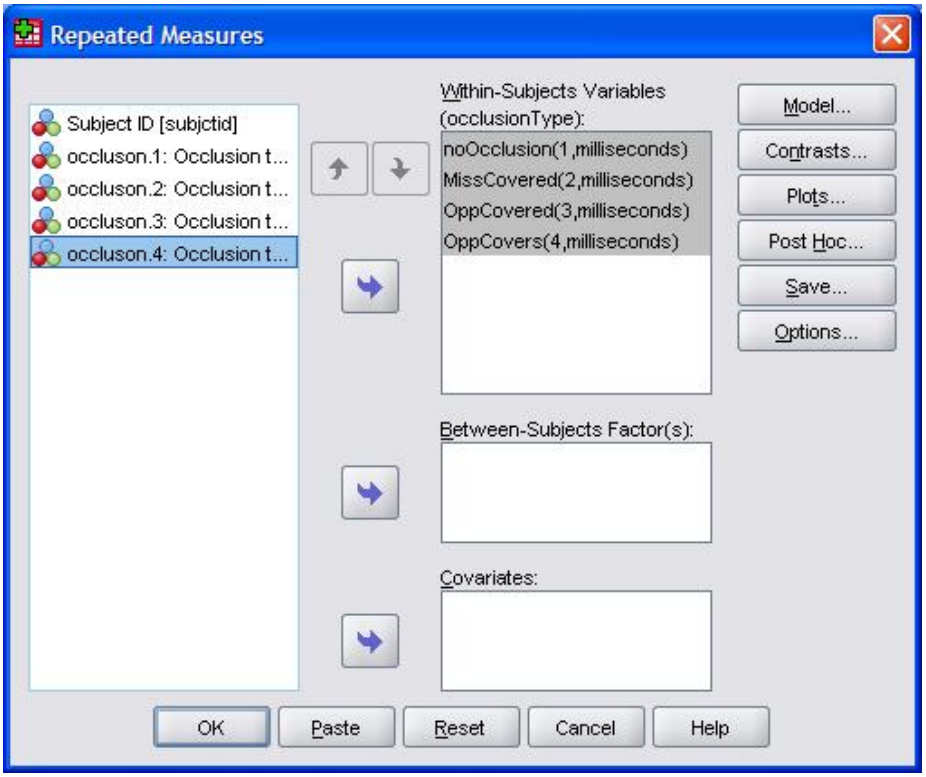


Figure 14.3: SPSS Repeated Measures dialog box.

Next you will see the Repeated Measures dialog box. On the left is a list of all variables, at top right right is the "Within-Subjects Variables" box with lines for each of the levels of the within-subjects variables you defined previously. You should move the k outcome variables corresponding to the k levels of the within-subjects factor into the "Within-Subjects Variables" box, either one at a time or all together. The result looks something like Figure 14.3. Now enter the between-subjects factor, if any. Then use the model button to remove the interaction if

desired, for a two-way ANOVA. Usually you will want to use the contrasts button to change the within-subjects contrast type from the default "polynomial" type to either "repeated" or "simple". If you want to do post-hoc testing for the between-subjects factor, use the Post-Hoc button. Usually you will want to use the options button to display means for the levels of the factor(s). Finally click OK to get your results.

Chapter 15

Mixed Models

A flexible approach to correlated data.

15.1 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 14, 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, whether intentional or unintentional. 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 14.1, 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, but rather are thought of as a random selection from a much larger set of levels. Subject effects are almost always random effects, while treatment levels are almost always fixed effects. 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 measurement occurring among all of the lower level units for each particular upper level unit.

15.2 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

the the data are in the tall format with one observation per row, and multiple rows per subject,

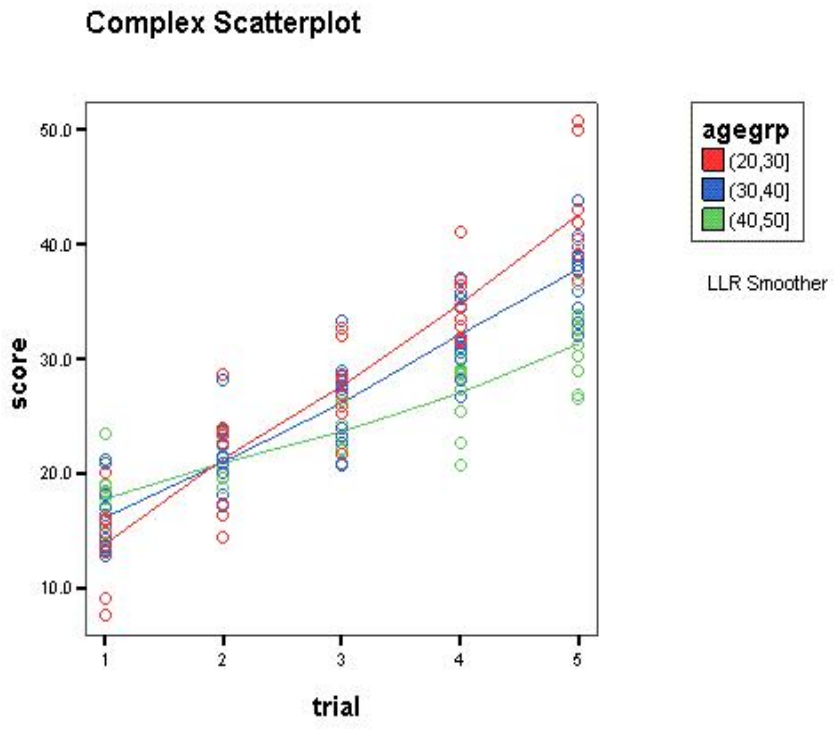


Figure 15.1: EDA for video game example with smoothed lines for each age group.

Some EDA is shown in figure 15.1. 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 as many subjects with most or all of their outcomes on the same side of their group's fitted line.

15.3 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 char-

acteristic of chapter 14. 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.

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 including the nature of the hierarchy, the fixed effects and the random effects.

In almost all situations several related models are considered and some form of model selection must be used to choose among related models.

