# Fundamentals of Research Design and Statistical Analysis

**John Ruscio**
The College of New Jersey

**Draft Date**
January 21, 2020

## Goals and Contents

This text is designed to assist students taking introductory-level courses in research methodology and statistics as well as students engaged in other research experiences that involve data analysis. This is not copyrighted material and you may save, print, copy, or distribute it. The material is organized into 21 chapters and 5 appendixes, as listed below. Coverage includes how to use SPSS to perform data analyses and how to report results in APA style. Comments, questions, and suggestions are warmly welcomed; please write to ruscio@tcnj.edu.

# 1. Basic Concepts

## Overview

Research design and data analysis go hand in hand. Choosing the most appropriate statistical tests depends on how a study was designed. For example, if there are separate groups of subjects representing different experimental conditions, known as a **between-subjects** design, scores must be compared using an **independent groups** statistical procedure. On the other hand, if subjects were tested repeatedly in all experimental conditions, known as a **within-subjects** design, scores must be compared using a **related samples** statistical procedure. In other words, there is a close correspondence between the kinds of research designs used to collect data and the kinds of statistical procedures used to analyze data.

Another link between design and analysis emerges from the fact that no statistic is self-interpreting. Drawing conclusions from statistical results requires careful consideration of how the data were collected. For example, finding that a treatment group scores better than a control group does not necessarily mean that the treatment works. There might be a design flaw, such as failure to control for the well-known **placebo effect** by which the power of suggestion itself leads to improvement, that caused the statistical difference observed between groups. Data analysis is not a mechanical procedure that will reveal whether a hypothesis is true or false. Statistics should help you reach informed conclusions, but evaluating hypotheses requires you to make judgments that go beyond the statistics alone.

In this chapter, we'll explore some of the basic concepts of research design and statistical analysis to set the stage for all that follows. To help make these concepts concrete, an illustrative study will be introduced.

The InnerChange Freedom Initiative (IFI) was an extensive, faith-based program designed to reduce criminal recidivism, the commission of crimes by individuals released from prison. In 2000, the results of research evaluating the effectiveness of IFI were distributed. The full report is available online[1], and excerpts from the Executive Summary are quoted below:

> *… The InnerChange Freedom Initiative (IFI) … is a largely volunteer driven program … the first full-scale attempt to offer comprehensive programming emphasizing education, work, life skills, values restructuring, and one-on-one mentoring in an environment where religious instruction permeates all aspects of the prison environment.*
>
> *… this study tracks the two-year post-release recidivism rates for those prisoners that entered the IFI program from April of 1997 through January of 1999, and were released from prison prior to September 1, 2000…*
>
> *… A total of 177 participants … formed the basis of the IFI study group. Comparison groups were selected from the records of inmates released during the evaluation period that met program selection criteria but did not enter the program. The comparison*

[1] http://web.archive.org/web/20030805160548/http://www.crrucs.org/8_research_pdf/innerchange_freedom.pdf

*groups were matched with IFI participants based on the following characteristics: race, age, offense type, and salient factor risk score. A total of 1,754 inmates were identified as the main comparison group for this study.*

*Anchored in biblical teaching, life-skills education, and group accountability, IFI is a three-phase program involving prisoners in 16 to 24 months of in-prison programs and 6 to 12 months of aftercare following release from prison…*

*Among the study's key findings are the following:*

*1. The IFI participants in this study include 75 prisoners who completed all phases of the program (called IFI Graduates), 51 who were paroled early, 24 who voluntarily quit the program, 19 who were removed for disciplinary reasons, 7 who were removed at the request of the staff, and 1 who was removed for serious medical problems …*

*2. 17.3% of IFI program graduates and 35% of the matched comparison group were arrested during the two-year post-release period. A program graduate is someone who completes not only the in-prison phases of IFI dealing with biblical education, work, and community service (usually lasting 16 months), but also includes an aftercare phase (usually lasting 6 months) in which the participant must hold a job and have been an active church member for 3 consecutive months following release from prison.*

*3. 8% of IFI program graduates and 20.3% of the matched comparison group were incarcerated during the two-year post-release period.*

*4. Considering all participants, including those inmates who did and did not complete all phases of the program, 36.2% of IFI participants were arrested compared to 35% of the matched group during the two-year tracking period. Among the total number of IFI participants, 24.3% were incarcerated compared to 20.3% of the comparison group during the two-year post-release period.*

## Exploratory vs. Hypothesis-Testing Research

Research can be designed either to develop ideas or to test them, and it is usually not possible to address both of these goals in a single study. The goal of **exploratory research** is to develop ideas. This is done by collecting information on a wide range of variables, perhaps with very little experimental control, and then examining the data in many ways to search for interesting patterns. One need not have any hypotheses to perform useful exploratory research, nor is strong evidence required to raise the possibility that observed trends may be worthy of further testing. Findings are tentative, and follow-up research is needed to replicate and better understand them.

The goal of **hypothesis-testing research**, which is sometimes referred to as **confirmatory research** to contrast it with exploratory research, is to subject ideas to rigorous tests. This is done by designing a study such that the evidence will either support or refute a hypothesis. This entails careful experimental control, collecting a large sample of data on a focused selection of variables, performing demanding statistical tests, or other techniques to help rule out alternative explanations for results. Findings that survive this rigorous testing can be trusted with greater confidence that they are correct.

The IFI study is an example of hypothesis-testing research. The investigators had already developed some ideas for how to reduce criminal recidivism and they wanted to test whether their faith-based program worked.

## Experimental vs. Correlational Research

In **experimental research**, the investigator manipulates one variable (e.g., treatment vs. placebo conditions) and tests for differences on an outcome variable. Often this is done through **random assignment** to experimental conditions. Provided that the sample is large enough, all variables other than the one manipulated by the experimenter should cancel out across conditions. The goal of experimental research is to draw causal conclusions, to test whether the manipulated variable causes changes in the outcome variable.

In **correlational research**, variables are measured rather than manipulated. Sometimes investigators study variables that cannot be manipulated (e.g., sex, age, race), and sometimes subjects volunteer for different conditions (e.g., patients who seek different types of psychotherapy). In both of these cases, the research is correlational rather than experimental. The goal of correlational research is to examine relationships among variables, and it's critical to keep in mind that correlation does not equal causation.

The IFI study included a treatment group and a matched comparison group, so in a superficial sense it might appear to be experimental research. However, prisoners volunteered for the IFI treatment, and the comparison group was formed through the review of records for other prisoners who did not volunteer to participate in the study. There was no random assignment to experimental conditions, and there are any number of ways that these groups could have differed besides the IFI treatment. Therefore, this is an example of correlational research.

## Populations and Samples

A **population** is all individuals of interest to a researcher, those to whom we would like to generalize the results of a study. A **sample** is the individuals that were actually included in the study. A truly **random sample**, in which every member of a well-defined population has an equal chance of being included in the sample, would be **representative** of its population.

It's extremely rare for researchers to obtain truly random samples. For example, it might be impossible to list all members of a population, to select people at random, or to obtain consent from everyone who is selected. The extent to which a sample differs systematically from a population is referred to as **sample bias**. Some samples are less biased, and therefore more representative, of a population than other samples. Very often, researchers use a **convenience sample** such as students enrolled in introductory psychology courses. As the name implies, convenience samples can make life easier for researchers. However, such samples are seldom very representative of the populations of interest.

In the IFI study, the population wasn't defined explicitly. This is common. One might infer that the population would be all prisoners. The sample consisted of the 177 program volunteers as well as the 1,754 members of the matched comparison group, for a total of 1,931 individuals. The extent to which this is a representative sample is debatable. On the plus side, these were actual prisoners rather than students or another more convenient sample. On the other hand, the treatment group consisted entirely of volunteers for a faith-based program, and they may not be very representative of all prisoners. In addition, because the sample was drawn from a single prison, it cannot represent the entire

population of all prisoners (e.g., it cannot simultaneously represent male and female inmates or those at minimum-, medium-, and maximum-security prisons).

## Parameters and Statistics

**Statistics** are numerical values that summarize a sample of data. **Parameters** are the corresponding numerical values for the population. By way of analogy, we can say that sample : statistic :: population : parameter.

Because we seldom, if ever, have access to data for every member of a population, parameters typically remain unknown. Using a sample of data, we calculate statistics to estimate parameters. The difference between the true value of a parameter and the observed value of a statistic is called **sampling error**. All statistics are subject to some degree of sampling error, and the goal is to reduce this to a minimum. The best way to do so is to collect as much data as possible because the main determinant of sampling error is sample size. The larger the sample, the less sampling error.

Usually, sampling error decreases as a function of the square root of sample size. For example, suppose you've polled $N = 100$ people to see how many plan to vote for candidate *X* in an upcoming election. This sample size would give a margin of error of about ±10% when estimating the parameter (what you want to know, the percentage of people in the population who plan to vote for candidate *X*) from your statistic (the percentage of people who reported this intention in your sample). To cut this margin of error in half, to ±5%, you'd need to increase the size of your sample to $N = 400$, not just $N = 200$. Four times as much data (400 / 100 = 4) yields twice as much precision (±10% / ±5% = 2).

In the IFI study, the most important statistics were the percentage of individuals in each group who were rearrested and reimprisoned during the two-year period following their release from prison. These statistics would provide pretty good estimates of their corresponding parameters because the size of each group was fairly large ($n = 177$ for the treatment group) to very large ($n = 1{,}754$ for the comparison group).

## Error and Bias

In everyday language, the terms "error" and "bias" are very similar. In research methods and statistics, they have distinct meanings. **Error** refers to something that is random, and **bias** refers to something that is systematic.

Sampling error illustrates a kind of random deviation. Assuming that one has a representative sample, there is an equal chance that any statistic calculated from the data would be an overestimate or an underestimate of the parameter. We don't expect the measured height of a random sample of college students (the statistic) to be a perfect estimate of the average for the entire population of students at that school (the parameter). There will be some sampling error, due entirely to the luck of the draw in who happens to be in the sample. However, we also don't expect our statistic (the sample average) to diverge from the parameter (the population average) in a systematic way. The fact that it's equally likely that the sample of students is taller or shorter than the population of all students demonstrates that sampling error is random.

Sample bias illustrates a kind of systematic deviation. When a sample is not representative of a population, statistics can be expected to depart from parameters in a systematic manner. For example, if our sample of students was drawn from psychology classes, it would probably include a larger proportion of women than in the school-wide population of all students. In this case, we would expect the average height in the sample to be lower than the population average. Sample bias allows you to predict the direction of the difference between a statistic and a parameter.

The simple point to remember is that "error" is random and "bias" is systematic.

## Descriptive and Inferential Statistics

**Descriptive statistics** are used to summarize data. Often, this involves indicating what are typical scores and how much variation there is in the sample. **Inferential statistics** are used to test hypotheses, to reach conclusions extending beyond a sample of data.

In the IFI study, the percentage of individuals in each group who were rearrested or reimprisoned were descriptive statistics. Though these figures differed across groups, it's impossible to tell just by looking at them whether the differences were more than sampling error alone could explain. In other words, the apparent differences between the groups in the sample may not correspond to actual differences between groups in the population. To test this would require inferential statistics, which were not presented in the executive summary.[2] Inferential statistics require us to determine how large a difference we would expect to occur by chance alone. This provides a context for judging whether or not we believe the observed difference provides compelling evidence that the treatment had some effect.

## Independent and Dependent Variables

An **independent variable** is either manipulated or measured, and it is used to predict scores on a **dependent variable**. Some sources differentiate between "true" independent variables, which are strictly manipulated in experimental designs, and **subject variables** that are measured. We will adopt the convention used in most statistics texts by using the term "independent variable" more inclusively.

Whether manipulated or measured, independent variables are conceptualized as causal factors. Dependent variables are conceptualized as outcomes. By way of analogy, we can say that independent variable : cause :: dependent variable : effect.

In the IFI study, the only independent variable was group membership, indicating whether each individual was in the treatment group or the matched comparison group. The primary dependent variables were rearrest and reimprisonment. The researchers wanted to test the hypothesis that members of the treatment group would have lower rates of rearrest and reimprisonment than members of the comparison group.

[2] We'll revisit these data to calculate the appropriate inferential statistics in a later chapter.

## Conceptual and Operational Definitions

Whereas the **conceptual definition** of a variable captures its essence and serve as a shorthand for ease of communication, the **operational definition** of a variable indicates precisely how it's measured in a particular study. For example, a researcher might want to study aggression. That's an important concept, and surely eveyone has some idea what aggression means. At the same time, it's pretty vague. Does the researcher have in mind physical aggression (e.g., hitting or kicking someone), verbal aggression (e.g., taunting or insulting someone), or relational aggression (e.g., excluding someone from a group). Even these are conceptual distinctions, and a complete operationalization of the variable requires specifying in detail how aggression was measured.

In the IFI study, the conceptual definition of the dependent variable was recidivism, the commission of crimes after release from prison. This was operationalized in two ways, both of which involved a period of two years following release from prison. Specifically, the investigators coded whether or not an individual was rearrested or reimprisoned during the post-release period. Each individual received a score of "yes" or "no" on each of these variables.

## Measurement Scales

The kinds of statistical analyses that we can perform are determined in part by the types of data we have collected and the ways that the variables were measured. The standard typology includes four scales of measurement: Nominal, ordinal, interval, and ratio.

### Nominal Scale

The simplest type of data consists of categories that cannot be placed into any meaningful order. There may be only two categories (e.g., male, female) or more (e.g., marital status, classified as married, divorced, separated, widowed, or never married). These variables are measured using a **nominal** scale. "Nominal" refers to the fact that categories can only be named, not organized further.

### Ordinal Scale

The next type of data consists of values that can be rank-ordered. For example, artworks can be subjectively rated as high, moderate, or low in creativity, and competitors in a race can be scored as finishing 1st, 2nd, 3rd, and so forth. These variables are measured using an **ordinal** scale. "Ordinal" refers to the fact that scores can be arranged in order, even though the actual differences between neighboring scores may be highly uneven (e.g., a smaller gap between 1st and 2nd place than between 2nd and 3rd place).

### Interval Scale

The next type of data consists of values that can be ranked and for which the differences between neighboring values are equivalent. For example, IQ tests are constructed such that the difference between scores of 90 and 100 is equivalent to the difference between scores of 100 and 110. This variable is measured using an **interval** scale. "Interval" refers to the fact that the gaps, or intervals, between scores are equivalent along the scale.

### Ratio Scale

The final type of data not only has the property of equal intervals, but also that there exists a true zero point. For example, physical measurements such as height and weight can be used to form ratios. Even though nobody can actually have a height or weight of zero, this value exists and is easy to conceptualize. The same cannot be said for a variable like IQ. No matter how we define intelligence, there is no such thing as its complete absence in any person. As much as we might be tempted to say that someone is twice as smart as someone else, this isn't meaningful in the same way that we can say that this adult is twice as tall as that child, or that one person is two-thirds the weight of another. Variables such as height or weight are measured using a **ratio** scale. "Ratio" refers to the fact that ratios of one score to another are meaningful.

In the IFI study, each of the variables was measured using a nominal scale. The independent variable consisted of membership in either the treatment or matched comparison group. The dependent variables consisted of scores of "yes" or "no" for rearrest and reimprisonment. It's important to differentiate between the dependent variables themselves and the statistical summary of the data. For an individual in this study, there were only two possible scores on each of the outcome measures—someone either was or was not rearrested, and either was or was not reimprisoned—and that's why these are nominal data. When the scores on the dependent variables are summarized, the rates could vary from 0% to 100%. That still doesn't make this quantitative data. The type of data depends only on how individuals' scores are scaled, not on how we later summarize the data for a larger group.

## Threats to Internal Validity

**Internal validity** is the extent to which a causal conclusion can be drawn from a study's findings. A study has strong internal validity when it has been designed to minimize **confounds**, sources of bias that lead to alternative explanations for the results. There are many potential threats to internal validity. Some of these are fairly unique to particular studies, but a handful are among the most commonly occurring problems. In this section, we'll review the threats to internal validity that are most important to consider.

In thinking about internal validity, it's critical to pay close attention to whether a potential confound results in a systematic bias across conditions in a study. If an uncontrolled influence would be expected to have different effects on outcomes in different conditions, that is a threat to internal validity because any observed differences across conditions may be attributable to the confound. On the other hand, if an uncontrolled influence would be expected to have the same effect on outcomes in different conditions, this is not a threat to internal validity. It does not pose an alternative explanation for any observed differences across conditions. For example, because some amount of sampling error is always present, statistics never estimate parameters with perfect precision. This doesn't pose a threat to internal validity, though, because it's a source of error rather than bias. In other words, sampling error adds random noise to the analysis, but it doesn't bias the findings in favor of one condition relative to another.

Many threats to internal validity can be prevented by careful research design in which the outside influence is not eliminated, but made to cancel out across conditions. For

example, **random assignment** of subjects to experimental conditions can be helpful to equate groups in many ways. Individual differences that can influence the dependent variable are not eliminated by randomization, but they are equated across conditions. Random assignment tends to be more effective as sample size increases because this makes it more likely that individual differences will cancel out across conditions. With very small samples, groups might still differ in important ways despite randomization.