The interpretation of the statistical output of a mixed model requires an understanding of how to explain the relationships among the fixed and random effects in terms of the levels of the hierarchy.

15.4 Analyzing the video game example

Based on figure 15.1 we should model separate linear relationships between trial number and game score for each age group. Figure 15.2, shows smoothed lines for each subject. From this figure, it looks like we need a separate slope and intercept 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

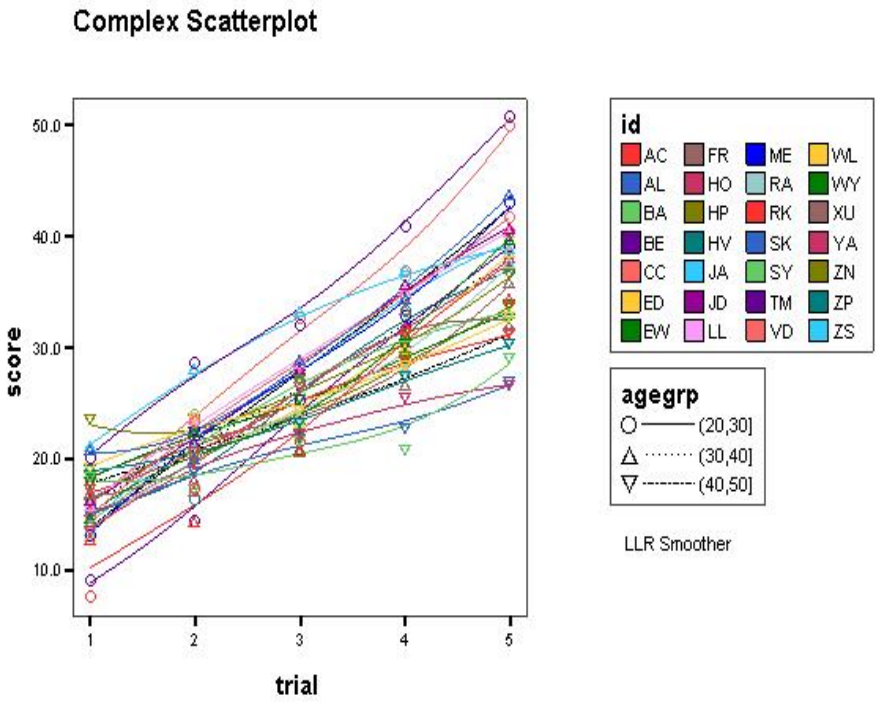


Figure 15.2: EDA for video game example with smoothed lines for each subject.

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.

15.5 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 15.3) 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.

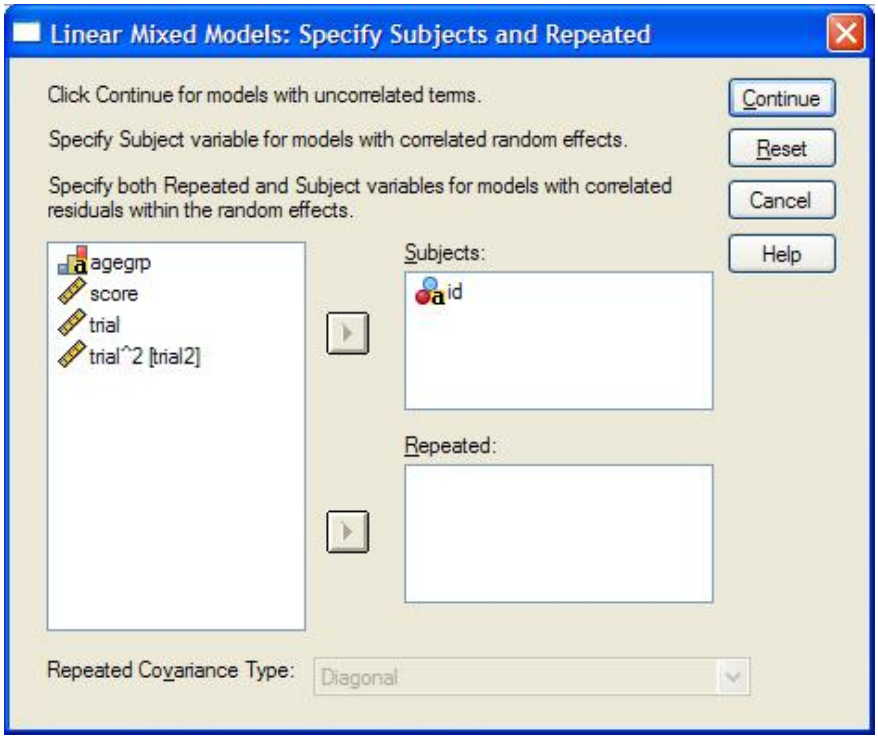


Figure 15.3: Specify Subjects and Repeated Dialog Box.

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.