### History

Events that take place between measurements in a study, but that are not related to the independent variable(s) under investigation, can bias the findings. Apparent changes over time might result from **history** effects rather than the operation of other causal influences being studied. For example, suppose that you happened to be studying changes in anxiety as children age, and the terrorist attacks of September 11 occurred right in the middle of your study. It might be impossible to differentiate the normal developmental trajectory of anxiety from changes caused by this outside event.

When comparing groups, randomly assigning subjects to conditions controls history effects because they should affect outcomes in each condition to the same extent. Even though random assignment does not reduce, let alone eliminate, history effects, it does mean that they should cancel out across conditions and no longer pose a threat to internal validity.

### Maturation

Over long periods of time, subjects tend to grow older, wiser, stronger, and healthier, and over short periods of time they can become tired, bored, and so forth. Apparent changes over time may be attributable to **maturation** effects. For example, performance on a test given early in an experimental session may be superior to performance on a later test not because of real differences in ability assessed by these tests, but because subjects are less attentive or motivated toward the end of a lengthy session.

As with history effects, random assignment to conditions will not eliminate maturation effects, but it can cause them to cancel out across conditions. When each subject will participate in more than one condition in a study, another useful design strategy is to randomize the order of conditions. This technique is called **counterbalancing**, and it can control maturational influences by ensuring that they will affect outcomes in each condition to the same extent. Like random assignment, counterbalancing doesn't eliminate maturation effects but it can cancel them out across conditions.

These strategies are not mutually exclusive because random assignment deals with differences between two or more groups (a between-subjects component of a design) and counterbalancing deals with differences within each group (a within-subjects component of a design). If your study's design includes both between- and within-subjects components, you can randomly assign subjects to different groups and counterbalance the order of conditions within each group. Both random assignment and counterbalancing are desirable features whenever they can be used, alone or in combination.

### Instrumentation

When the nature or process of measurement differs either across conditions or over the course of a study, this can bias the results. Problems of **instrumentation** can involve differences in definitions, scoring rules, rating criteria, or the functioning of equipment.

Sound research designs will include precautions to ensure that measurement techniques remain constant across conditions and duration of a study. For example, the agreement of independent raters should be checked periodically to ensure that they continue to apply the same criteria when observing behavior. Straying from the initial criteria is known as **rater drift**. Similarly, any device used to present stimuli or record responses should be tested regularly to ensure that it is functioning properly.

### Selection

Individuals in different conditions may differ from one another at the outset of a study in ways that are confounded with the independent variable. For example, suppose that a researcher compares the earnings of husbands and wives to test for gender differences. There is a **selection** problem here, namely a confound with age. Husbands tend to be a few years older than their wives, and earnings also tend to increase with age. If husbands earn more, it would be impossible to tease apart the influences of gender and age as causes.

Selection effects are especially problematic when subjects have not been randomly assigned to conditions. When people choose their own conditions (e.g., seeking vs. not seeking counseling for a psychological problem), or when people bring pre-existing differences to a study (e.g., different levels of self-esteem), there may be many plausible alternative explanations for any differences observed across groups.

Random assignment to conditions yields groups that should not systematically differ. As we have seen, this neither eliminates nor reduces the influence of individual differences, but it should cancel them out across conditions. When subjects cannot be randomly assigned, measuring potential confounding variables allows you to at least include them in the analysis to test for their influence. For example, recording the ages of husbands and wives would enable you to perform analyses that test for both gender and age differences.

### Mortality

Over the course of a study, some subjects may die, drop out, or refuse to continue their participation. This threat to internal validity is known as **mortality**. Because the potential bias is due to missing data and not necessarily death, this threat is also known as **dropout** or **attrition**.

If mortality is caused by factors unrelated to the focus of a study, it reduces sample size but may pose no threat to internal validity. For example, in a longitudinal study some subjects may move away and fail to provide new contact information. If it can be assumed that those who move do not differ systematically from those who do not move on any of the variables being studied, then this is a random source of missing data rather than a bias confounding the results.

If mortality is caused by factors related to the focus of a study, and in particular if there is **differential dropout** across conditions, this can pose a threat to internal validity. For example, in a study of a new drug treatment using a treatment vs. placebo group design, if those who experience side effects of the treatment refuse to continue taking it and drop out of the study at a higher rate than those who received the placebo, this will bias the results.

The drug may appear beneficial only because those who had negative reactions dropped out of the study, leaving only those who had positive reactions in the treatment group. Alternatively, in a study of a powerful drug, subjects who experience no side effects of any kind may suspect that they're in the placebo group. Not only does this reduce the value of having a placebo group to control for the power of suggestion, but it also might lead some subjects in that group to discontinue their participation because they believe it's a waste of their time.[3]

Precautions against mortality include random assignment to conditions (so that dropout should be equalized across groups), minimizing the number of follow-up measurements, and taking steps to stay in touch with subjects and motivate them to return for follow-up sessions. If the concern is that attrition might yield too small a sample of data, which threatens the power of statistical tests but not the internal validity of the study, an investigator might want to begin with an especially large sample size to allow for substantial dropout.

### Reactivity

The act of measuring or observing behavior can influence that behavior. One common source of such **reactivity** is the repeated testing of subjects, because practice at or knowledge of previous tests can affect subsequent performance. Reactivity is also known as a **testing** or **repeated testing** problem.

Likewise, simply knowing what a researcher is studying can influence subjects' behavior. **Experimenter bias** refers to any influence on subjects—intentional or otherwise—that biases the results in support of the researcher's hypothesis. **Demand characteristics** are subtle cues provided to subjects that hint at expected responses or behaviors.

Several kinds of precautions can be taken to guard against reactivity. In a **blind** study, subjects do not know what condition they were assigned to, thereby eliminating the possibility that such knowledge could affect their behavior. In a **double-blind** study, neither the subjects nor the experimenters who interact with them know who was assigned to which condition. This further reduces the potential for bias because experimenters cannot leak information about subjects' assignment to conditions, not even accidentally.

As we'll see, there are many important decisions that must be made to analyze data in the most appropriate manner. How the data analyst makes critical choices can affect the results. For this reason, there is even a **triple-blind** procedure in which neither the subjects, the experimenters, nor the data analysts know who was assigned to which condition in a study. For example, data analysts can be asked only to compare outcomes for conditions A and B, without knowing which is the treatment and which the control group.

Other precautions against reactivity include training experimenters to avoid biases or demand characteristics, minimizing the number of times that the same tests are used, and observing subjects covertly rather than overtly.

---

[3] For this reason, researchers often use a so-called "active placebo" that doesn't contain the active ingredient of the treatment but does mimic its side effects.

## Statistical Regression

Subjects selected on the basis of extreme scores on one variable will tend to score at less extreme levels on another variable. This phenomenon of **statistical regression** is expected whether the two variables are the same measure collected at two points in time (e.g., IQ tested at ages 20 and 21) or different measures collected at the same point in time (e.g., IQ and extraversion). Someone who receives a very high score at age 20 will be expected to score above average, but less so, at age 21. Likewise, someone who receives a very low score at age 20 will be expected to score below average, but less so, at age 21. For this reason, statistical regression is often referred to as **regression toward the mean**, where "mean" refers to an average score.

Statistical regression is a more subtle phenomenon than the other threats to internal validity. It can be difficult to grasp and easy to misunderstand. The reason that statistical regression occurs is that an observed score—what we see in our data—reflects the sum of two factors, **true score** and **measurement error**. True score is like talent, fairly stable over time, and measurement error is like luck, randomly distributed over time.

If you select people with the very highest scores on, say, an IQ test, part of why they happened to score that high on that occasion will be talent, or exceptional levels of general mental ability, but part will be good luck, perhaps guessing correctly more often than would usually be the case. High ability alone gets people near the top, but among all those with high ability, those at the very top are those who also experienced good luck. When you test the same people again, though, luck won't tend to repeat itself. The people with the very highest scores the first time around will still score above average, but not as far above average. They will regress toward their own mean, their true score. A new, and partially overlapping, set of people will attain the very highest scores this time. They also have high ability, but they experienced the best luck this time around.

All of this is equally true for those who score on the low end. It takes a combination of low ability and bad luck to be among the very lowest of scorers on any particular occasion.

Confusion about statistical regression often stems from misunderstanding one of two things. First, regression operates at the level of individuals. A person's performance will tend to regress toward his or her own personal mean, or true score, not toward a group mean. Second, regression is distinct from genuine change. A person's performance can improve over time, but statistical regression still occurs relative to this changing mean. All of this can be challenging to grasp in the abstract, so let's consider a concrete example.

Imagine that Zeke is fairly new to basketball and wants to improve his free-throw shooting. He decides to practice by shooting a lot of free throws every day. From day to day, the percentage of shots that he makes will vary. There are two factors at work, talent and luck. Suppose that when Zeke begins he can make, on average, 30% of his free throws. This mean is a measure of Zeke's talent. If he makes 50% one day, well above his mean, that's in large part because of good luck. We can expect he'll do worse the next day because luck is random and doesn't tend to repeat itself. Specifically, we expect Zeke to be closer to his talent, his mean, the next day: 30%. Likewise, if he makes only 10% one day, that unusually bad performance is largely due to bad luck, and he'll probably do better, closer to his mean of 30%, the next day.

As time goes by, Zeke's daily practice will begin to lift his average. Suppose he reaches the point, after weeks or months of effort, that his daily average is now 50%. Day to day, his

outcomes are now much better than before, but they'll still vary randomly around his new mean of 50%. If Zeke makes 70% (or 30%) one day, we should expect him to be closer to 50% the next day because the good (or bad) luck is unlikely to repeat itself.

This is regression toward the mean. Zeke's performance outcomes will vary randomly (the luck component, or error) around his mean (the talent component, or true score). When any particular performance deviates very far above or below his mean, that's largely due to luck, and the next performance probably will be closer to his mean, because that's his real level of talent.

Naturally, this applies to everyone, not just Zeke. If you had 100 people attempt free throws and you selected those who scored in the top 10% to be retested the next day, you should expect this group to perform well—but not as well as the first time. Part of what landed people in the top 10% one day is luck, which won't repeat itself for the same people every day. A slightly different 10% of these 100 people will make it to the top the next day.

This is how statistical regression poses a problem in research. When individuals are selected based on extreme scores on one measure (e.g., depression at time 1), their scores on another measure (e.g., depression at time 2) will be less extreme due to regression toward the mean. This can easily be mistaken for genuine change (e.g., less depression) even though it reflects nothing more than the fact that measurement error is like luck and doesn't tend to repeat itself.

One solution to this problem is to select subjects at all levels, rather than only at extreme levels, on a measure. This is not always feasible. For example, when studying disease or disorder, this usually requires sampling only individuals functioning relatively poorly. When subjects must be selected based on extreme scores, randomly assigning them to conditions controls for statistical regression because this phenomenon should affect outcomes in each condition to the same extent. For example, randomly assigning equally depressed subjects to treatment and control conditions will hold constant the amount of statistical regression in each group. Any differences observed across conditions could then be more safely attributed to treatment effects rather than regression toward the mean.

## Evaluating External Validity

Whereas internal validity involves the soundness of causal conclusions drawn from a study's findings, **external validity** involves the generalizability of those findings to the populations, settings, outcomes, and time frames of genuine interest. Often, there is a trade-off between internal and external validity such that steps taken to strengthen one of these comes at the cost of weakening the other. Because peer reviewers tend to place much more emphasis on the causal conclusions that can be drawn from research than on the generalizability of findings, investigators tend to shore up internal validity, even at the expense of external validity.

This trade-off is understood well in the realm of clinical psychology. There's an important distinction between an **efficacy study**, which has strong internal validity, and an **effectiveness study**, which has strong external validity. In an efficacy study, subjects must meet stringent eligibility criteria (e.g., being diagnosed with a single mental disorder), therapy is delivered in a controlled manner (e.g., from a treatment manual), and data are analyzed only for subjects who complete all therapy sessions. Exerting experimental

control in these ways improves the ability to draw conclusions regarding cause and effect. At the same time, this limits the generalizability of those conclusions to clinical practice.

Effectiveness studies are quite different. Eligibility criteria are less stringent, which allows a more representative sample of patients to be studied. Therapy is delivered in a more natural manner, better reflecting the personalization of treatment in practice. So-called "intention-to-treat" analyses examine data for all patients enrolled in the study, which includes those who chose to discontinue treatment for any reason. In all of these ways, effectiveness studies make it more difficult to draw causal conclusions, but they do make it easier to generalize the results to clinical practice.

Whenever we want to apply scientific theory and research, we should consider how well findings are likely to generalize to the populations, settings, outcomes, and time frames of interest. Though many discussions of external validity focus all or most of their attention on generalizability to populations, all four domains are important.[4]

## Populations

The first step in thinking about external validity is to determine the most appropriate population for study. For example, who are the people that a line of applied research is intended to help? Ideally, a representative sample would be recruited from this population. In practice, it is common for a sample of convenience to be drawn from an unspecified population that differs substantially from the population of interest.

In the IFI study, this was handled very well by recruiting a sample of actual prisoners. The results should generalize reasonably well to similar populations of prisoners. For different populations (e.g., female or minimum-security prisoners), it's less clear how well the findings would generalize.

## Settings

The next step in thinking about external validity is to determine the most appropriate setting. Where should the study take place? Ideally, the setting would mimic the ways that subjects would experience something in the real world. In practice, it is common to collect data in a laboratory than in a more naturalistic setting.

In the IFI study, this was also handled very well by performing the research in actual prisons. The setting could not be more naturalistic, given the goals of this applied research.

## Outcomes

Another step in thinking about external validity is to determine the most appropriate measures to collect. What are the outcomes of interest? Ideally, the measures would be those with the most real-world significance. In practice, it is common to collect "proxy" measures.[5] Such data are easier to collect, but more distant from the outcomes of interest.

In the IFI study, the outcome of greatest interest is recidivism, or the commission of new crimes once released from prison. Because the criminal justice system is not omniscient, it's impossible to know for sure who has or has not committed crimes. Many

---

[4] Loyka, C., Ruscio, J., Edelblum, A. B., Hatch, L., Wetreich, B., & Zabel, A. (in press). Weighing people rather than food: A framework for examining external validity. *Perspectives on Psychological Science*.

[5] The term "proxy" refers to a substitute, as in "voting by proxy," a practice in which one person designates a representative to vote on his or her behalf when he or she can't be present.

crimes are undetected or unreported, for example. The researchers recorded whether each subject was rearrested and reimprisoned, and these proxy measures should provide a reasonable (if imperfect) assessment of recidivism that would generalize fairly well beyond this study.

### Time Frames

The final step in thinking about external validity is to determine the most appropriate time frame for a study. How long should a study last? Can it be performed in a single experimental session or are multiple sessions necessary? Should follow-up data be collected? Ideally, a study would be of sufficient duration to represent a treatment as it would be applied in practice, and follow-up data would be collected for a long enough period to test for enduring, rather than fleeting, effects. In practice, it is common to test only a simplified, condensed version of a treatment and to collect little, if any, follow-up data. Perhaps for the simple reason that they're much easier to conduct, single-session experiments are the norm, not the exception, even in applied research.

In the IFI study, the treatment phase of the study lasted for 16 to 24 months, with 6 to 12 months of aftercare following release from prison. No compromises were made when implementing the IFI program. With respect to outcomes, recidivism was tracked during a two-year follow-up period. Findings regarding the difference (or lack thereof) between the IFI and matched control groups may or may not generalize to even longer time frames, but two years is a pretty good starting point. For example, the researchers used a sufficiently long follow-up period to avoid floor effects (recidivism rates close to 0% in each group), which means that any actual differences between groups had the opportunity to reveal themselves.

## Problems

The following problems refer to a study designed to test whether superficial, visual cues influence the amount of food we eat.[6] Fifty-four undergraduate students volunteered to eat a soup-only lunch at a restaurant-style table in a research lab. Subjects were randomly assigned to eat what was described as a new recipe of tomato soup from either a normal bowl or a self-filling bowl that was surreptitiously rigged to slowly refill itself as soup was consumed. Enough time was allowed for subjects to eat as much soup as they liked, and the number of ounces of soup each person ate was measured and recorded. Subjects were also asked how many ounces of soup they thought they'd eaten. Though there was no statistically significant difference in perceived consumption across conditions, on average subjects ate significantly more soup from the self-filling bowl ($14.7 \pm 8.4$ oz.) than from the normal bowl ($8.5 \pm 6.1$ oz.). The researchers concluded that "people use their eyes to count calories and not their stomachs."

1. Is this research exploratory or hypothesis-testing? How can you tell?
2. Is this research experimental or correlational? How can you tell?

[6] Wansink, B., Painter, J. E., & North, J. (2005). Bottomless bowls: Why visual cues of portion size may influence intake. *Obesity Research, 13,* 93-100.