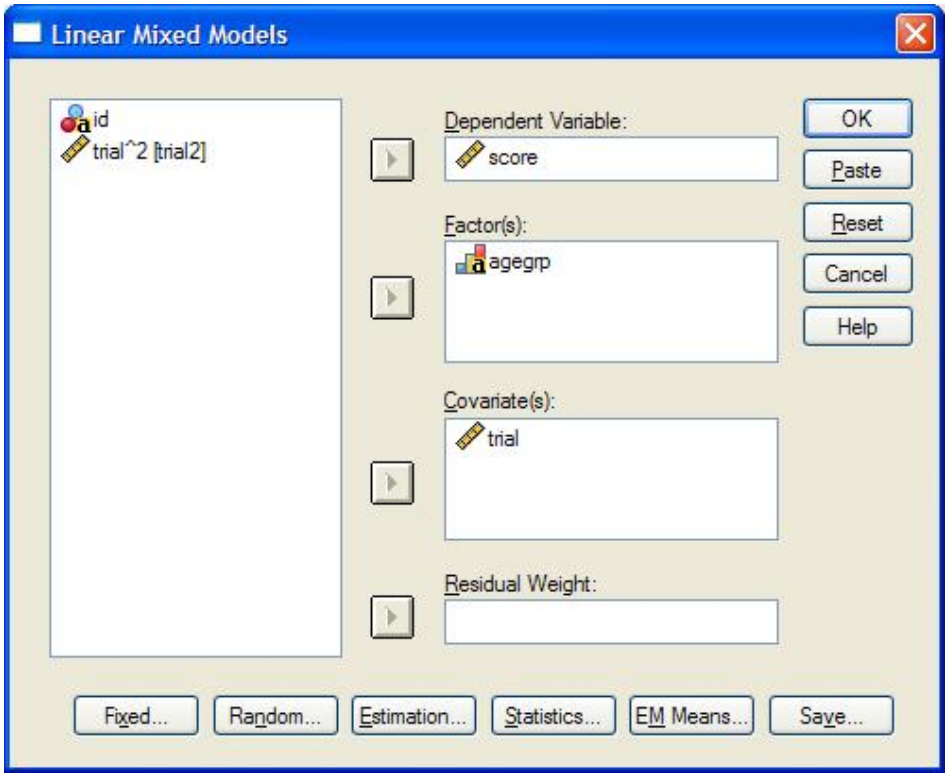


Figure 15.4: Main Linear Mixed Effects Dialog Box.

The main "Linear Mixed Models" dialog box is shown in figure 15.4. (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 15.5. 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 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.

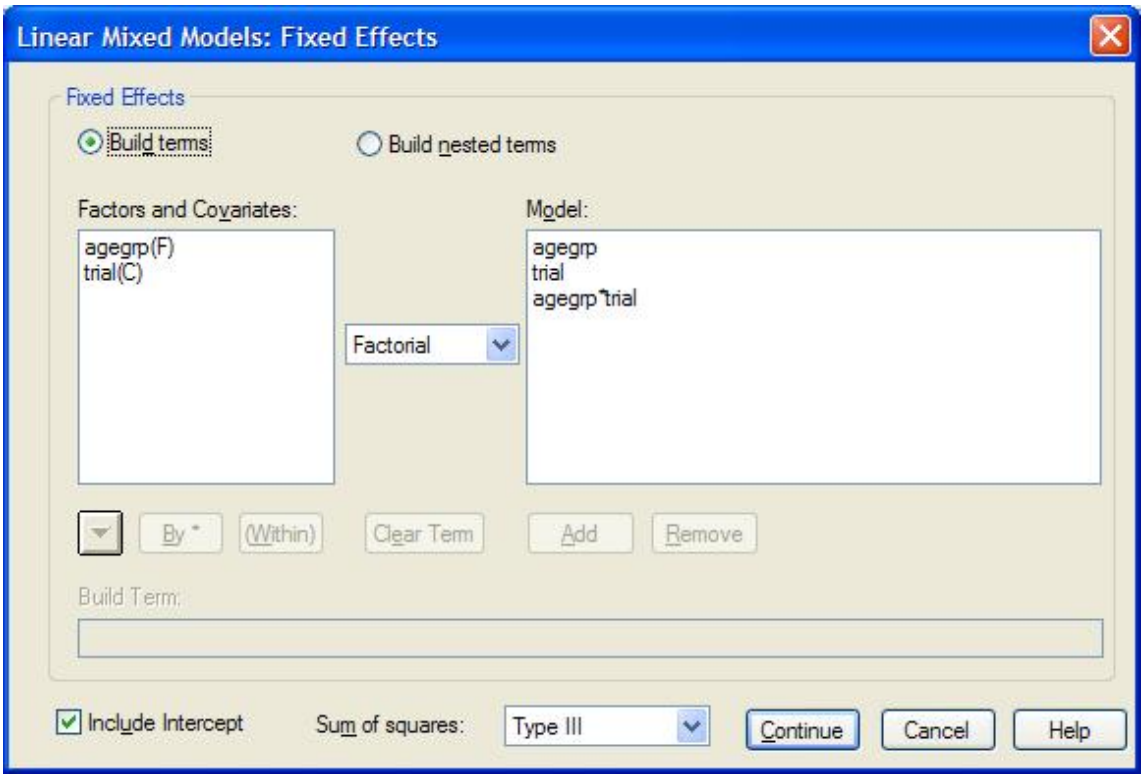


Figure 15.5: Fixed Effects Dialog Box.

For the video 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 co-

efficient 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 15.6), 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.

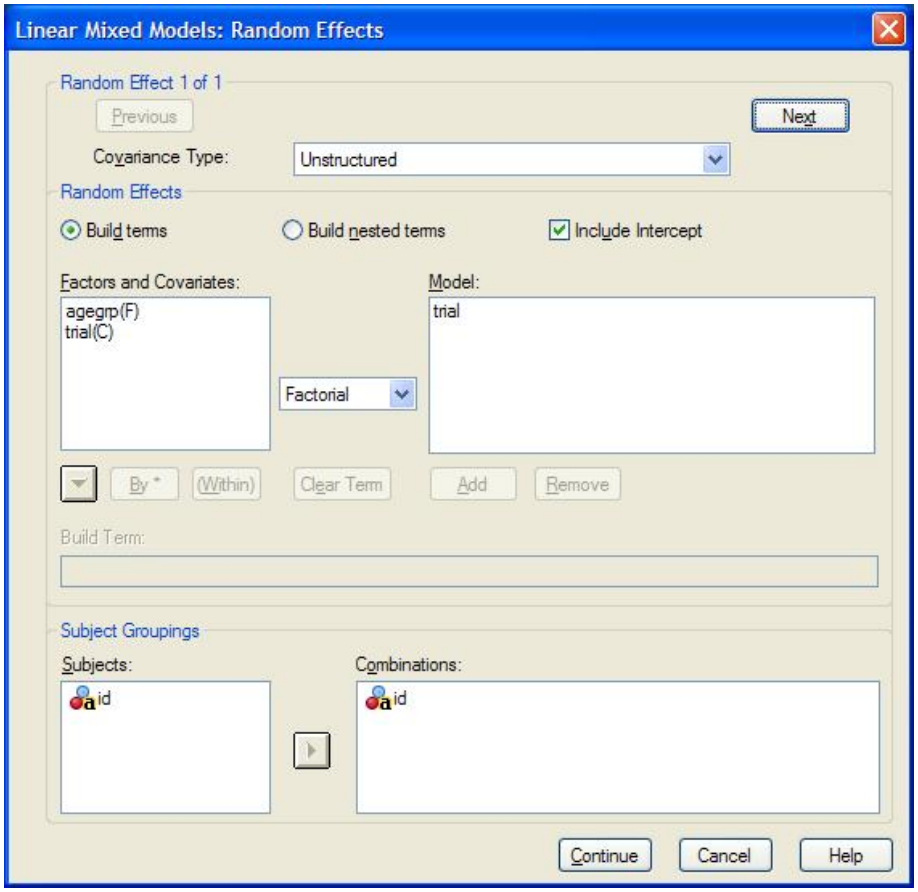


Figure 15.6: Random Effects Dialog Box.

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. We can look at it like this: 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 recommend 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 video 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, averages for each level of age group are really meaningless.

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.

15.6 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 15.1 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), but lines beyond the number of estimated parameters are essentially blank (with

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	agegrp	3		2	
	trial	1		1	
	agegrp * trial	3		2	
Random Effects	Intercept	1	Identity	1	id
Residual				1	
Total		9		8	

Table 15.1: Model dimension for the video game example.

parameters labeled as redundant and a period in the rest of the columns). We can see that we have a single random effect, which is an intercept for each level of id (each subject). 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. 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. See the section on model comparison for more about information criteria.

Next comes the Fixed Effects tables (tables 15.2 and 15.3). The tests of fixed effects has 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 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. We can see that age group 40-50 is the "baseline" (because SPSS chooses the last category). 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.

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 15.2: 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 15.3: Estimates of Fixed Effects for the video game example.

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 15.4). 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)

					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 15.4: Estimates of Covariance Parameters for the video game example.

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 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.

15.7 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.

15.7.1 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 15.5. 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.

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 15.5: BIC for model selection for the video game example.

The BIC approach to model selection is a good one, although there are some 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.

15.7.2 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, usually 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.

15.8 Classroom example

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. 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 placebo 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 15.7. 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 for its effect.

Here are some results:

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

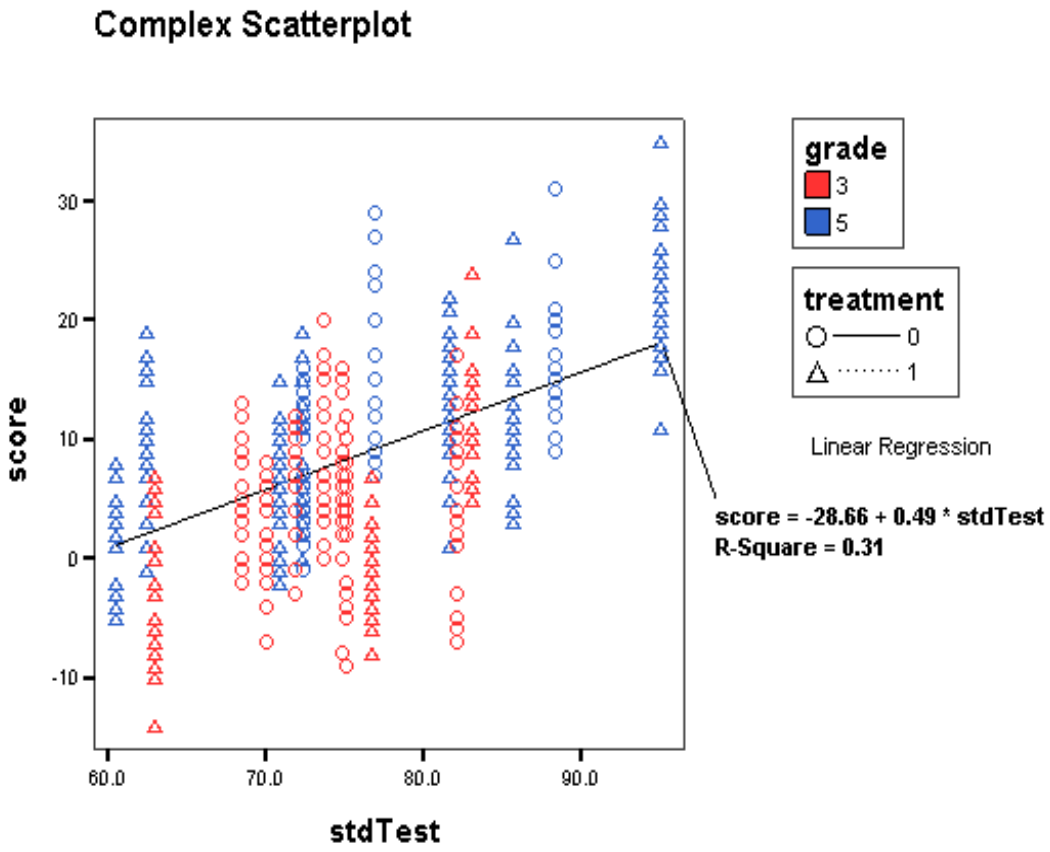


Figure 15.7: EDA for school example

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 15.6: 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 15.7: Estimates of Fixed Effects for the school example.

					95% Conf. Int.	
		Std.	Wald		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 15.8: Estimates of Covariance Parameters for the school example.

rise of 35 points.

The "Tests of Fixed Effects" table, Table 15.6, 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 15.7, 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 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).

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 15.7 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 15.8 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 large effect that either 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. (It would be good to center the stdTest values by subtracting their mean, then rerun the mixed model analysis; this would allow the Intercept to represent the average gain for a fifth grader with active treatment, i.e., the baseline group). 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 a data hierarchy.

Chapter 16

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.

16.1 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.

16.1.1 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 prediction equation $E(Y) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$, 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.

And for nominal categorical variables with more than two levels, the prediction is totally arbitrary and meaningless.

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.