3. What is the population? Was this stated explicitly or did you have to infer it?
4. What is the sample? To what extent do you believe this sample is representative of the population?
5. What is the independent variable, conceptually? How was it operationally defined? What is its scale of measurement?
6. What are the dependent variables, conceptually? How was each operationally defined. What is the scale of measurement for each?
7. What descriptive statistics are presented?
8. Would you expect there to be a lot of sampling error surrounding these statistics? Why or why not?
9. Did the investigators use any inferential statistics? How can you tell?
10. Comment on the internal validity of this study. Review the list of common threats to internal validity and explain whether each of these is, or is not, a concern here.
11. Comment on the external validity of this study. Review the four domains to consider and explain whether each of these is handled well, or poorly, here.

* * *

The following problems refer to a study of links between obesity, physical activity, and caloric intake.[7] Using a very large probability sample of National Health and Nutritional Examination Survey data collected from U.S. adults between 1988 and 2010, the authors performed a wide range of statistical analyses to examine the relationships between many variables. They found that both waist circumference and body mass index (BMI) increased by about 0.3% to 0.4% per year, which was a statistically significant change. Among women, there was an increase from 19.1% to 51.7% who reported no leisure-time physical activity; the increase for men was from 11.4% to 43.5%. Both of these increases were statistically significant. There was no significant change in daily caloric intake, as estimated by trained dietary interviewers using a 24-hour recall technique. Both waist circumference and BMI trends were associated significantly with physical activity levels, but not with caloric intake. The researchers concluded that the increasing prevalence of obesity may have more to do with decreases in physical activity than with changes in food consumption.[8]

12. Is this research exploratory or hypothesis-testing? How can you tell?
13. Is this research experimental or correlational? How can you tell?
14. What is the population? Was this stated explicitly or did you have to infer it?

---

[7] Ladabaum, U., Mannalithara, A., Myer, P. A., & Singh, G. (2014). Obesity, abdominal obesity, physical activity, and caloric intake in US adults: 1988 to 2010. *American Journal of Medicine, 127,* 717-727.

[8] A report based on data from Britain reached a similar conclusion: "The rise in obesity has been primarily caused by a decline in physical activity at home and in the workplace, not an increase in sugar, fat, or calorie consumption." See Snowdon (2014), "The Fat Lie," Institute for Economic Affairs.

15. What is the sample? To what extent do you believe this sample is representative of the population?
16. What are the independent variables, conceptually? How was each operationally defined? What is the scale of measurement for each?
17. What are the dependent variables, conceptually? How was each operationally defined? What is the scale of measurement for each?
18. What descriptive statistics are presented?
19. Would you expect there to be a lot of sampling error surrounding these statistics? Why or why not?
20. Did the investigators use any inferential statistics? How can you tell?
21. Comment on the internal validity of this study. Review the list of common threats to internal validity and explain whether each of these is, or is not, a concern here.
22. Comment on the external validity of this study. Review the four domains to consider and explain whether each of these is handled well, or poorly, here.

* * *

23. If the margin of error for a poll of $N = 400$ is ±5%, how many people would need to be polled for the margin of error to be ±1%?
24. In the IFI study, there were 177 members of the treatment group and 1,754 members of the matched comparison group. Do the unequal group sizes pose a problem related to sample bias? How about sampling error?

**Problems 1 – 11 are due at the beginning of class.**

# 2. Describing Data

## Overview

The first step in working with data should always be to take a careful look at it. Before we use descriptive statistics to summarize responses, let alone using inferential statistics to test hypotheses, we need to know a lot about the data. Do the data differ across a series of discrete categories (e.g., sex, race, marital status)? If so, it's pretty simple to indicate which categories occur more or less frequently. Qualitative data are easy to summarize.

Alternatively, do the data differ along a continuum of values (e.g., IQ scores, ratings on a 7-point Likert scale)? In this case, what's the shape of the distribution of scores? Are they bunched together or widely dispersed? Are there are any **outliers**, atypical scores either caused by data entry mistakes or representing unusual responses that could exert too much influence on results? Quantitative data are more complex to summarize, and because it's not always possible to present tables or graphs to show all scores, numerical summaries can help paint a mental picture of what the distribution looks like. This chapter will emphasize two features of score distributions: **central tendency**, which locates the center of a distribution, and **variability**, which indicates how widely scores are spread.

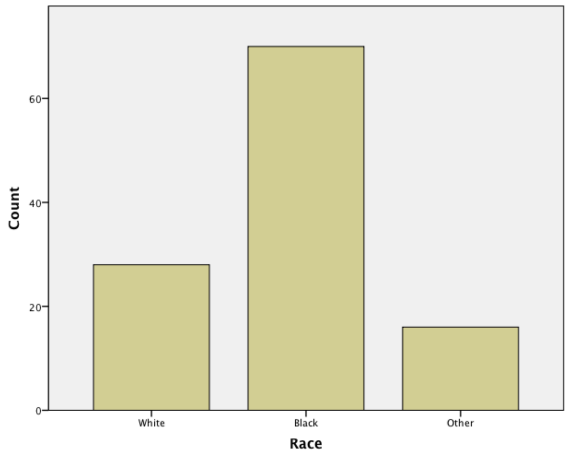
Particularly when we're interested in using descriptive statistics as estimates of population parameters, well-chosen measures of central tendency and variability will be highly **stable**. This means that they would be consistent when calculated for random samples from the same population. The less stable a measure, the more it varies from sample to sample, and the more poorly it estimates its population parameter.

## Qualitative Data

For a strictly categorical variable (e.g., anything measured using a nominal scale), a **frequency table** or a **bar chart** is simple to construct. Here are examples of each for a sample of 114 inmates at a federal corrections facility who were released on parole:

**Race**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | White | 28 | 24.6 | 24.6 | 24.6 |
| | Black | 70 | 61.4 | 61.4 | 86.0 |
| | Other | 16 | 14.0 | 14.0 | 100.0 |
| | Total | 114 | 100.0 | 100.0 | |




In the frequency table, the first column lists the categories (White, Black, Other). The "Frequency" column shows the number of cases belonging to each category. The "Percent" column shows the percent of cases for each category. These percentages are calculated out of the total number of cases ($N$ = 114). The "Valid Percent" column calculates percentages after excluding any missing data. Because there are no missing values in this instance—race is provided for all 114 cases—the values are identical for percent and valid percent. Finally, the "Cumulative Percent" column accumulates the percentages across categories. This is not meaningful for qualitative data and should be ignored in this case.

For qualitative data, a **bar chart** plots the frequency of scores for each category. The bars themselves should not touch one another. This separation between the bars in a bar chart signals to the informed reader that the data are qualitative. The gaps imply a discontinuity between the categories that the bars represent. If you're describing your data and have enough space to include a graph, a bar chart can communicate very intuitively and effectively to readers.

If you don't have the luxury of that much space, you might need to stick to a text-based summary written from the values in a frequency table. Because there are so few categories, it would be reasonable to list them all, indicating the number and/or percentage of cases belonging to each. For example, you could report that the sample of 114 parolees includes 28 white individuals (24.6%), 70 black individuals (61.4%), and 16 individuals of other races (14.0%). If the number of categories is too large for you to devote space to listing them all, you can list those that occur most often and then indicate how many cases belong to all other categories.

## Quantitative Data

In this same parole data set, quantitative scores are available for the Lifestyle Criminality Screening Form (LCSF). The LCSF assesses factors related to a criminal lifestyle using 14 items, and scores can range from 0 to 22. Whereas race represents qualitative differences between individuals—variation across discrete categories that cannot even be rank-ordered—LCSF scores vary along a continuum from low to high. Several kinds of tools can be helpful to understand and summarize a quantitative variable like this, including a frequency table, graphs, and descriptive statistics.

## Frequency Table

The frequency table for these 114 parolees' LCSF scores is shown below:

**Lifestyle Criminality Screening Form**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 0 | 1 | .9 | .9 | .9 |
| | 1 | 4 | 3.5 | 3.5 | 4.4 |
| | 2 | 4 | 3.5 | 3.5 | 7.9 |
| | 3 | 7 | 6.1 | 6.1 | 14.0 |
| | 4 | 12 | 10.5 | 10.5 | 24.6 |
| | 5 | 11 | 9.6 | 9.6 | 34.2 |
| | 6 | 14 | 12.3 | 12.3 | 46.5 |
| | 7 | 13 | 11.4 | 11.4 | 57.9 |
| | 8 | 13 | 11.4 | 11.4 | 69.3 |
| | 9 | 7 | 6.1 | 6.1 | 75.4 |
| | 10 | 9 | 7.9 | 7.9 | 83.3 |
| | 11 | 12 | 10.5 | 10.5 | 93.9 |
| | 12 | 3 | 2.6 | 2.6 | 96.5 |
| | 13 | 3 | 2.6 | 2.6 | 99.1 |
| | 15 | 1 | .9 | .9 | 100.0 |
| | Total | 114 | 100.0 | 100.0 | |

Though scores can range up to 22, nobody in this sample scored above 15. There were very few scores at the extremes, with most clustered closer to the middle. The "Cumulative Percent" values are meaningful for quantitative data such as these. They convert each score in the table to a **percentile**, which indicates how many scores were at or below that value. For example, 83.3% of scores were at or below 10.

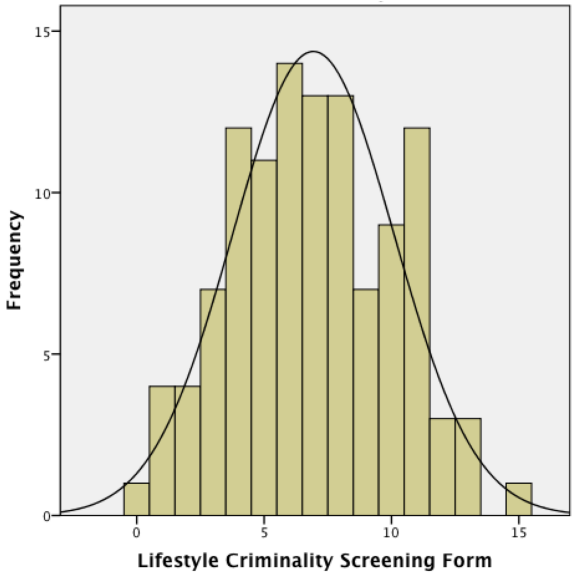
When a quantitative variable spans a very large number of distinct values, it can be useful to construct a **grouped** frequency table. Whereas the example shown above is **ungrouped**, meaning that each score that occurs in the data is listed in its own row in the table, a grouped table will list ranges of scores for each row of the table. The computer program used to generate these frequency tables (SPSS, introduced in the next chapter) does not provide grouped tables. Other software does, in which case a useful rule of thumb for a good number of rows to list in the table is the whole number closest to the square root of the sample size. For example, with $N = 500$ cases, a table can be constructed using 22 rows (the square root of 500 is 22.36).

One final note on frequency tables is that whereas SPSS lists percentages, other software might present equivalent information labeled as proportions, relative frequencies, or probabilities. Each of these is simply percentage / 100 (e.g., 83.3% is equivalent to a proportion, relative frequency, or probability of 83.3 / 100 = .833).

## Graphs

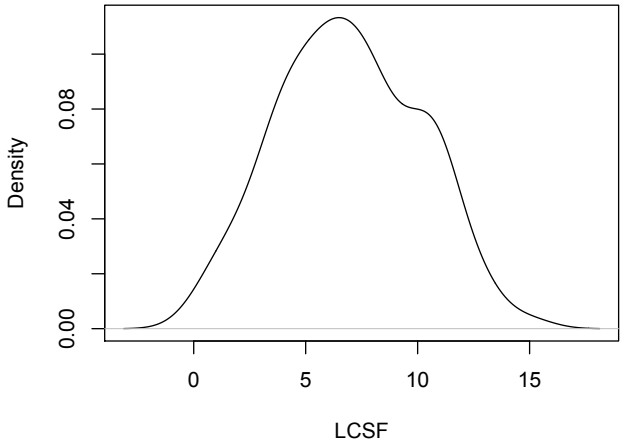
To display the frequencies for qualitative data, we used a bar chart. For quantitative data, we use a **histogram**. Once again, the frequency of scores is plotted, but this time the bars themselves touch one another to signal to the informed reader that the data are quantitative, varying along a continuum rather than belonging to discrete categories. The number of bars in a histogram, like the number of rows in a frequency table, can be as large as the number of distinct values (as in an ungrouped table) or something smaller (as in a grouped table). When the sample size is very large, a good rule of thumb is to set the

number of bars equal to a value close to the square root of *N*. Here is a histogram for LCSF scores:




This was generated using SPSS, which provides the option of superimposing a hypothetical normal curve. This can be helpful to determine whether the data are approximated well by a normal distribution. More will be said about shapes of distributions shortly.

Another graphical display for quantitative data is a **density plot**, which is basically a smoothed version of a histogram. Though SPSS does not generate density plots, other programs do. Here's an example for LCSF scores:



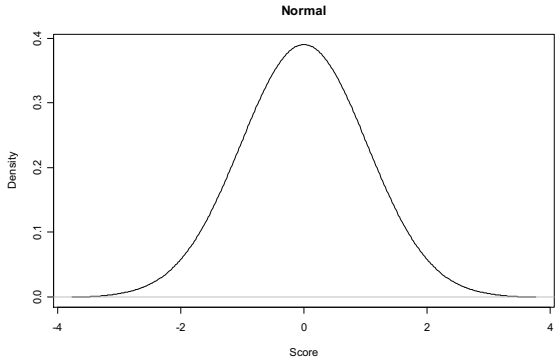

Not only does a density plot reveal the shape of a distribution at least as clearly as a histogram, but also it is simpler to sketch or read. Rather than plotting a series of bars, a single, smooth curve is plotted. Because a density plot is scaled such that the area under the curve equals 1, the *y* axis doesn't correspond to frequencies (as it does for a bar chart or histogram). You can safely ignore the scale along the *y* axis of a density plot; only the shape is important.

## Shapes of Distributions and Descriptive Statistics

Though the distribution of scores for a variable could take an infinite variety of forms, a handful of shapes are especially common. When reading or writing about research, it's important to know the terminology used to describe the shapes of distributions. Also, the most appropriate kinds of descriptive statistics used to summarize the data depend on the shape of the distribution. The most common distributions are considered in this chapter, along with measures of central tendency and variability to use with each.

### Normal

Probably the best-known distribution in statistics is the **normal curve**, also known as a **bell curve** or more technically as a **Gaussian distribution**. The normal curve is **symmetric**, meaning it's a mirror image of itself from left to right. There is a single peak, and the frequency of scores tapers off as you approach the **tails** of the distribution (the far left and far right portions). Here's an example of a normal distribution:




Some variables' distributions do follow a normal curve (e.g., height of adult men or adult women), but many others do not. Just because statistical analyses often assume that scores are drawn from a normal distribution, that doesn't make it true. It's important to check the shape of a distribution to see whether the assumption is satisfied.

The best way to summarize a normal distribution of scores numerically is to report its **mean** and **standard deviation**. The mean ($M$) is a familiar measure of central tendency, calculated simply as the sum of all scores divided by the number of scores:

$$M = \Sigma X / N,$$

where $X$ is an individual score and $N$ is the sample size. For example, for the set of five scores 1, 2, 3, 4, 5, you'd get $M = (1 + 2 + 3 + 4 + 5) / 5 = 3$.

The term "mean" usually refers to what is more precisely called the arithmetic mean or, in everyday language, the "average". There are actually many kinds of averages or means that can be calculated. For example, you can multiply all $N$ scores and then take the $N^{th}$ root of the product, and this is known as the geometric mean.[9] Unless otherwise specified, you can assume that "mean" refers to the arithmetic mean. For a normal distribution of scores, the mean is the most stable measure of central tendency.

[9] Whereas the arithmetic mean of the three scores 1, 5, 25 = 31 / 3 = 10.33, the geometric mean is the cube root of $1 \times 5 \times 25$, which equals 5.00.

The standard deviation (*SD*) is the typical distance from a score to the mean. In other words, if you were to select a score from a distribution at random, a good guess as to how far it will be from the mean is one *SD*. Though the term "standard deviation" might seem a bit off-putting at first, this measure is actually very well-named. "Standard" means typical, and "deviation" refers to distance. The formula for the *SD* shows this more precisely:

$$SD = \text{sqrt}(\Sigma(X - M)^2 / (N - 1))$$

To understand how this formula produces a measure of "typical distance" between a score and the mean, it can be broken down into the following steps:

1. Calculate each score's **deviation score**, or distance to the mean. That's what $X - M$ represents.
2. Square the deviation scores. That's what $(X - M)^2$ represents. The reason we square the deviation scores is so that they won't just cancel out when we average them (in the next step). Some deviation scores are negative (to the left of the mean), some are positive (to the right of the mean), and if you average them you'll get 0. Squaring them first solves that problem.
3. Take the average of the squared deviation scores. That's what the expression $\Sigma(X - M)^2 / (N - 1)$ accomplishes. Basically, you're adding up squared deviation scores and dividing by the number of them. If you're curious about why we divide by $N - 1$ rather than $N$ to take this average, check the footnote.[10]
4. Finally, take the square root of the average of the squared deviation scores. This step simply reverses the effect of squaring the deviation scores back in step 2. What you're left with is the typical distance from a score to the mean.

Here's what this looks like for scores of 1, 2, 3, 4, and 5 (for which $N = 5$ and $M = 3$).