16.2 Testing independence in contingency tables

16.2.1 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.

Note that if the outcome has only two possibilities then we only need the probability of one level of the variable rather than the full probability distribution (list of possible values and their probabilities) for each level of the explanatory variable. Of course, this is true simply because the probabilities of all levels must add to 100%, and we can find the other probability by subtraction.

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.

16.2.2 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 16.1 is a (fake) contingency table showing the results of asking 271 college students what their favorite music is and what their favorite ice

		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 16.1: Basic ice cream and music contingency table.

cream flavor is. This table was created in SPSS by using the Cross-tabs menu item under Analysis / Descriptive Statistics. 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 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 that 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

		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 16.2: Basic ice cream and music contingency table with row percents.

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 a form of EDA. We will need formal statistical analysis 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 row of the table, and **column percents** add to 100% (in the bottom margin) for each column of the table.

For example, table 16.2 shows the ice cream and music data with row percents. In SPSS the Cell button brings up check boxes for adding row and/or 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.

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 whether all of the row probabilities are similar to each other and to the marginal row probabilities. 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.

16.2.3 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 indicates that the null sampling distribution can not be computed exactly for a small sample size, but as the sample size increases, the null sampling distribution approaches the shape of a particular known distribution, which is the chi-square distribution in the case of the X^2 statistic. So 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 clear (or meaningful to ask) 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 16.3 shows the expected counts for the ice cream example. For example, using the percents in the bottom margin of table 16.2, if the two

		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 16.3: Expected counts for ice cream and music contingency table.

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 16.3. Note that there is no reason that the expected counts should be whole numbers, even though observed counts must be.

By combing the observed data of table 16.1 with the expected values of table 16.3, 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 conclusions 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.

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 16.4. 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 16.4: 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.

16.3 Logistic regression

16.3.1 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" and the other "failure". Also, conventionally, we use code 1 to represent success and code 0 to represent failure. Then we can look at logistic regression as modeling the success probability as a function of the explanatory variables. Also, for any group of subjects, the 0/1 coding makes it true that the mean of Y represents the observed fraction of successes for that group.

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: $\beta_0 +$

$x_1\beta_1 + \cdots + x_k\beta_k$. 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, while eta can take any value. The solution to this dilemma is to use the means model

$$\log\left(\frac{E(Y)}{1-E(Y)}\right) = \log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \eta.$$

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. The ratio of success to failure probabilities is called the odds. Therefore our means model for logistic regression is that the log of the odds (or just "log odds") of success is equal to the linear combination of explanatory variables represented as eta. In other words, for any explanatory variable j, if $\beta_j > 0$ then an increase in that variable is associated with an increase in the chance of success and vice versa.

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(S)=η, where S indicates success, i.e., Y=1.

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

$\Pr(Y=1)$	$\Pr(Y=0)$	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 16.5: Relationship between probability, odds and log odds.

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 16.5, 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 and denominator by whatever number gives whole numbers in the answer. In this case, we could multiple 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). Alternatively, in Windows calculator (scientific view) enter 1.099, then click the Inv (inverse) check box, and click the "ln" (natural log) button. (The "exp" button is *not* an equivalent calculation in Windows calculator.) For your handheld calculator, you should look up how to do this using 1.099 as an example.

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 prob-

ability 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 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 success 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.

The error model for logistic regression is that for each fixed combination of explanatory variables the distribution of success follows the binomial distribution, with success probability, p, determined by the means model.

16.3.2 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 Visual Binning in SPSS), and then to plot the mean of the explanatory variable in each bin vs. the

observed fraction of successes in that bin. Figure 16.1 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.

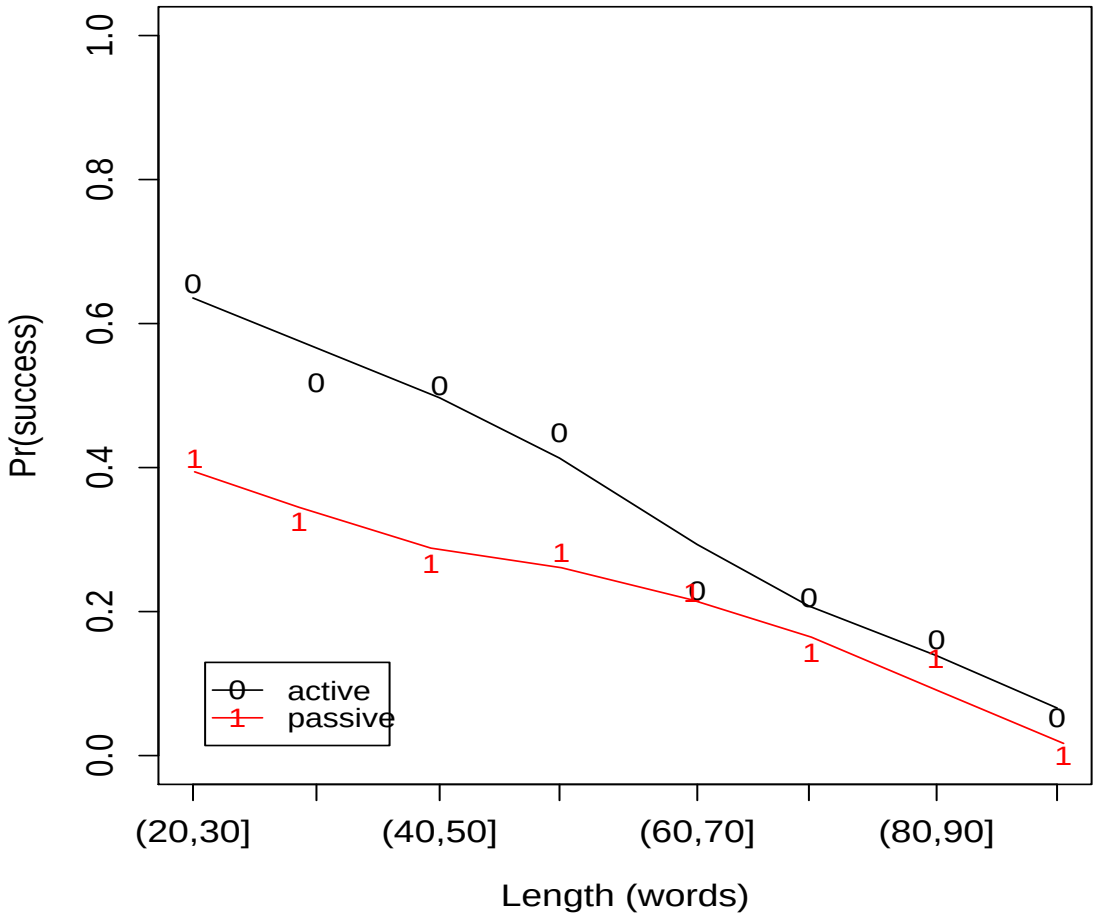


Figure 16.1: EDA for forum message example.