1. Calculate deviation scores: $1 - 3 = -2$; $2 - 3 = -1$; $3 - 3 = 0$; $4 - 3 = 1$; $5 - 3 = 2$.
2. Square the deviation scores: $(-2)^2 = 4$; $(-1)^2 = 1$; $0^2 = 0$; $1^2 = 1$; $2^2 = 4$.
3. Average the squared deviation scores: $(4 + 1 + 0 + 1 + 4) / (5 - 1) = 2.50$.
4. Take the square root of this average: $\text{sqrt}(2.50) = 1.58$. This is the *SD*.

Like the mean, the stability of the standard deviation is excellent for a normal distribution.

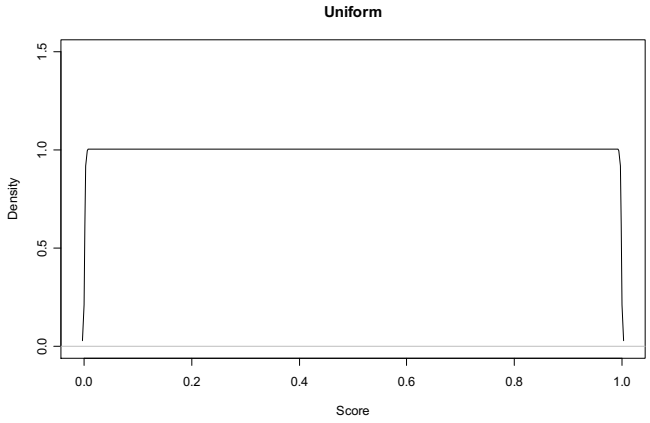
In the parole data, the *M* and *SD* would be good measures of central tendency and variability for LCSF scores because they were approximately normally distributed. In this case, $M = 6.93$ and $SD = 3.16$. In other words, the middle of the distribution is near a value of 7, and the typical distance from a score to the mean is about 3.

---

[10] The reason we divide by $N - 1$, rather than just $N$, to take the average of the squared deviation scores is that the *SD* as a descriptive statistic is a biased estimator of its population parameter. In the highly unusual situation in which you're calculating the standard deviation for a population of scores, you'd divide by $N$. In the much more common situation in which you're calculating the standard deviation for a sample of scores, you divide by $N - 1$ to correct for the bias. Most computer programs always divide by $N - 1$, assuming you're working with a sample rather than a population of scores.

## Uniform

Another symmetric distribution is referred to as **uniform**. This occurs when there are fairly equal numbers of cases at each score, and it produces a **flat** or **rectangular** shape when graphed. Here's an example of a uniform distribution:

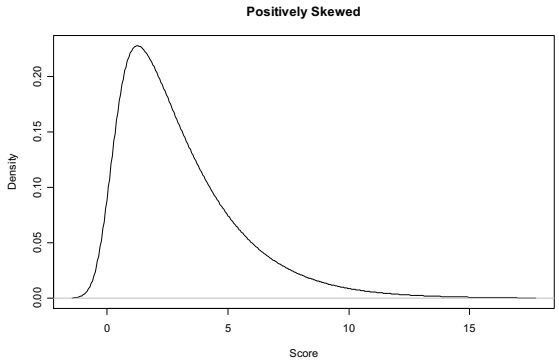



Uniform distributions don't tend to occur as often as normal distributions. Variables measured on an ordinal scale sometimes have uniform distributions. This happens when the ranks represent the position within a full set of scores (e.g., ranks of 1, 2, 3, …, *N* for *N* scores). Sometimes data are converted to ranks in order to achieve a uniform distribution. For example, when a variable's distribution badly violates the assumption of normality that underlies many statistical analyses, it can be helpful to convert it to ranks. This at least ensures symmetry, and a uniform distribution may be closer to normality than the original distribution.

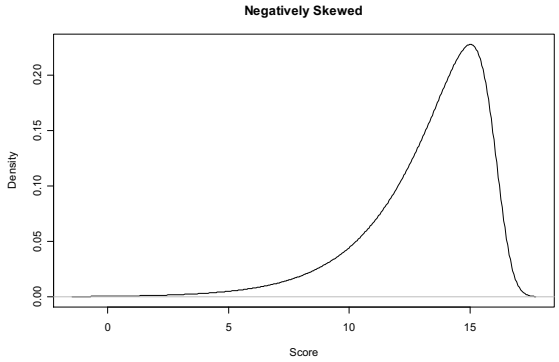
The mean and standard deviation remain good measures of central tendency and variability for uniform distributions.

## Skewed

**Asymmetric** distributions are also known as **skewed**, and the direction of the asymmetry is indicated as well. If most of the scores are low, with a long, thin tail spreading out across higher scores, this is known as **positive skew**, also described as a **right-skewed** distribution or one that is **skewed to the right**. Here's an example of a positively skewed distribution:



If instead most scores are high and the long, thin tail spreads out across lower scores, this is known as **negative skew**, also described as a **left-skewed** distribution or one that is **skewed to the left**. Here's an example of a negatively skewed distribution:




One way to remember the distinction between positive and negative skew is to think about which direction the long, thin tail is pointing. If it points toward positive numbers that's positive skew, and if it points toward negative numbers that's negative skew.

Many variables are at least somewhat skewed. For example, timed variables tend to be skewed. When measuring reaction time to a stimulus, most people will be reasonably quick. A few will take much longer, perhaps because of inattention or impairment. That produces a positively skewed distribution. In contrast, the time students take to complete a classroom test is likely to be negative skewed. Most students will use all or most of the time allowed, but a few will finish substantially earlier. Both of these timed variables are also bounded on one side, a feature that tends to produce a skewed distribution. Reaction time cannot be less than 0, and time to complete a test must be less than the class period.

The degree of skew can be very slight, or it can be enormous. There's no accepted standard for determining when a distribution no longer qualifies as approximately normal and should be considered skewed. That's a judgment call. It can be helpful to use qualifiers such as slightly skewed, moderately skewed, or strongly skewed to express the amount of asymmetry perceived in a distribution.

When it comes to descriptive statistics for skewed distributions, the mean and standard deviation are poor choices. Both can be heavily influenced by extreme scores on one side of a distribution, or by a skewed distribution even if there are no outliers. The mean tends to be pulled in the direction of a long, thin tail or extreme scores. To help understand why this is so, consider that if you were to build a physical model of a histogram, the mean would be the balancing point. Try to balance your model on any point to the left or right of the mean, and it would tip to one side or the other. If you add an extreme score to a distribution, the mean has to move pretty far in that direction to maintain the balance. Likewise, the long, thin tail of a skewed distribution also forces the mean in that direction to maintain balance. For similar reasons, the standard deviation can be affected quite a bit by outliers or skew.

Whereas the stability of the mean and standard deviation is mathematically optimal for a normal distribution and good for pretty much any symmetric distribution (e.g., uniform), it's much poorer as the amount of skew or the severity of outliers increases.

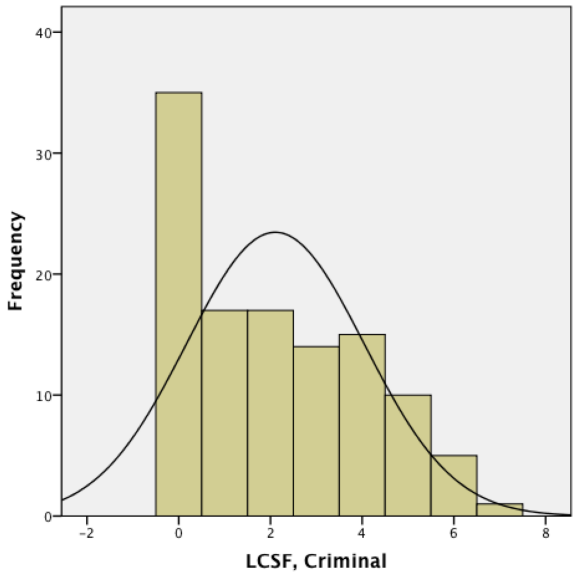
Fortunately, alternative measures of central tendency and variability are more stable with asymmetry or outliers. These alternatives are based on positions within a distribution

rather than averages or distances. The **median** (*Mdn*) is defined as the middle score or, if there is an even number of scores, the mean of the middle two scores. The simplest way to calculate this is to rank-order a set of scores, then cross off pairs at both ends until all that remains is a single score in the middle (if *N* is odd) or two scores in the middle that need to be averaged (if *N* is even). This yields a stable measure of central tendency that's highly robust to outliers or asymmetry. For example, you could multiply the largest data point by 1,000,000 and this would have no effect on the median. It's based only on the middle score (or, if *N* is even, the middle pair of scores).

A good measure of variability to accompany the median is the **interquartile range** (*IQR*), which spans the middle 50% of scores. Put another way, it's the range that runs from the first quartile (Q1, below which lies 25% of the scores) to the third quartile (Q3, above which lies 25% of the scores). You can calculate the *IQR* as Q3 – Q1, or you can report it in a more informative way as running from Q1 to Q3. This yields a stable measure of variability that's highly robust to outliers or asymmetry. For example, multiplying the largest data point by 1,000,000 would have no effect on the *IQR*.

The median and interquartile range are not only robust to skew and outliers, but also they're pretty easy to understand. If you list scores in order, the median is the one in the middle. The interquartile range encloses the middle half of the scores.

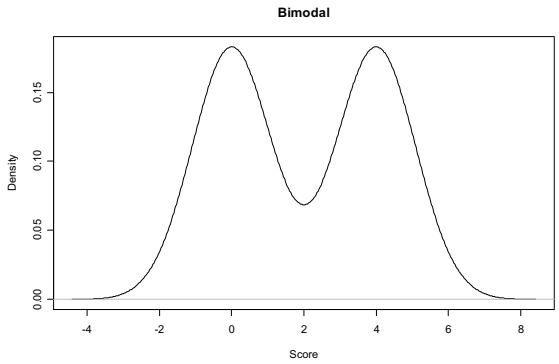
In the parole data, the *IQR* would be a good measure of variability for LCSF-Criminal scores because they were positively skewed:



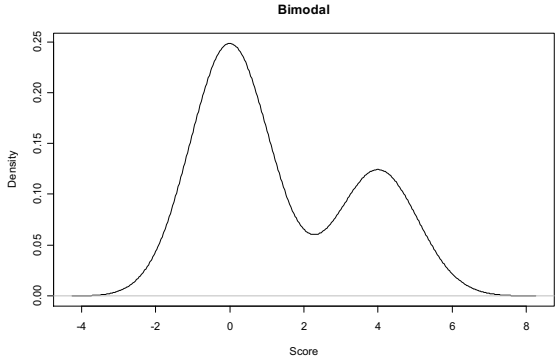

In addition to reporting that the median value was 2, the *IQR* was 0 to 4.

## Bimodal and Multimodal

A final type of distribution worth knowing is **bimodal**, which refers to a curve with two peaks. Technically, to qualify as bimodal the peaks need to reach equal heights. Here's an example of such a bimodal distribution:




Most researchers would be willing to describe any distribution with two clearly distinct peaks as bimodal, even if the heights of these peaks differ. Here's another example of a bimodal distribution, this time with peaks of unequal height:




Bimodality occurs when there is a mixture of two subgroups of subjects whose scores differ substantially. The relative heights of the two peaks correspond to the relative sizes of the two groups. The first bimodal distribution (above) shows what you'd expect with equal-sized groups, and the second shows what you'd expect if there were many more people in the lower-scoring than the higher-scoring group.

Even though subgroups are mixed all the time in samples of data, bimodal distributions are relatively rare. The reason is that the groups have to differ by a lot for bimodality to occur. For example, simply mixing adult men and women in a sample and measuring their heights will not produce a bimodal distribution. Even though men are taller than women, on average, there is too much overlap in the heights of these groups to produce a bimodal distribution. When you do see a bimodal distribution, that means that the groups differ substantially.

It is possible not only to have bimodal distributions, but also to have curves with more than two peaks, or **multimodal** distributions. Because the subgroups would have to differ by a lot for more than two peaks to emerge, multimodality occurs even less often than bimodality.

When considering the possibility of multiple modes, it's important not to overinterpret the "lumpiness" of chance. Any distribution, especially in a small sample, will have peaks and valleys. When making a judgment about the shape of a curve, look at the big picture and try not to jump to conclusions of bimodality or multimodality based on small differences in height between adjacent bars in a histogram. A density plot can be helpful in

avoiding the temptation to perceive meaningful peaks, which correspond to subgroups, where only chance variation exists.

For example, look back at the histogram for LCSF scores shown earlier. There are three places where the height of one bar is less than that of its neighbors on both sides (at scores of 4, 6, and 11). It would be mistaken to conclude that this is a multimodal distribution based only on these relatively minor fluctuations. Rather than reporting that subgroups differ substantially from one another, it's much more reasonable to conclude that this distribution is approximately normal. The deviations from normality are fairly minor and to be expected given the modest sample size. Reading the density plot underscores this conclusion. The curve is fairly symmetrical, with no large gaps. There is somewhat of a lump on the right side, but this, too, is to be expected given the sample size. Peaks would need to be much better-defined to indicate bimodality or multimodality.

If you do conclude that a distribution has two or more modes, it might be misleading to report either the mean or the median as a measure of central tendency. Though these measures would identify the middle of the distribution (as long as it's fairly symmetric), they would not represent a typical score. For example, in the first bimodal curve shown above, the mean (and median) would be 2. That's actually a somewhat unusual score, as most scores are closer to either 0 or 4. Reporting the locations of the separate modes is a better approach. For example, you might report that the first distribution shown above is bimodal, with about equal numbers of scores clustered around values of 0 and 4. Likewise, you might report that the second distribution shown above is bimodal, with many scores clustered around a value of 0 and a smaller number of scores clustered around a value of 4. Once you have identified the modes, no measure of variability is particularly useful.

## Symmetry, the Mean, and the Median

A final note on the relationship between symmetry, the mean, and the median can be helpful to keep in mind. In a symmetric distribution, the mean will equal the median. In an asymmetric distribution, on the other hand, the mean will be drawn toward the long, thin tail. The greater the degree of skew, the further apart the mean and median will be. Thus, if all you know is the mean and median of a distribution, you can infer something about whether, and to what extent, it's skewed. For example, if the mean and median of a large set of IQ scores are 140 and 110, respectively, you can tell that this distribution is highly positively skewed. Some individuals with very high IQs must be pulling up the mean, relative to the median. If, instead, the mean and median of a large set of IQ scores are 118 and 121, respectively, this would suggest that there is only the slightest negative skew in the distribution. This doesn't necessarily indicate that the distribution is normal, but it does suggest it's highly symmetric.

## Problems

Suppose that you observe the following frequencies (and percentages) for 40 college students' majors:

| Major | Frequency | Percent |
|---|---|---|
| Natural science | 9 | 22.5% |
| Social science | 5 | 12.5% |
| Humanities | 9 | 22.5% |
| Business | 11 | 27.5% |
| Undeclared | 6 | 15.0% |
| Total | 40 | 100.0% |

1. Are these data qualitative or quantitative?
2. Would it be appropriate to report any measures of central tendency or variability for these data? If so, which ones would you report? If not, why not?
3. Write a brief description of these data.

* * *

The following problems refer to these scores on a high school biology test:

60, 65, 70, 70, 75, 80, 80, 85, 85, 85, 90, 90, 90, 95, 95, 95, 95, 95, 100, 100

4. What is the mean (*M*) for these scores?
5. What is the standard deviation (*SD*) for these scores?
6. What is the median (*Mdn*) for these scores?
7. What is the interquartile range (*IQR*) for these scores?
8. If you were to graph this distribution, what would its shape be? (Hint: Consider the relationship between symmetry, the mean, and the median.)
9. Construct a histogram for these scores. On the *x* axis of your graph, label scores of 60, 65, 70, 75, … 90, 95, 100. On the *y* axis of your graph, label frequencies from 0 to whatever is the largest frequency you observe for any score in the sample.
10. Based on the shape of the distribution shown in your histogram, what would be the best measures of central tendency and variability for these data? Why?
11. Why might someone want to use a different measure of central tendency? Consider who has an interest in test scores and why they might like to see a high or a low average.

* * *

On a survey distributed to a large, random sample of students at a four-year college, one of the questions asked the students to indicate their class year (freshman, sophomore, junior, or senior).

12. If you were to plot a histogram of the responses, what would you expect its shape to be?
13. Based on the shape of the distribution you expect to observe in your histogram, what would be the best measures of central tendency and variability for these data? Why?

On the same survey, another question asked the students to indicate their attitude toward abortion on a 7-point Likert scale ranging from 1 ("Strongly pro-life") to 7 ("Strongly pro-choice").

14. If you were to plot a histogram of the responses, what would you expect its shape to be?
15. Based on the shape of the distribution you expect to observe in your histogram, what would be the best measures of central tendency and variability for these data? Why?

* * *

16. Members of a local union threaten to go on strike for higher pay. The union's president reports they're paid, on average, $50,000. The governor's office reports that the state pays them, on average, $80,000. Assuming both statistics are correct, what could cause this discrepancy? (Hint: Consider how "average pay" might be operationalized.)
17. Suppose that you could obtain data on the annual income (i.e., salaries, wages, tips) of every U.S. citizen. What would you expect the shape of the income distribution to be? Why?
18. Suppose that you could accurately adjust the income data to remove all taxes paid and add in the dollar value of benefits (e.g., welfare payments, rent subsidies, food stamps). How would you expect the shape of this adjusted income distribution to compare to that of the original income distribution?