EDA for continuous explanatory variables can take the form of categorizing the continuous variable and plotting the fraction of success vs. failure, possibly separately for each level of some other categorical explanatory variable(s).

16.3.3 Fitting a logistic regression model

The means model in logistic regression is that

$$\text{logit(S)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

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 Pr(S)=1 as x_i increases and Pr(S)=0 as x_i decreases. For a negative β_i, the curve starts near Pr(S)=1 and falls toward Pr(S)=0. 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 some extreme values of that variable.

It is worth mentioning here that the magnitude of β_i is related to the steepness of the rise or fall, and the value of the intercept relates to where the curve sits left to right.

The fitting of a logistic regression model involves the computer finding the best estimates of the β values, which are called b or B values as in linear regression. Technically logistic regression is a form of generalized (not general) linear model and is solved by an iterative method rather than the single step (closed form) solutions of linear regression.

In SPSS, there are some model selection choices built-in to the logistic regression module. These are the same as for linear regression and include "Enter" which just includes all of the explanatory variables, "Backward conditional (stepwise)" which starts with the full model, then drops possibly unneeded explanatory variables one at a time to achieve a parsimonious model, and "Forward conditional

Dependent Variable Encoding

Original Value	Internal Value
Not a quick reply	0
Got a quick reply	1

Table 16.6: Dependent Variable Encoding for the forum example.

(stepwise)" which starts with a simple model and adds explanatory variables until nothing "useful" can be added. Neither of the stepwise methods is guaranteed to achieve a "best" model by any fixed criterion, but these model selection techniques are very commonly used and tend to be fairly good in many situations. Another way to perform model selection is to fit all models and pick the one with the lowest AIC or BIC.

The results of an SPSS logistic regression analysis of the forum message experiment using the backward conditional selection method are described here. A table labeled "Case Processing Summary" indicates that 500 messages were tested. The critical "Dependent Variable Encoding" table (Table 16.6) shows that "Got a quick reply" corresponds to the "Internal Value" of "1", so that is what SPSS is currently defining as success, and the logistic regression model is estimating the log odds of getting a quick reply as a function of all of the explanatory variables. *Always check the Dependent Variable Encoding.* You need to be certain which outcome category is the one that SPSS is calling "success", because if it is not the one that you are thinking of as "success", then all of your interpretations will be backward from the truth.

The next table is Categorical Variables Codings. Again *checking this table is critical* because otherwise you might interpret the effect of a particular categorical explanatory variable backward from the truth. The table for our example is table 16.7. The first column identifies each categorical variable; the sections of the table for each variable are interpreted entirely separately. For each variable with, say k levels, the table has k lines, one for each level as indicated in the second column. The third column shows how many experimental units had each level of the variable, which is interesting information but not the critical information of the table. The critical information is the final $k-1$ columns which explain the coding for each of the $k-1$ indicator variables created by SPSS for the variable. In our example, we made the coding match the coding we want by using the Categorical button and then selecting "first" as the "Reference category". Each

		Frequency	Parameter coding (1)
Male gender?	Female	254	.000
	Male	246	1.000
Passive voice?	Active voice	238	.000
	Passive voice	262	1.000

Table 16.7: Categorical Variables Codings for the forum example.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.597	8	0.800
2	4.230	8	0.836

Table 16.8: Hosmer-Lemeshow Goodness of Fit Test for the forum example.

of the $k-1$ variables is labeled "(1)" through "(k-1)" and regardless of how we coded the variable elsewhere in SPSS, the level with all zeros is the "reference category" (baseline) for the purposes of logistic regression, and each of the k-1 variables is an indicator for whatever level has the Parameter coding of 1.000 in the Categorical Variables Coding table. So for our example the indicators indicate male and passive voice respectively.

Correct interpretation of logistic regression results in SPSS critically depends on correct interpretation of how both the outcome and explanatory variables are coded.

SPSS logistic regression shows an uninteresting section called "Block 0" which fits a model without any explanatory variables. In backward conditional model selection Block 1 shows the results of interest. The numbered steps represent different models (sets of explanatory variables) which are checked on the way to the "best" model. For our example there are two steps, and therefore step 2 represents the final, best model, which we will focus on.

One result is the **Hosmer and Lemeshow Test of goodness of fit**, shown in Table 16.8. We only look at step 2. The test is a version of a goodness-of-fit chi-square test with a null hypothesis that the data fit the model adequately. Therefore, a p-value *larger* than 0.05 suggests an adequate model fit, while a small p-value indicates some problem with the model such as non-monotonicity, variance inappropriate for the binomial model at each combination of explanatory variables, or the need to transform one of the explanatory variables. (Note that Hosmer and Lemeshow have deprecated this test in favor of another more recent one, that is not yet available in SPSS.) In our case, a p-value of 0.836 suggests no problem with model fit (but the test is not very powerful). In the event of an indication of lack of fit, examining the Contingency Table for Hosmer and Lemeshow Test may help to point to the source of the problem. This test is a substitute for residual analysis, which in raw form is uninformative in logistic regression because there are only two possible values for the residual at each fixed combination of explanatory variables.

The Hosmer-Lemeshow test is a reasonable substitute for residual analysis in logistic regression.

The Variables in the Equation table (Table 16.9) shows the estimates of the parameters, 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.

16.3.4 Tests in a logistic regression model

The main interpretations for a logistic regression model are for the parameters. Because the structural model is

$$\text{logit}(S) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

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.

	B	S.E.	Wald	df	Sig.	Exp(B)
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
Constant	1.384	0.308	20.077	1	<0.005	3.983

Table 16.9: Variables in the equation for the forum message example.

The coefficient estimate results from the SPSS section labeled "Variables in the Equation" are shown in table 16.9 for the forum message example. It is this table that you should examine to see which explanatory variables are included in the different "steps", i.e., which means model corresponds to which step. Only results for step 2 are shown here; step 1 (not shown) indicates that in a model including all of the explanatory variables the p-value for "male" is non-significant (p=0.268).

This model's prediction equation is

$$\text{logit}(S) = \beta_0 + \beta_{\text{length}}(\text{length}) + \beta_{\text{passive}}(\text{passive})$$

and filling in the estimates we get

$$\widehat{\text{logit}}(S) = 1.384 - 0.035(\text{length}) - 0.744(\text{passive}).$$

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. As usual for an intercept, the interpretation of the estimate is meaningful if setting all explanatory variables to zero is meaningful and is not a gross extrapolation. Note that a zero log odds corresponds to odds of $e^0 = 1$ which corresponds to a probability of $\frac{1}{1+1} = 0.5$. Therefore it is almost never valid to interpret the p-value for the intercept (constant) in logistic regression because it tests whether the probability of success is 0.5 when all explanatory variables equal zero.

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. You should ignore the p-value for the intercept.

For a k-level categorical explanatory variable like "passive", SPSS creates $k-1$ indicator variables and estimates $k-1$ coefficient parameters labeled $B_{\text{x(1)}}$ through $B_{\text{x(k-1)}}$. In this case we only have $B_{\text{passive(1)}}$ because $k = 2$ for the passive variable. As usual, $B_{\text{passive(1)}}$ 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

$$\widehat{\text{logit}}(S) = B_{0L} - 0.744(\text{passive})$$

where $B_{0L} = 1.384 - 0.035L$ for some fixed message length, L. Then we exponentiate both sides to get

$$\widehat{\text{odds}}(S) = e^{B_{0L}} e^{-0.744(\text{passive})}.$$

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 $\widehat{\text{odds}}(S) = e^{B_{0L}}$ and for passive voice $\widehat{\text{odds}}(S) = 0.475e^{B_{0L}}$. 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.

So the usual way to interpret the effect of a categorical variable on a binary outcome is to look at "exp(B)" and take that as the multiplicative change in odds when comparing the specified level of the indicator variable to the baseline level.

If B=0 and therefore exp(B) is 1.0, then there is no effect of that variable on the outcome (and the p-value will be non-significant). If exp(B) is greater than 1, then the odds increase for the specified level compared to the baseline. If exp(B) is less than 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. This is just what we want so that we can keep the probabilities between 0 and 1. If we incorrectly claim that for each one-unit increase in x probability rises, e.g., by 0.1, then this becomes meaningless for a baseline probability of 0.95. But if we say that, e.g., the odds double for each one unit increase in x, then 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). Note that all new probabilities are valid, and that a doubling of odds corresponds to a larger probability change for midrange probabilities than for more extreme probabilities. This discussion also explains why you cannot express the interpretation of a logistic regression coefficient on the probability scale.

The estimate of the coefficient for an indicator variable of a categorical explanatory variable in a logistic regression is in terms of exp(B). 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, we can say that 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 constant. It is easier for people to understand if we change to the odds scale. Then exp(B) represents the *multiplicative* change in the odds of success for a one-unit increase in x with all other explanatory variables held constant.

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 or by 0.35 for each additional ten words. It is better to use exp(B) and say that the odds are multiplied by 0.966 (making them slightly smaller) for each additional word.

It is even more meaningful to describe the effect of a 10 word increase in message length on the odds of success. Be careful: you can't multiply exp(B) by ten. There are two correct ways to figure this out. First you can calculate $e^{-0.35} = 0.71$, and conclude that the odds are multiplied by 0.71 for each additional ten words. Or you can realize that if for each additional word, the odds are multiplied by 0.966, then adding a word ten times results in multiplying the odds by 0.966 ten times. So the result is $0.966^{10} = 0.71$, giving the same conclusion.

The p-value for each coefficient is a test of $\beta_x = 0$, and 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(B) is the multiplicative change in odds of success for a one-unit increase in the explanatory variable.

16.3.5 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.

For our forum messages example, 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 16.2 which is a nice summary of the means model.