**Problems 1 – 11 are due at the beginning of class.**

# 3. Overview of SPSS and APA Style

## Overview

SPSS is a user-friendly program for managing, exploring, and analyzing data. There are three types of SPSS files, each appearing in its own window: data, syntax, and output. This chapter shows how to use them, followed by a brief review of APA style for reporting statistical results.

## SPSS Data Files

In SPSS, data are stored in a spreadsheet with one row per subject, one column per variable. A new **data file** will simply be a blank spreadsheet into which you can enter data or copy and paste it from another spreadsheet program, such as Excel.

Any data that you want to analyze should be entered numerically. Use numerical codes to represent categories. For example, you can code experimental conditions as 1 = treatment, 2 = control. This will work for any number of categories, and it makes no difference what codes you choose for each category. Entering data using codes is quicker and less error-prone than using text, but the codes will not be memorable or meaningful. Fortunately, you can enter **value labels**. Each value label is text that corresponds to one of your code numbers; you only have to enter the labels once. Value labels are stored with the data file and used in any output you generate. For example, notice that the frequency table for race, shown earlier, doesn't list categories of 1, 2, and 3; instead, it uses the value labels of "White", "Black", and "Other".

SPSS also stores and displays labels for variables. For example, "LCSF" is a variable in this dataset, but that abbreviation is not as informative as the variable label "Lifestyle Criminality Screening Form". It's a good habit to enter a **variable label** for every variable in a dataset, even when the meaning might seem obvious.

The easiest way to set up a new data file is to begin in the "Variable View". You can get there via a tab near the bottom of the data window. This is not the data spreadsheet itself, but a table that lists the variables in the dataset and their characteristics. Start by entering variable names, with no spaces allowed, one variable per row in this table. Then, for each variable, enter a variable label in the "Label" column of the table. Finally, for each categorical variable, enter the value labels. Clicking on a cell in the "Values" column causes a "…" button to appear, and clicking on that button brings you to a dialogue box in which you can add, edit, or remove value labels for that variable.

Once you've set up the data file in the "Variable View", go back to the "Data View" via a tab near the bottom of the data window. You'll see that your variables appear as column headings, and if you place the pointer on one of them for a moment, the variable label will appear. This spreadsheet is where you enter the data. Remember that it's organized such that all each row will contain the data for one subject. Leave cells blank to indicate any missing data.

SPSS saves data files with a .sav extension. You can use the "Save As…" option to save your data in other formats, such as an Excel file.

## SPSS Syntax Files

The syntax window is where you tell SPSS what you want it to do with your data. From the "File" menu, choose "New" and then "Syntax" to open a new syntax window. This will be a blank text file into which you can enter commands. SPSS saves a **syntax file** with a .sps extension, but it contains nothing other than plain text.

Throughout this book, commands for managing, exploring, and analyzing data will be introduced and illustrated. For example, the "freq" command (short for "frequencies") can be used to generate frequency tables, histograms, and descriptive statistics. For a qualitative variable *X*, the following command would generate a frequency table:

```
freq vars = X
```

For a quantitative variable *Y*, the following command would generate more output:

```
freq vars = Y
/histogram normal
/stats all
/per 25 75
```

In addition to a frequency table, the first subcommand (each subcommand begins with a "/") generates a histogram; adding "normal" after "histogram" requests a superimposed, hypothetical normal curve that can help you judge the extent to which this approximates the observed distribution. The second subcommand generates a table of descriptive statistics that would include the mean, standard deviation, and median. The third subcommand includes the 25$^{th}$ percentile (Q1) and the 75$^{th}$ percentile (Q3) in the table of descriptive statistics. These define the interquartile range, which you'd want to report if you decide to use the median as your measure of central tendency.

You can also list multiple variables on the "freq" command. For example, if you have three variables *X*, *Y*, and *Z*, any of these variations on the first line of the "freq" command (leaving subcommands the same as before) will provide full output for all three variables:

```
freq vars = X Y Z
```

```
freq vars = X to Z
```

```
freq vars = all
```

Many, but not all, SPSS commands can be accessed via pull-down menus and dialogue boxes. For several reasons, using commands is a smarter choice:

- Typing commands can be much easier than navigating your way through menus and dialogue boxes. This is especially true as you learn to use SPSS, because you can copy and paste commands from another source (such as this book) and skip the menus and dialogue boxes altogether.
- Because SPSS seldom changes its commands, even as it changes its menus and dialogue boxes all the time, working with the commands now will prove helpful if you need to use SPSS in the future.
- Finally, saving a syntax file provides you with a record of what you have done and a way to easily re-run analyses if you later add (or remove) some data.

Typing a command in the syntax window has no effect until you run it. You can run one command at a time, or a lengthy series of commands, by highlighting what you want to run and either pressing Ctrl-R, clicking the green ► button near the top of the syntax window, or selecting an option from the "Run" menu.

## SPSS Output Files

The objects—text, tables, and graphs—produced by running SPSS commands are displayed in the output window. This will be opened for you when you run the first command of an SPSS session. You can save an **output file**, which would have a .spo extension, but that is neither necessary nor convenient for most purposes. An output file can only be opened by SPSS, so you'd have to be running the program to access your output. If you are running the program, it's probably easier to simply re-run a command to recreate the output you want rather than searching for it in what might be a very lengthy output file. As you'll see, SPSS sometimes generates a lot of output.

When using SPSS to view your output, you can click within the pane on the left to move directly to a particular object. You can highlight objects to print by selecting them in either the listing pane on the left or the display pane on the right. Alternatively, you can copy and paste objects into Word or another program to save or print them more conveniently.

## APA Style for Reporting Statistical Results

This section contains an overview of most elements of APA style that you need to know to report statistical results. For more detailed information, see the latest edition of the *Publication Manual* (especially pp. 32-35 and 111-123). The page numbers provided in brackets below refer to locations in the 6th edition, published in 2009.

Computer software often has default settings that are inconsistent with some elements of APA style. For example, if you don't change settings on Word, you'll probably break rules #2 and #3, below. It's your job to adhere to APA style. Computer defaults are no excuse.

In the list provided below, the first seven items about format apply throughout a paper in APA style. The remaining items are more specific to the reporting of statistical results. Many of the items on this list refer to concepts or statistics appearing in later chapters, so don't be concerned if you're unfamiliar with them now.

### General Guidelines

1. *Font*. Use a standard, readable font, such as Times New Roman in 12-point size. [p. 228]
2. *Margins*. Leave margins of 1" on all sides of the page. [p. 229]
3. *Line spacing*. Double-space all text, with no extra space around headings or between paragraphs. [p. 229]
4. *Justification*. Do not right justify text. [p. 229]
5. *Indenting*. New paragraphs are indented ½". [p. 229]

6. *Bold print, underlining*. Bold print is used only for matrix variables or headings, do not use it for any other purpose. Likewise, do not underline anything; use italics for emphasis. [pp. 104-106, 118]
7. *Capitalization*. Do not capitalize names of groups or conditions in a study. [p. 104]

### Statistics Guidelines

8. *Italics*. Statistics such as *N*, *M*, *SD*, *Mdn*, *IQR*, *t*, *F*, *p*, *d*, and *r* are italicized. Do not italicize Greek letters (e.g., μ, σ, α, $\eta^2$ and $\chi^2$) or numbers or symbols appearing along with statistical results (e.g., "(*M* = 4.37, *SD* = 2.13)", not "*(M = 4.37, SD = 2.13)*"). [pp. 119-123]
9. *Decimals and rounding*. The general rule is to report statistics to two decimal places; there are several specific guidelines worth noting. [pp. 113-114]
    A. You don't need to add decimals to whole numbers that are not statistics (e.g., report that there were "23 men and 45 women", not "23.00 men and 45.00 women"; state that a scale ranged from "1 (strongly disagree) to 7 (strongly agree)", not "1.00 (strongly disagree) to 7.00 (strongly agree)").
    B. If computer output contains fewer decimal places than you will report, add zeros as needed (e.g., "*M* = 4.2" becomes "*M* = 4.20" and "*M* = 4" becomes "*M* = 4.00").
    C. Provide exact *p* values (to 3 decimal places) rather than writing "*p* < .05" or "*p* > .05." If computer output gives a *p* value of .000, report this as "*p* < .001" because *p* cannot literally equal 0.
    D. When extra decimal places are available, do not drop the extra digits—round off to the nearest value (e.g., "*M* = 1.866" becomes "*M* = 1.87", not "*M* = 1.86").
    E. When a statistical value is less than 0, place a 0 before a decimal point only if the statistic can exceed 1 (e.g., "*d* = 0.70"), not otherwise (e.g., "*p* = .028").
10. *Spaces*. Leave one space between statistical symbols, punctuation, and numerical values (e.g., "*M* = 4.37", not "*M*=4.37"), but no space between statistical abbreviations (such as *t*, *F*, *r*), and the parentheses that include the degrees of freedom (e.g., "*t*(35) = 2.12", not "*t* (35) = 2.12"). [p. 118]
11. *Numbers that begin sentences*. Write out any number that begins a sentence (e.g., "Two hundred and fifty responses…", not "250 responses…"). [p. 112]
12. *SPSS variable names*. Do not use SPSS variable names in your writeup, unless the variable name happens to be identical to the clearest way to identify the variable (e.g., a variable named "age").
13. *Names of statistical tests*. Unless you are using an unusual test—and nothing in this book qualifies—do not state what statistical test was used. This should be clear from the context. [pp. 116-117]
14. *Statistical information*. Provide all pertinent statistical information for each test. For example, *t* test results include the *t* value, *df*, *p* value, and Cohen's *d*; *F* test results include the *F* value, both *df*, *p* value, and $\eta^2$; correlation results include the *r* value, *df*, and *p* value. [pp. 33-34, 116-17]

15. *Explanation*. Explain the nature of any statistically significant differences. Simply stating that an effect was observed is insufficient. For comparisons across multiple conditions, the *M* and *SD* for each one should be included. For an interaction effect, describe how the factors combined to predict the outcome. [pp. 33, 116]

## APA Style for Describing Data

As an illustration of writing in APA style, consider how to describe data. The clearest, most concise way to do so depends on the type of data. For qualitative variables, you can report the number and percentage of subjects in each category. For quantitative variables, you can report the range of scores, the shape of the distribution, and the appropriate measures of central tendency and variability. If it is not obvious, the way that each variable was defined and assessed should be explained along the way.

The following paragraph shows how one might describe some of the parole data in APA style. Note that when referring to the size of a total sample the abbreviation is *N*, and when referring to the size of a subsample of cases the abbreviation is *n*.

Among the inmates (*N* = 114) at a federal corrections facility who were released on parole, the most common race was black (*n* = 70; 61.4%), followed by white (*n* = 28; 26.4%), and other (*n* = 16; 14.0%). The Lifestyle Criminality Screening Form (LCSF) contains 14 items, and scores can range from 0 to 22. In this sample, scores ranged from 0 to 15 in an approximately normal distribution (*M* = 6.93, *SD* = 3.16). Scores on the LCSF-Criminal subscale, which contains only crime-related items, ranged from 0 to 7 in a positively skewed distribution (*Mdn* = 2, *IQR* = 0 to 4).

## Problems

1. What are the three types of SPSS files and what is each one used for?
2. What are the steps to setting up a new SPSS data file?
3. Suppose you have three variables that indicate the number of diagnostic criteria met for Major Depressive Disorder, Generalized Anxiety Disorder, and Post-Traumatic Stress Disorder. The variables are MDD, GAD, and PTSD.
   a. If all you want are tables of frequencies for the three variables, what would the command look like?
   b. If you also want histograms, statistics, and quartiles for the three variables, what would the command look like?

4. The following statistical report contains many mistakes with respect to APA style:

*40 people completed a survey, including 25.00 ( 62.50 % ) **Women** and 15.00 ( 37.50 % ) **Men**.When asked about their political party affiliation, 17.00 people ( 42.50 % ) chose the **democratic** party, 15.00 ( 37.50 % ) chose the **republican** party, and 8.00 ( 20.00 % ) indicated that they were **independents**.*

*Individuals' ages ranged from 22 to 82 in a positively skewed distribution (Mdn= 44.1, IQR= 36.2 to 50.5). After removing three outliers (ages 78, 81, and 82, well above the next highest age of 63), the distribution was approximately normal (M= 42.3052, SD= 9.8117).*

a. Explain which of the guidelines described in this chapter were violated.
b. Based on the problems you identified, retype this report in APA style.

* * *

5. Create a new SPSS data file and set it up to contain three variables, labeled as follows:
   - SAT = SAT Score (Math + Verbal)
   - Classper = HS Class Percentile
   - Major = Type of Major

     1 = Natural Science
     2 = Social Science
     3 = Humanities
     4 = Business
     5 = Undeclared

6. Enter data for 40 students and save the data file as “Academics.sav”.

| SAT Score (Math + Verbal) | HS Class Percentile | Type of Major | SAT Score (Math + Verbal) | HS Class Percentile | Type of Major |
|---|---|---|---|---|---|
| 1170 | 92 | 4 | 1400 | 58 | 2 |
| 1290 | 89 | 3 | 1340 | 90 | 5 |
| 1190 | 96 | 5 | 1270 | 94 | 2 |
| 1080 | 85 | 4 | 1110 | 96 | 1 |
| 1100 | 95 | 2 | 1120 | 81 | 4 |
| 1160 | 94 | 4 | 1050 | 79 | 4 |
| 1330 | 90 | 1 | 1530 | 88 | 5 |
| 1180 | 81 | 4 | 1250 | 74 | 2 |
| 1140 | 92 | 3 | 1190 | 90 | 1 |
| 1190 | 76 | 1 | 1330 | 95 | 1 |
| 1190 | 87 | 5 | 1030 | 85 | 1 |
| 1100 | 63 | 1 | 1290 | 92 | 2 |
| 1140 | 89 | 3 | 1360 | 91 | 3 |
| 1190 | 71 | 3 | 1400 | 91 | 3 |
| 1430 | 96 | 3 | 1010 | 88 | 4 |
| 1050 | 87 | 4 | 1300 | 98 | 3 |
| 1200 | 81 | 5 | 1270 | 82 | 5 |
| 1170 | 79 | 4 | 890 | 85 | 4 |
| 1250 | 89 | 3 | 1130 | 86 | 4 |
| 1280 | 84 | 1 | 1320 | 95 | 1 |

7. Open a new SPSS syntax file. Enter and run the “freq” command to generate a frequency table for type of major. Enter and run the “freq” command to generate a frequency table, histogram (with normal curve), and descriptive statistics for SAT score and class percentile. Save the syntax file as “Academics.sps”.
8. Examine the results. Are some majors more popular than others? What are the shapes of the distributions for SAT scores and class percentiles? Based on these shapes, what would be the most appropriate measures of central tendency and variability for each variable?
9. Write a report in APA style that describes all three variables.

**Problems 1 – 4 are due at the beginning of class.**

# 4. Standard Scores

## Overview

From study to study, and even from one variable to another within a study, it can be difficult to compare scores to one another. What's missing is a common metric. In this chapter, we'll learn how to convert scores to a common scale using *z* scores. In addition, we'll see how to quantify how rare or how common a particular score is within a normal distribution.

As useful as that can be, it's even more useful to do the same thing for a sample of scores. This chapter will extend the use of *z* scores to sample means, which forms a bridge from descriptive to inferential statistics. We can develop expectations for what will happen if samples are drawn at random from a population and then determine whether the sample we actually observe deviates from these expectations. A *z* score for a sample can tell us whether it seems to be unusual, and therefore whether something other than random sampling is required to account for research findings.

## Standard Scores

Suppose someone takes an IQ test and scores 120. Many people know that the average scores on an IQ test is 100, so 120 is above average. By how much? Is this a very rare score, or fairly common? The IQ scale isn't sufficiently familiar to nonspecialists, in the way that measures of height or weight are, to provide an intuitive sense for how exceptional or commonplace an individual's score is.

The solution to this problem is to convert a **raw score**, the value as measured, to a **standard score**, a value on a common metric. The most popular type of standard score is the ***z* score**, which is defined such that the mean is 0 and the standard deviation is 1. The sign of a *z* score indicates whether the score is above average (a positive *z* score) or below average (a negative *z* score). The absolute value of a *z* score indicates how far the score is from the average, in *SD* units. A *z* score of +1.00 means that someone scored one *SD* above average.

To convert a raw score *X* to a *z* score, use the following formula:

$$z = (X - \mu) / \sigma,$$

where $\mu$ is the population mean and $\sigma$ is the population standard deviation.[11] For example, IQ tests are scored such that in the general population, $\mu = 100$ and $\sigma = 15$. To convert the raw score of $X = 120$ into a *z* score:

$$z = (120 - 100) / 15 = 20 / 15 = 1.33$$

The sign is positive, so this represents an above-average score. The absolute value tells us that the score is 1.33 *SD*s above average.