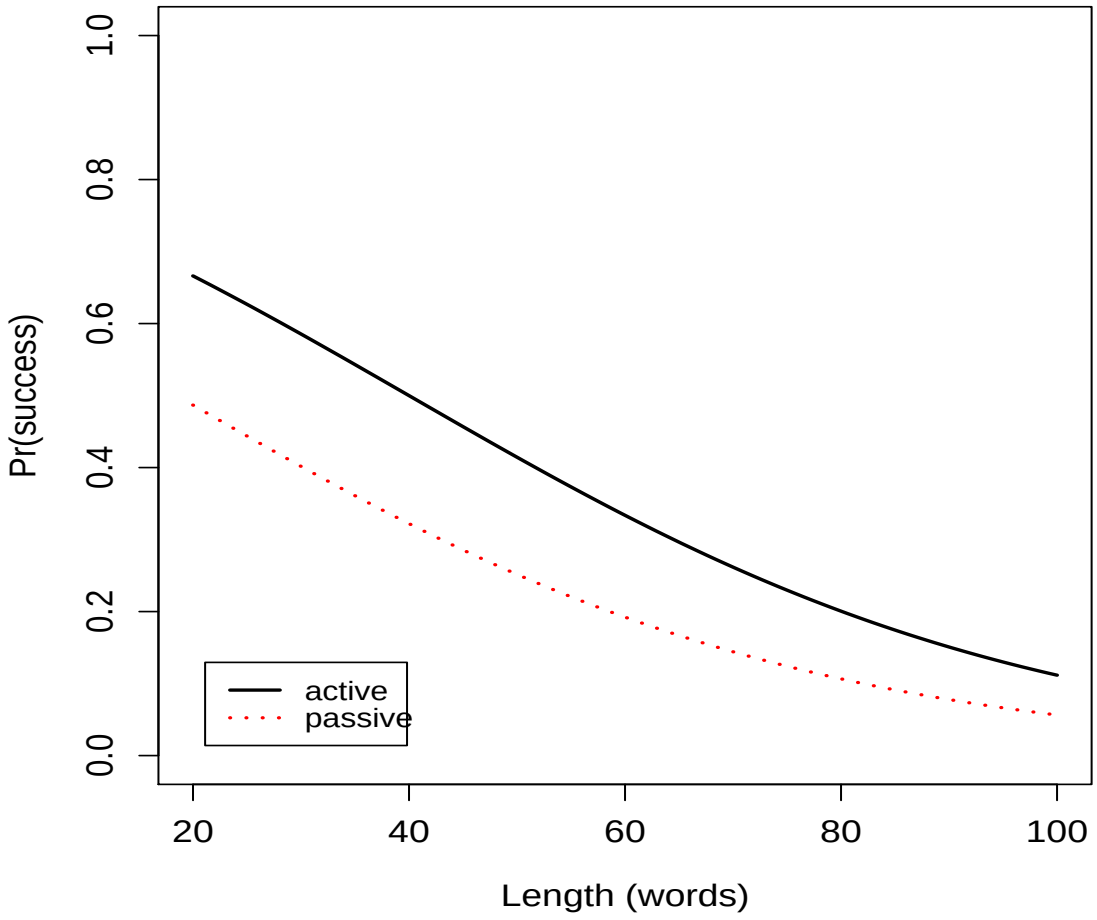


Figure 16.2: Model predictions for forum message example.

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.

16.3.6 Do it in SPSS

In SPSS, Binary Logistic is a choice under Regression on the Analysis menu. The dialog box for logistic regression is shown in figure 16.3. Enter the dependent variable. In the "Covariates" box enter both quantitative and categorical explanatory variables. You do *not* need to manually convert k-level categorical variables to indicators. Select the model selection method. The default is to "Enter" all variables, but you might want to switch to one of the available stepwise methods. You should always select "Hosmer-Lemeshow goodness-of-fit" under Options.

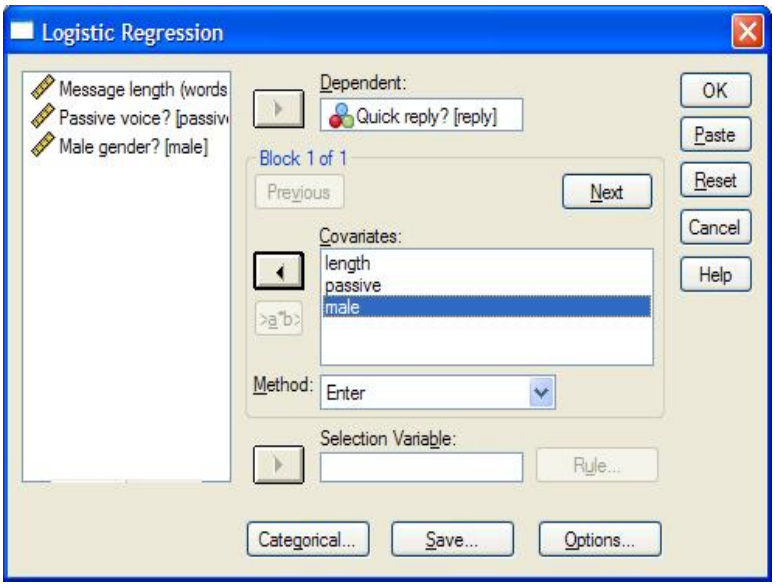


Figure 16.3: SPSS dialog box for logistic regression.

If you have any categorical explanatory variables listed in the "Covariates" box, click on "Categorical" to open the dialog box shown in figure 16.4. Move only the

categorical variables over to the "Categorical Covariates" box. The default is for SPSS to make the last category the baseline (reference) category. For variables that are already appropriately named indicator variables, like passive and male in our example, you will want to change the "Reference Category" to "First" to improve the interpretability of the coefficient tables. Be sure to click the "Change" button to register the change in reference category.

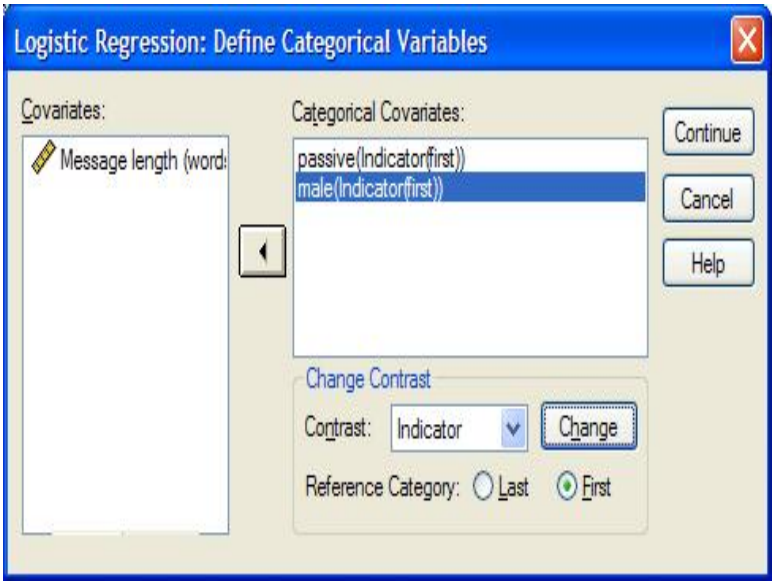


Figure 16.4: SPSS Categorical Definition dialog box for logistic regression.

The interpretation of the SPSS output is shown in the preceding sections.

Chapter 17

Going beyond this course

Index