[11] The Greek letters $\mu$ and $\sigma$ are pronounced "mu" and "sigma", respectively.

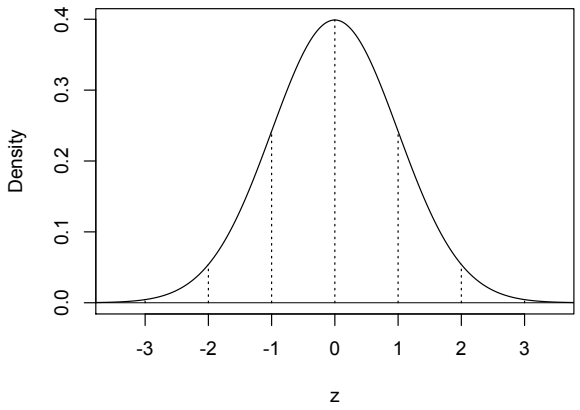
As another example, consider an IQ score of 70, the threshold applied to help diagnose mental retardation. Converting this to a *z* score reveals how far below average this is:

$z = (70 - 100) / 15 = -30 / 15 = -2.00$

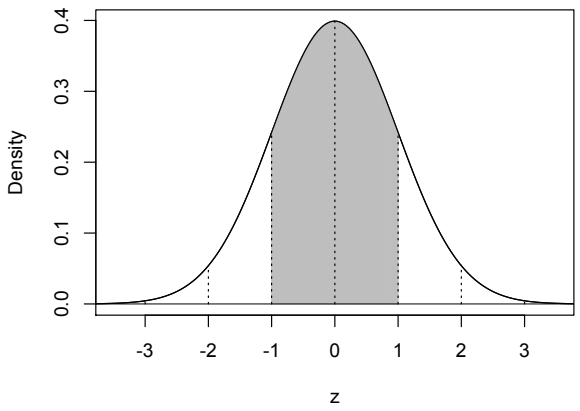
The sign is negative, confirming that this is a below-average score, and the absolute value shows that this is 2 *SD*s below average. How rare is such a score? The next step in helping to quantify that requires an assumption that IQ scores are normally distributed in the population.

## Normal Curves

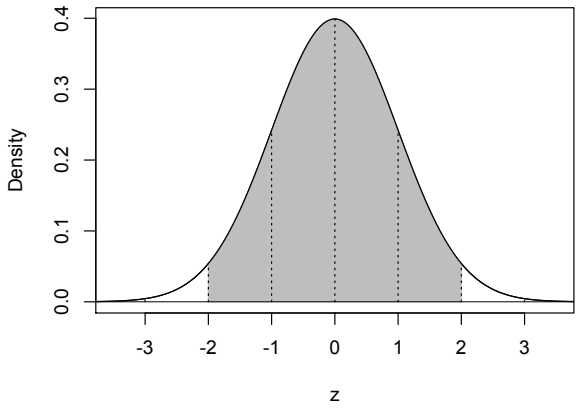
Though people sometimes refer to "the normal curve," there is actually a family of normal curves, not just one. All normal curves share an identical form, the classic bell shape. Most scores are located close to the center, with frequencies tapering off toward the tails. A normal curve is symmetric, and nearly all scores lie within a few *SD*s on either side of the mean. Here's a density plot for a normal distribution of *z* scores:



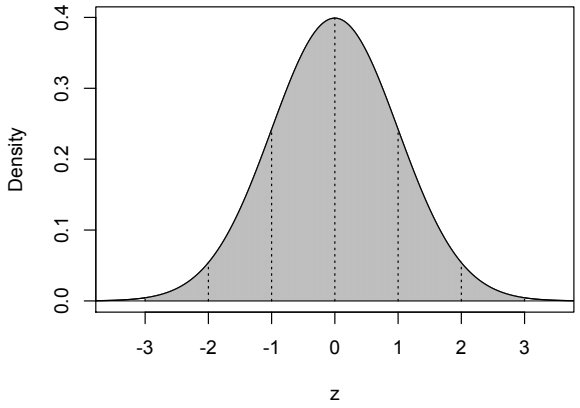

The dotted lines are not part of the normal curve itself. They've been added to highlight some helpful reference points within the distribution. Specifically, they're plotted at each *SD* unit along the *x* axis. These reference points correspond to a few percentages that are worth remembering. First, about 68% of scores fall within 1 *SD* of the mean (i.e., in the range from $z = -1.00$ to $z = 1.00$). That's roughly two-thirds of scores in the ±1 *SD* range. Here's what this looks like:



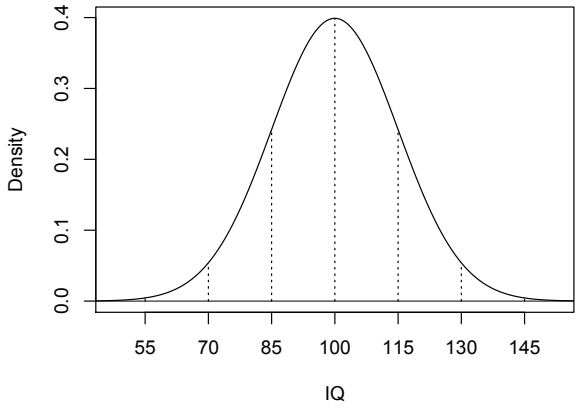
Another useful approximation to remember is that about 95% of all scores fall within the ±2 *SD* range. In other words, it's fairly unusual for a score to be more than 2 *SD* away from the mean. Here's what this looks like:




By the time you extend this to the ±3 *SD* range, about 99.7% of all scores are included. Here's what this looks like:

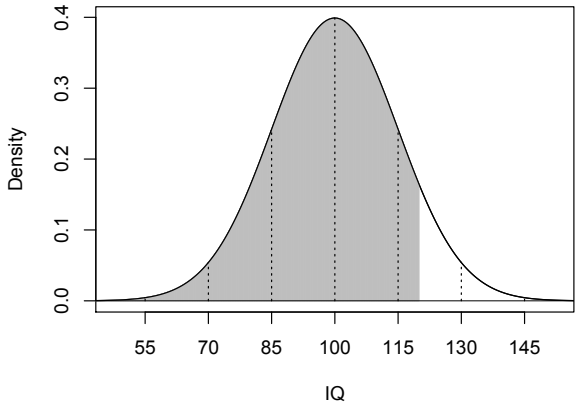



All normal curves share an identical shape and include the same proportions of scores within the same *SD* intervals. Normal curves differ from one another according to the scale along the *x* axis, which is determined by μ and σ. For example, a normal distribution of IQ scores looks like this:



The only difference between this and the normal curve for *z* scores is that instead of using $\mu = 0$ and $\sigma = 1$, we now use $\mu = 100$ and $\sigma = 15$. The shape of the curve is unchanged.

Now we can revisit the question of how unusual a particular IQ score is. For an IQ of 120, which corresponds to a *z* score of 1.33, we can estimate its percentile based on a simple approximation. First, take a look at a graph plotting this value:




We know that 50% of all scores fall below the mean. That leaves the region extending from the mean to $z = 1.33$. If about 68% of scores fall within ±1 *SD* of the mean, then half of them (34%) fall between the mean and $z = 1.00$. Adding this to the 50% of scores below the mean gives 84% of all scores below a *z* score of 1.00. A *z* score of 1.33 is a little higher still, so the percentile might be closer to 90%. In other words, an IQ of 120 is higher than the IQ scores for about 90% of the population. In a moment, we'll see how to identify the precise value.

For an IQ of 70, which corresponds to a *z* score of -2.00, we can even more easily estimate its percentile based on an approximation. If 95% of all scores fall within ±2 *SD* of the mean, then only 5% lie beyond this region, with about 2.5% in the upper tail and 2.5% in the lower tail. In other words, about 2.5% of IQ scores in the population are below 70.

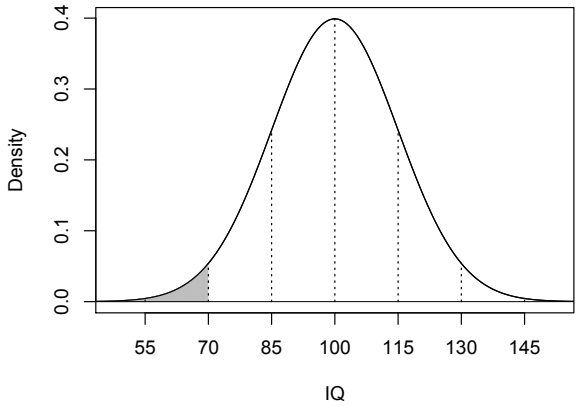
## Unit Normal Table

Though we can estimate the percent of scores in various regions of a normal distribution using some approximate values, it's not hard to obtain a precise value. The final tool for doing this is called the **unit normal table**, which lists the proportion of the normal curve in the tail beyond any particular *z* score (see Appendix A). This is called a unit normal table because it's based specifically on the normal distribution for a population with $\mu = 0$ and $\sigma = 1$, the scale of *z* scores. The fact that $\sigma = 1$ puts the "unit" in "unit normal table".

To look up a *z* score in this table, you go to the row that contains the beginning of the *z* score (whole number and first decimal place) and the column that contains its second decimal place. For example, we calculated earlier that an IQ score of 120 equals a *z* score of 1.33. If you look in the row labeled "1.30" and the column labeled ".03", which combine to form the *z* score of 1.33, the value listed in the table is .09176. Remember, though, that this is the proportion of scores in the tail of the distribution, which in this case means the proportion of scores falling above $z = 1.33$. To calculate the proportion that falls below 1.33, simply subtract the tail from the area under the whole curve, which is 1.00. This gives

1.00 – .09176 = .90834. To convert that proportion into a percent, just move the decimal two places to the right to get 90.834%. That's very close to the percentile that was estimated earlier. It's reassuring to see this correspondence, and doing a rough approximation before calculating anything is always a good idea. If your calculated answer is far from your estimation, this suggests that you calculated something wrong. For example, it's easy to misplace a decimal point or negative sign. But if you make a mistake like that, your answer is likely to be way off, and you'll notice the mistake because it doesn't come close to your approximation.

We can use the same table to find the percentile for an IQ of 70, which can be graphed like this:

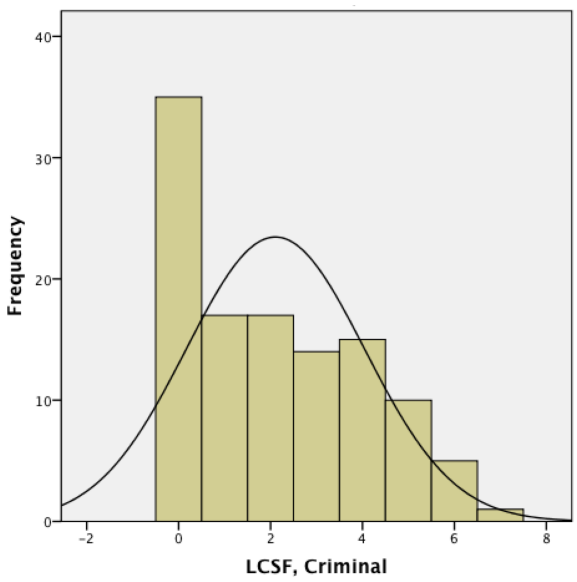


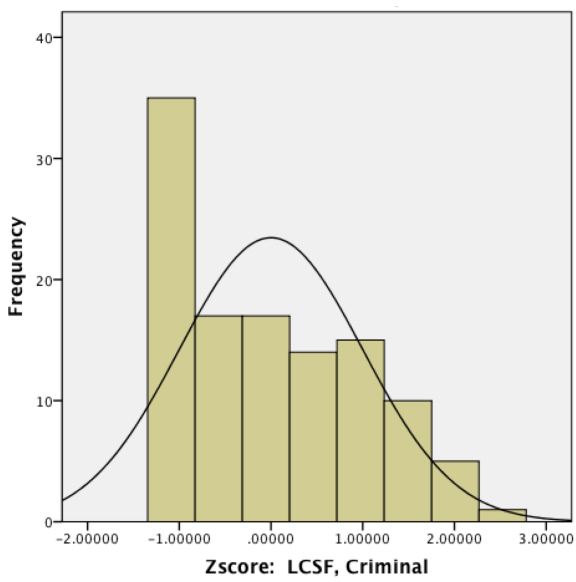

Earlier we calculated that $z$ = -2.00, so we look in the row labeled "2.00" and the column labeled ".00", which combine to form 2.00. Because the normal curve is symmetric around $z$ = 0, the unit normal table works the same way for positive or negative $z$ scores. The proportion of scores below $z$ = -2.00 is listed as .02275. Moving the decimal two places to the right converts that to a percentile of 2.275%, once again close to what was estimated earlier (2.5%).

## Standardizing and Normalizing

There's an easy mistake to make when thinking about what happens when an entire distribution of scores is converted to $z$ scores, or **standardized**. Specifically, it's easy to mistakenly believe that once raw scores have been converted to $z$ scores, their distribution becomes normal. In fact, the act of standardizing a distribution has no effect on the shape of a distribution. If the raw scores were skewed, the standardized scores will skewed, too.

Standardizing scores is a **linear transformation**, which means that the equation of a straight line is used to turn a raw score into a $z$ score. This preserves not only the rank-ordering of cases, but also their relative distances from one another. If none of the scores are squeezed closer together or spread further apart from one another, the shape of the distribution will not be affected by the transformation. All that would happen is that the scale on the $x$ axis of a histogram or density plot would change. For example, here are histograms for the LCSF-Criminal scores from the parole dataset. First is the distribution of raw scores, followed by the distribution of standardized scores:




Naturally, standardizing scores will change the *M* and *SD* of the distribution. If raw scores are converted to *z* scores, for example, the new values will be *M* = 0 and *SD* = 1, by definition. But this doesn't mean that the distribution itself becomes normal in shape. No matter how much a distribution diverges from normality, standardizing it will yield *M* = 0 and *SD* = 1 within the new, equally non-normal distribution.

The reason this is important to understand is that the existence of a *z* score doesn't necessarily mean you can use the unit normal table to calculate what proportion of scores fall above or below it. Using this table requires the assumption of a normal distribution.

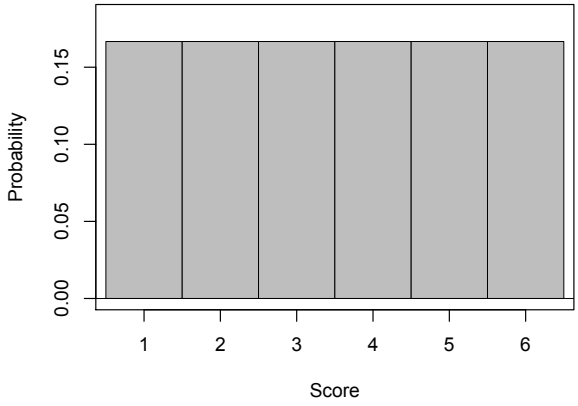
If the goal is to **normalize** a set of scores, to change the shape of the distribution so that it better approximates normality, then a **nonlinear transformation** is required. Some scores will be squeezed closer together, and others spread apart relative to one another. For example, taking the square root of all scores would have a more pronounced effect on the larger ones than on the smaller ones. Scores of 1, 4, 9, 16, 25 would become 1, 2, 3, 4, 5. The distances between successive scores changed from 3, 5, 7, and 9 to 1, 1, 1, and 1. All gaps shrunk, but larger gaps shrunk more. This could be helpful if, for example, an original distribution was positively skewed and you wanted to reduce the skew.

There are an infinite variety of nonlinear transformations that one might use to try to normalize a distribution. Determining when and how to use them is not the goal here. The point is simply to understand the critical difference between standardizing—which changes only the scaling of a variable, not the shape of its distribution—and normalizing—which changes the scaling of a variable and the shape of its distribution. Don't be fooled by the presence of *z* scores into assuming a variable is normally distributed, that's something that still has to be checked.

## Sampling Distributions

Standard scores are an extremely useful way to place individual scores into context and help determine how common or rare they are. Because research seldom involves the study of single subjects, it's even more useful to convert the mean for an entire sample of subjects into a standard score.

To see how this is done, let's begin with a simple hypothetical question: What would you expect to happen if you rolled a die an infinite number of times, recording the number rolled each time? What would the distribution of scores look like? With a moment's thought, this isn't very difficult to figure out. The distribution would be uniform, with equal frequencies for scores of 1, 2, 3, 4, 5, and 6. Here's a histogram that shows the scores according to their probability, or relative frequency, of occurring:




This is a theoretical distribution, what you'd expect for a population. What would be the central tendency and variability of the distribution? That's not as simple as figuring out the shape of the distribution, but it's not that hard, either. The distribution is symmetric, so a good choice of measures would be the mean and standard deviation. The former is easy to calculate: Because the scores are all equally likely, $\mu = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.50$. Calculating $\sigma$ isn't so hard, either, though there are more steps. As explained in an earlier chapter, you'd need to follow a four-step process to calculate that $\sigma = 1.71$.[12]

What we've done is established a theoretical distribution that shows us what we would expect to happen if we sampled from it at random. In this case, we've sampled individual scores. Rolling one die is like studying one person. What happens if we roll a pair of dice and record their mean? That's like taking a sample of $N = 2$ people in a study, and examining their average response.

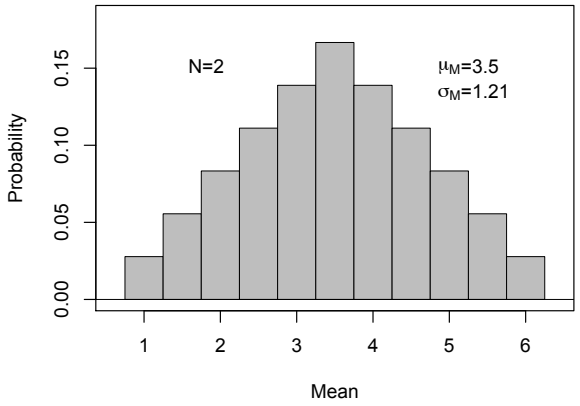
What we're doing now is looking at what's called a **sampling distribution**. This shows us how a statistic is distributed if you calculate it for many samples from the same population. It allows us to see how much sampling error there would be, how much random variation would occur just due to the luck of the draw based on which cases happen to be selected for a study. The standard deviation of a sampling distribution is called the **standard error** of that statistic. Just like the standard deviation for a sample is the typical distance from a score to the mean, the standard error for a sampling distribution is the typical distance from a statistic (calculated in a sample) to the population mean for that statistic.

For example, if we graph a distribution of sample means, the average of those is labeled $\mu_M$. The "$\mu$" indicates we're taking a population mean of something, and the subscript "M" indicates that the statistic whose mean we're examining is sample means. The standard error of this distribution is labeled $\sigma_M$. The notation is the same as for $\mu_M$ in that "$\sigma$"

[12] We're using $\mu$ and $\sigma$ rather than *M* and *SD* because this is a population, not a sample. Remember that this means we'd divide by *N* rather than *N* – 1 to calculate the standard deviation.

indicates we're taking a population standard deviation of something and the subscript "M" indicates the statistic whose variability we're examining is sample means.

So let's get back to the smallest samples possible, *N* = 2, and see what happens when we roll two dice. Here's the histogram for this sampling distribution of the mean:



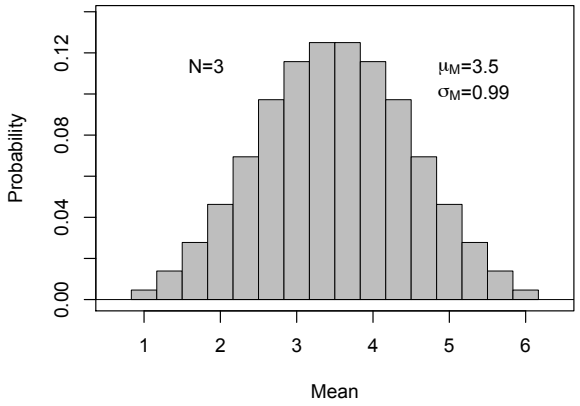

Notice that the *x* axis is now labeled with a statistic—in this case, "Mean"—rather than "Score". This shows that we're no longer looking at a distribution of individual scores. This is a distribution of sample means.

Three things about this sampling distribution deserve mention. First, whereas the population distribution of individual scores was uniform, the sampling distribution of the mean is not. There's only a single way to get an average roll of 1.00 (1 + 1) or 6.00 (6 + 6), but there are six ways to get an average roll of 3.50 (1 + 6, 2 + 5, 3 + 4, 4 + 3, 5 + 2, or 6 + 1). Thus, the distribution is peaked in the center and tapers off toward both ends.

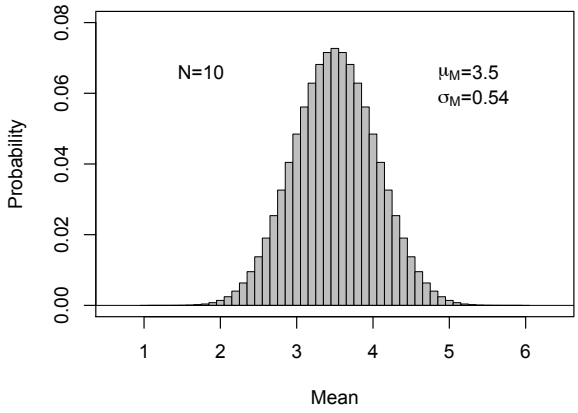
Second, $\mu_M$ equals $\mu$. The average of all the sample means is the same as the average in the original population of scores from which the samples are drawn.

Third, $\sigma_M$ is smaller than $\sigma$. Sample means very less than do individual scores. For example, roll a die once and there's a 1 in 6 chance of getting a 1 or a 6. Roll two dice and there's only a 1 in 36 chance that their average will be 1.00 or 6.00. Extreme scores become less common as sample size increases.

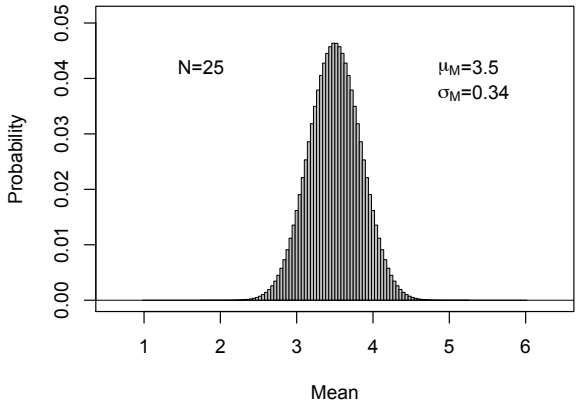
Let's look at a few more graphs to see these patterns develop. Here's the histogram showing what you'd expect if you rolled *N* = 3 dice:




Next, let's jump up to what you'd expect if you rolled *N* = 10 dice:




Finally, here's what you'd expect if you rolled $N = 25$ dice:




Several patterns become clear. In the following subsections, three important observations about the sampling distribution of the mean are described. Collectively, these are referred to as the **central limit theorem**.

## Shape

With larger samples, the sampling distribution of the mean becomes more normal in shape. Importantly, this will happen regardless of the shape of the population distribution from which scores are sampled. In the example shown above, even though the population distribution is uniform, by the time you get to samples of $N = 25$, the sampling distribution of the mean is almost perfectly normally distributed.

If the population distribution is itself normal, then even with the smallest of samples the sampling distribution will also be normal. To the extent that the population distribution deviates from normality, such as being skewed, larger samples will be needed for the sampling distribution of the mean to approximate normality well. One rule of thumb is that this usually happens by the time you reach $N = 30$.

## Central Tendency

The center of the sampling distribution of the mean will always be the population mean ($\mu_M = \mu$). Sample means will vary from one to the next, but they are an unbiased estimate of the population mean. In each of the histograms shown above for rolling dice, $\mu_M = \mu = 3.50$.

### Standard Error

The variability of the sampling distribution of the mean decreases with sample size. In other words, sample means estimate population means more accurately when they're based on larger samples. Collecting more data reduces the amount of sampling error. The usual relationship between sampling error and sample size holds here: The standard error of the mean ($\sigma_M$) decreases as a function of the square root of the sample size. Specifically, $\sigma_M = \sigma$ / sqrt($N$). You can easily verify that this formula yields the variability shown in each of the histograms shown above for rolling dice. For example, with samples of $N = 25$, $\sigma_M$ = 1.71 / sqrt(25) = 0.34.

## *z* Scores for Samples

Recall that for an individual score $X$, we can convert it to a $z$ score using this formula:

$$z = (X - \mu) / \sigma$$

This tells us how many standard deviations above or below the population mean the score is. For a sample mean $M$, the $z$ score formula is adapted like this:

$$z = (M - \mu) / \sigma_M$$

The sample mean ($M$) is substituted for an individual's score ($X$), and the standard error of the mean ($\sigma_M$) is substituted for the standard deviation ($\sigma$). The resulting value for $z$ tells us how many standard errors above or below the population mean a sample mean is.

For example, earlier we found that an IQ score of 120 is higher than about 90.8% of normally distributed scores in the population with $\mu = 100$ and $\sigma = 15$. We calculated that by converting the IQ score to a $z$ score (1.33 in this case) and using the unit normal table to find the proportion of scores below that level. Suppose we take a random sample of $N = 25$ people from this population and find their mean IQ to be 120. Is this likely to occur? If not, can we say anything about how rare an outcome this would be?

The first step, once again, is to calculate $z$. To do that, we need to know the standard error of the mean for $N = 25$. That's easily calculated:

$$\sigma_M = \sigma / \text{sqrt}(N) = 15 / \text{sqrt}(25) = 3.00$$

The next step is to calculate $z$:

$$z = (M - \mu) / \sigma_M = (120 - 100) / 3.00 = 6.67$$

The final step is to find out what proportion of scores are below that level. Thanks to the central limit theorem, we can safely assume that the sampling distribution of the mean is normally distributed. Because of that, we can use the unit normal table. In this case, as you can see, the table doesn't even have entries for $z$ scores more extreme than ±4.00. The reason is that the proportion of scores in the tail is vanishingly close to 0 when you get that far from the mean. For our $z$ value of 6.67, the proportion in the tail is very close to 0. That means it's next to impossible for this to happen, for a random sample to score this high or higher.

An IQ score of 120 or higher is somewhat exceptional for an individual. Just under 10% of all people in a population with $\mu = 100$ and $\sigma = 15$ will have IQs that high. But if we take a

random sample of 25 people from this same population, it's almost impossible for their mean to be as high as 120. You can see the same thing happening with the illustrations involving rolling dice presented earlier. The chance of rolling a 1 on a single die is 1/6, or a probability of .17. The chance of getting a mean of 1.00 when rolling $N = 10$ or $N = 25$ dice is extremely close to 0.

The important point is that, all else being equal, the larger the sample size, the less likely it is for an extreme sample mean to occur by chance. And we can quantify just how unlikely it is to observe a sample mean as or more extreme than any particular value. This is how we'll begin to use inferential statistics to test hypotheses in the next chapter.

## Problems

The following problems refer to Zeke's test scores in five different classes:

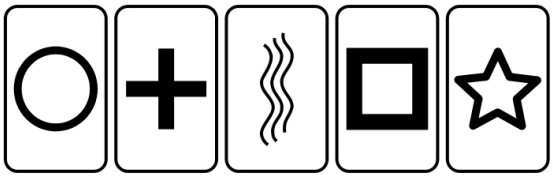
| Class | Score | μ | σ |
|---|---|---|---|
| Chemistry | 80 | 65 | 10 |
| English Lit. | 85 | 79 | 12 |
| History | 87 | 75 | 16 |
| Psychology | 85 | 70 | 15 |
| Social Work | 86 | 90 | 12 |

1. Why are μ and σ listed in this table, rather than *M* and *SD*?
2. In a sense, Zeke seems to have done very well on the test in his Social Work class. In another sense, why might it be misleading to say that?
3. For each class, sketch a normal curve with the class μ and σ. Then, add a vertical line to mark where Zeke's score falls. Shade the region of the graph below Zeke's score. Finally, estimate what percentage of the graph has been shaded. This is a visual estimate of the percentile. Even if you do this only roughly, it will be very helpful in checking the answers you get later by doing calculations.
4. Calculate a *z* score for Zeke's performance in each class.
5. Use the unit normal table to calculate the percentile for each of these test scores. Remember that the proportion listed in the table is for the tail of the distribution, which for positive *z* scores will be the proportion of scores above (not below) that point. Check your answers against the estimates you made in #3. Minor discrepancies are expected, but if the answer is very far from the estimate you should double-check your work to figure out what went wrong.
6. Based on the percentiles, in which class did Zeke score highest relative to the other students in that class?
7. Based on the percentiles, in which class did Zeke score lowest relative to the other students in that class?

8. Based on the percentiles, in which class did Zeke score closest to the average of the other students in that class?
9. Based on the $\mu$ and $\sigma$ for the Social Work class, how can you tell whether the distribution of scores is symmetric or skewed?
10. Suppose that scores in the Social Work class are standardized by converting to *z* scores. How would that affect the shape of the distribution?

* * *

The following problems refer to a study in a parapsychology laboratory. The investigator would like to test for a particular type of ESP known as "precognition." He uses a deck of Zener cards, which is popular in ESP research: five cards each of five different symbols (shown below), for a total of 25 cards. A subject is shown the deck of cards and then asked to shuffle them. Once shuffled, the task is to "see" in advance what symbol is on each card as the experimenter picks it up from the pile. Each prediction is recorded, and then the card is revealed to the subject so that the prediction can be scored correct or incorrect. If ESP did not exist, and subjects simply guessed a symbol at random on each trial, you would expect the average number of correct responses to be $\mu = 5.00$ (25 cards × 20% chance of guessing correctly for each = 5), with a standard deviation of $\sigma = 2.00$.[13]



11. Suppose that an individual scores $X = 8$ cards correct. What is the *z* score for this person's level of performance?
12. What is the probability of doing at least this well if the person was only guessing? Use the unit normal table to figure this out.
13. Based on your answer to #12, does this impress you as unusually good performance, better than guessing cards at random? Why or why not?
14. Suppose that a sample of $N = 9$ individuals averages $M = 8$ cards correct. What is the *z* score for this level of performance?
15. What is the probability of doing at least this well if these individuals were only guessing? Use the unit normal table to figure this out.
16. Based on your answer to #15, does this impress you as unusually good performance, better than guessing cards at random? Why or why not?
17. Revisit the design and procedure of this research. Can you think of any ways that subjects could "cheat," or score more than 5 cards correct even though they do not possess ESP?

[13] For the curious, this is calculated as sqrt($N \times p \times q$), where $N$ is the number of trials (here, 25 cards per subject), $p$ is the probability of a correct guess on each trial, and $q = 1 - p$. Thus, sqrt($25 \times .20 \times .80$) = 2.00.

18. How can the design and procedure be improved to prevent the tricks you identified in #17?

* * *

19. In statistics, we learn to use *z* scores because they so clearly communicate (1) whether a score is above or below average (the sign of the *z* score shows this) and (2) how far from average (the size of the *z* score shows this). On the other hand, standardized tests (e.g., SAT, ACT, IQ) usually are scaled such that for the population of individuals who take the test, score distributions will approximate normality with $\mu$ much larger than 0 and $\sigma$ much larger than 1. Why are standardized tests never scaled in *z* score units?

**Problems 1 – 16 are due at the beginning of class.**

# 5. Statistical Decision Making

## Overview

With a way to determine how much a sample mean differs from a population mean, we're now in a position to begin to test hypotheses. The key question to ask is whether the difference we observe between the means is more than we'd expect due to sampling error alone. In simpler terms, could chance account for our findings? If not, then we have some evidence that something more than chance, something systematic, is responsible. This chapter introduces the framework that will be used for all inferential statistics.

As a running example, we'll revisit the precognition experiment from an earlier chapter. Using an improved procedure in which subjects are not allowed to touch the cards, which remain face down until all 25 guesses have been recorded, Zeke tests 16 subjects who believe they possess precognitive ability (the type of ESP that involves seeing the future). Here are their scores, the number of correct responses:

3, 8, 8, 9, 6, 6, 5, 4, 5, 6, 8, 8, 6, 3, 5, 2

In what follows, we'll examine how inferential statistics are used to reach a decision about whether these data support Zeke's hypothesis that precognition exists. The test we'll use is called a **one sample *z* test**, and we'll see it put into practice in the four steps of statistical decision making.

## Step 1: Construct the Statistical Hypotheses

The first step in statistical testing is to establish the **statistical hypotheses**. It's important to understand that these are not the same as what one might be called the **researcher's hypothesis**, what the researcher is actually proposing. In this case, the researcher's hypothesis is that precognition exists. The statistical hypothesis that will be tested using data is called the **null hypothesis** ($H_0$), and it corresponds to the absence of any systematic effect. In this case, the null hypothesis is that guessing at random can account for subjects' performance, that neither ESP nor any other systematic effect is required. Naturally, Zeke hopes to obtain data inconsistent with $H_0$, which would allow him to reject it. If that happens, he would tentatively accept the **alternative hypothesis** ($H_1$), which corresponds to the presence of a systematic effect.

In this case, $H_0$ would represent performance no better than chance-level guessing. How many cards would someone be expected to get right just by guessing? Well, with 5 types of cards, you'd have a 1 in 5 chance of guessing one correctly. Guessing for all 25 cards would yield an expected $\mu = 25 \times (1/5) = 5$ cards correct. $H_0$, then, would include any value less than or equal to 5. $H_1$ would represent performance better than this, including any value greater than 5. The statistical hypotheses can be stated concisely:

$H_0$: $\mu \leq 5$
$H_1$: $\mu > 5$

Several important points about this notation need to be understood.

## Population Parameters

Statistical hypotheses involve population parameters (here, μ) and not sample statistics (such as *M*). This is what it means to do inferential statistics: We use statistics calculated for a sample of data to test hypotheses that deal with populations. That's why it would be mistaken to define the statistical hypotheses as $H_0$: $M \leq 5$ and $H_1$: $M > 5$. If all we wanted to know is whether a sample mean is greater than 5, we wouldn't need inferential statistics at all. We'd just calculate the sample mean ($M = 5.75$ for these 16 subjects), compare it to 5, and reach a conclusion. But that ignores the whole point of doing a statistical test, to examine the role of sampling error. Is the finding that $M = 5.75$ really far enough from $\mu = 5$ to reject $H_0$? Might these means differ purely because of sampling error? We need to ask whether the observed difference is more than we'd expect by chance. If the possibility of this being a fluke is sufficiently small, then we can reject $H_0$ and tentatively accept $H_1$.

## All Possible Outcomes

Statistical hypotheses must cover all possible outcomes. In this case, if μ is less than or equal to 5, that would support $H_0$. If μ is greater than 5, that would support $H_1$. All possibilities are covered. It would be mistaken to define the statistical hypotheses as $H_0$: $\mu < 5$ and $H_1$: $\mu > 5$. This fails to account for the possibility of $\mu = 5$.

## Directional and Nondirectional Hypotheses

Many statistical tests—including the *z* test—allow you to construct either **directional** or **nondirectional** hypotheses. As the names imply, the key is whether you're predicting the direction of an effect. In this case, we'd use directional hypotheses because only performance better than chance would support the existence of precognition. Performance equal to or worse than chance defines $H_0$: $\mu \leq 5$. This leaves performance above chance to define $H_1$: $\mu > 5$.

The present example is actually unusual in that nondirectional hypotheses are far more common in data analysis. They're the default. This is partly because it's wise to be open to the possibility that an effect might occur in either direction (e.g., even a well-intentioned treatment might cause harm), and partly because it keeps researchers honest. Scientists are only human, and they're subject to the same kinds of cognitive and motivational biases as anyone else. They like to obtain support for their hypotheses. The problem with directional hypotheses is that you never know whether someone predicted the direction before looking at the data. As we'll see, there's an unfair statistical advantage if one looks at the data first, notes the direction of the results, and then "predicts" that with a directional hypothesis. Because it's seldom possible to know for sure whether an investigator made a directional prediction before seeing the data, making nondirectional hypotheses the norm prevents anyone from taking advantage of this kind of statistical cheating.

Nondirectional statistical hypotheses would take the form of $H_0$: $\mu = 5$ and $H_1$: $\mu \neq 5$. In this case, that really wouldn't make sense because it would entail rejecting $H_0$ even if performance was worse than chance-level guessing. This is a rare instance where logical considerations argue against using nondirectional hypotheses and you could justify using directional hypotheses instead. In short, consider nondirectional hypotheses the default, and only use directional hypotheses if you believe there is very strong justification.

### Equality and $H_0$

$H_0$ includes equality. It's called the null hypothesis for just this reason, that it includes the possibility of no effect. Whether you use directional or nondirectional hypotheses, the possibility of $\mu = 5$ must be included as part of $H_0$.

### Defining $H_0$ and $H_1$

A final point may seem superficial, but it's important nonetheless. Writing "$H_0$:" means "the null hypothesis is defined as…" A common mistake is to write something like "$H_0 = 5$". This is meaningless because it contains no population parameters. The null hypothesis itself is not a numerical value. Rather, it's an expression stated in terms of one or more population parameters.

## Step 2: Establish the Decision Threshold

The second step in statistical testing is to establish a **decision threshold**. How strong do the results have to be to reject $H_0$? Consider, by analogy, the decision that a jury must reach in a criminal trial. The defendant is considered innocent until proven guilty "beyond a reasonable doubt". The null hypothesis is innocence, and it can only be rejected if the evidence surpasses this decision threshold.

Whereas that's a subjective decision, in statistical testing we set an objective threshold using what's called an **$\alpha$ level**.[14] The $\alpha$ level is the probability that a result could occur by chance if $H_0$ is true. This is where sampling distributions come into play. We know what results we'd expect to observe for randomly selected samples from a specified population. If the results for our sample of data fall well out in one of the tails of this distribution, that suggests they're incompatible with $H_0$. For example, earlier we saw that although it's possible to roll 10 dice and get an average score of 6.00, that's extremely unlikely. In fact, the only way it could happen is if all 10 dice came up as 6s, and there's only a $(1/6)^{10}$ chance of that, which is a probability of .0000000165. When we see such a minute probability that something would happen by chance, we conclude that something other than chance is responsible. In this case, perhaps the dice are loaded.

In research, we don't require a probability as low as .0000000165 to reject $H_0$. Usually $\alpha$ is set at .05, meaning that if the probability of obtaining results as (or more) extreme than what we observe is less than .05, we'll reject $H_0$. Sometimes, to be more rigorous, the $\alpha$ level is set at .01. This makes it harder to reject $H_0$, but it also means that when we do reject $H_0$ we have stronger evidence against it. More will be said about this trade-off later.

Once you've chosen your $\alpha$ level—and you should use .05 unless there's a strong reason to lower it—you determine the **critical region** of the sampling distribution of expected results if $H_0$ is true. The critical region is where the results must fall to enable you to reject $H_0$. For a $z$ test, you want to know how large $z$ must be to reject $H_0$.

Recall that a nondirectional hypothesis, also known as a **2-tailed test**, is the norm. This means that the critical region falls in both tails of the sampling distribution, each containing a proportion of scores equal to $\alpha / 2$. For a nondirectional test with $\alpha = .05$, that would leave a proportion of .025 in each tail. Using the unit normal table, you'd look for the entry

[14] The Greek letter $\alpha$ is pronounced "alpha".

closest to .025 and find that this corresponds to $z = 1.96$. Because this is a 2-tailed test, the critical region includes the left tail ($z < -1.96$) and the right tail ($z > 1.96$). The simplest way to express the critical region for the nondirectional $z$ test with $\alpha = .05$ would be $|z| > 1.96$. The absolute value notation indicates that the size of $z$ must exceed 1.96, regardless of sign, thereby including both tails. For a nondirectional $z$ test with $\alpha = .01$, the critical region would be $|z| > 2.58$.

For a directional hypothesis, also known as a **1-tailed test**, you'd locate the $z$ score beyond which the proportion of cases equals your $\alpha$ level. For a directional $z$ test with $\alpha = .05$, this would be a $z$ score of 1.64 (or -1.64, depending on which direction you're predicting.) The critical region would be defined as $z > 1.64$ (or $z < -1.64$). For $\alpha = .01$, the critical region would be a $z > 2.33$ (or $z < -2.33$).

Notice that the decision thresholds for 1-tailed tests are less extreme than those for 2-tailed tests. With $\alpha = .05$, you'd only need to find $z > 1.64$ rather than $z > 1.96$ to reject $H_0$. This is the statistical advantage of using directional hypotheses. The same is true for using larger rather than smaller $\alpha$ levels. For 2-tailed $z$ tests, using $\alpha = .05$ means you'd only need to find $|z| > 1.96$ to reject $H_0$, but using $\alpha = .01$ means you'd need to find $|z| > 2.58$. This is the statistical advantage of using larger $\alpha$ levels. So why don't researchers tend to use 1-tailed tests with larger $\alpha$ levels? The short answer is that this could be abused.

If we could trust that researchers would never look at the data before performing their statistical tests, 1-tailed tests and larger $\alpha$ levels might be more acceptable, more common. But we can't be sure that everyone will be so honest. The scientific method, which includes the practice of statistical decision making, is designed to prevent human biases from affecting conclusions to the greatest extent possible. To keep us from fooling ourselves, we restrict the options in statistical testing. This minimizes the impact of any temptation to peek at the data and then establish a decision threshold that allows us to reject $H_0$. Thus, 2-tailed tests with $\alpha = .05$ are expected unless you can provide a strong justification for doing otherwise. These are arbitrary norms—particularly the choice of $\alpha = .05$, which is not special in any way—but following them supports the integrity of the scientific method.

## Step 3: Collect Data and Compute the Statistic

The first two steps in statistical testing can, and arguably should, be done in advance of data collection. You don't need to have any data in hand to state the statistical hypotheses and establish a decision threshold. The third step is to gather your data and calculate the statistic.

To calculate a $z$ value for our data, we use the equation from an earlier chapter:

$$z = (M - \mu) / \sigma_M$$

The numerator is simple, that's $5.75 - 5 = 0.75$. This represents the difference between what we've observed in our data and what we expect in the population if precognition doesn't exist. For the denominator, recall that $\sigma_M = \sigma / \text{sqrt}(N)$. In this case, $\sigma = 2$, so the denominator is $2 / \text{sqrt}(16) = 0.50$. This represents the typical distance between a sample mean and the population mean if precognition doesn't exist. The ratio between the observed difference between the means of 0.75 and the typical distance to the population

mean of 0.50 gives us a $z$ value of 0.75 / 0.50 = 1.50. In other words, the difference between the means we observed in our data is 1.50 times as large as what we'd expect by chance.

## Step 4: Reach a Decision

The final step in statistical testing is to make a decision. This is based entirely on whether the statistic falls in the critical region. Recall that for a 1-tailed $z$ test with $\alpha = .05$, $z$ needs to exceed 1.64 to reject $H_0$. We found that $z = 1.50$, therefore we do not reject $H_0$. By retaining $H_0$, we acknowledge that chance-level guessing can account for our results.

Whenever we reach a statistical decision, there's some possibility that it's mistaken. This is because we're using probability, which will not provide us with certainty. There are two kinds of mistakes that we can make.

If we reject $H_0$ when it's actually true, this is known as a **Type I error**. We'd be concluding that a systematic effect exists when really it does not. This is analogous to a **false alarm**, such as when a smoke detector sounds the alarm when there's no fire. Perhaps the smoke detector is too sensitive to small amounts of smoke that can occur when there's no fire (e.g., crumbs in the toaster oven giving off some smoke during routine operation). The same thing can happen in statistics: By using $\alpha = .05$, we're tacitly accepting that even when $H_0$ is true, we'll reject it 5% of the time. The usual practice in statistical testing provides some sensitivity to detect real effects, but it will result in a lot of false alarms when there are not.

If we retain $H_0$ when it's actually false, this is known as a **Type II error**. We'd be concluding that there is no systematic effect when really there is. This is analogous to a **miss**, such as when a smoke detector fails to sound the alarm when there is a fire. Perhaps the smoke detector isn't sensitive enough to the smoke produced by a fire (e.g., the battery needs replacing). The same thing can happen in statistics: By requiring that results be in the critical region in order to reject $H_0$, we'll fail to reject it sometimes even when it's really false. Especially in small samples, statistical testing is fairly insensitive to real effects, missing them quite often.

## Minimizing Type I and Type II Errors

Because we rely on probabilities to reach statistical decisions, we cannot eliminate the possibility of making Type I and Type II errors. However, we can take steps to minimize the risks of one, the other, or both kinds of error.

### Setting the $\alpha$ Level

When you choose an $\alpha$ level, you're implicitly making a trade-off between Type I and Type II errors. Changing $\alpha$ can't reduce both risks at the same time, but it does trade a lower likelihood of one kind of mistake for a greater likelihood of the other. Normally, you leave $\alpha$ at .05, but there are circumstances under which you might adjust it.

If you're doing exploratory research, you might be more concerned with missing effects that deserve further study (Type II errors) than with suggesting further study for apparent effects that turn out to be mistaken (Type I errors). If the primary goal is to suggest ideas rather than to rigorously test them, you might consider using a larger $\alpha$ level. For example,

if you raise $\alpha$ from .05 to .10, you're demanding weaker evidence to reject $H_0$. For a 2-tailed $z$ test, the size of the critical region increases from $|z| > 1.96$ (fairly small tails) to $|z| > 1.64$ (somewhat larger tails). By making it easier to reject $H_0$, you reduce the chance of a Type II error. At the same time, though, you increase the chance of a Type I error. It's like adjusting your smoke detector so that it's more sensitive to smoke: You won't miss as many fires, but you'll get more false alarms.

If you're doing hypothesis-testing research, you might be most concerned with providing false support for nonexistent effects (Type I errors). If the primary goal is to rigorously test ideas, you might consider using a smaller $\alpha$ level. For example, if you lower $\alpha$ from .05 to .01, you're demanding stronger evidence to reject $H_0$. For a 2-tailed $z$ test, the size of the critical region shrinks from $|z| > 1.96$ (fairly small tails) to $|z| > 2.33$ (even smaller tails). By making it harder to reject $H_0$, you reduce the chance of a Type I error. At the same time, though, you increase the chance of a Type II error. It's like adjusting your smoke detector so that it's less sensitive to smoke: You won't get as many false alarms, but you'll miss more actual fires.

### Increasing Sample Size

When you collect more data, you reduce the amount of sampling error. This makes it easier to detect an effect if one really exists, meaning that you reduce the chances of a Type II error. The probability of making a Type I error is controlled entirely by the choice of an $\alpha$ level, so increasing sample size is a way to reduce the chance of one kind of mistake (missing a real effect) without increasing the chance of the other kind of mistake (a false alarm). This is one of many reasons why it's desirable to collect as much data as is feasible in research.

### Replication

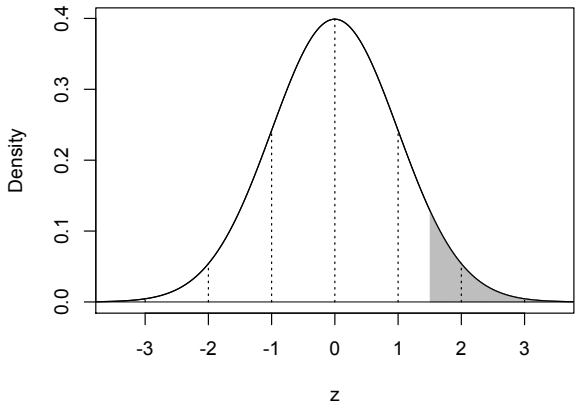
Another way to curb the number of mistaken decisions is through the practice of **replication**. This is a cornerstone of scientific method precisely because it can help to root out mistaken conclusions. In statistics, there's always some sampling error. Ultimately, the best way to see whether an effect exists or not is to test it many times and look for consistency in the results. There are many ways to check the consistency of results. You can repeat a study precisely, called **exact replication**, to see whether the findings hold up within a tolerable margin of error. Alternatively, you can change the methodology so that important constructs are operationalized differently, sometimes called **conceptual replication**. Even within a single study, you can incorporate multiple measures of key variables to check for consistency in the results.

The results of a single statistical test should never be considered strong evidence for or against a researcher's hypothesis. Far more persuasive is consistent evidence observed across many tests, especially if these involve different measures obtained in different samples in studies performed using different designs by members of different research labs. Each qualitative difference in methodology helps to rule out the possibility that an apparent effect is an artifact some element of the research. Consistency of results across diversity of methods provides the most compelling evidence.

## *p* Values

Rather than specifying an α level to identify the critical region, computerized data analysis is easier. SPSS, for example, simply provides what's called a ***p* value**. This is the probability of obtaining results as or more extreme than those observed in your sample of data. Using the data from this chapter, the *z* value of 1.50 corresponds to a *p* of .067. How is this calculated, and what does it mean?

It's calculated by using the unit normal table and finding what proportion of all *z* scores are beyond *z* = 1.50. Here's the relevant graph, with the upper tail shaded:




The area in one tail, found in the table for *z* = 1.50, is .06681. What this means is that if you select samples of *N* = 16 from a population of scores with μ = 5 and σ = 2, only about 6.7% of those samples would have *z* > 1.50.

How is this easier than identifying a critical region? You don't need to use the unit normal table at all. You just compare the *p* value directly to your α level. The decision rule is simple, and it will apply to all the inferential statistics we review in this book:

Statistical decision rule: If $p < \alpha$ you reject $H_0$, otherwise you retain $H_0$

A statistic that falls far out into the tail of a sampling distribution that represents $H_0$ yields a low *p* value. This means it's unlikely to have occurred just due to chance, so you can reject $H_0$. On the other hand, if a statistic falls closer to the middle of the sampling distribution expected under $H_0$, it yields a large *p* value. That means it might very well have occurred by chance, so you retain $H_0$.

One final note is that whereas we've done a 1-tailed test for these data, 2-tailed tests are the norm. For a 2-tailed test, the *p* value represents the area in both the upper and lower tails of the sampling distribution. Because this sampling distribution is symmetric, all that you have to do is double the area in one tail. In this case, for example, .06681 × 2 = .13362. The means the *p* value for a 2-tailed test would have been .134 rather than .067. Either way, we would retain $H_0$ because both of these *p* values are larger than α = .05.

## APA Style

For a one sample *z* test, you should include the sample *M* and *SD*, the population μ that represents $H_0$ plus the value of σ used in the statistical test, the *z* value, the *p* value, and

Cohen's *d*. This last value is a measure of effect size, and it will be described in the next chapter.

The key to reporting statistical results clearly and concisely is to begin by writing in plain English, and then adding statistical details as support. Whenever possible, try putting these details in parentheses or tacked onto the end of the sentence. Treat them like you would citations or footnotes. Tuck them away where an interested reader can find them, but don't let them intrude or distract from what you're really trying to say.

For any inferential statistic, you should indicate whether or not the test result is **statistically significant**. To be statistically significant means that your decision is to reject $H_0$. When you retain $H_0$, the result is not statistically significant.

Any *z* test can be reported in a single sentence. Here's an example for these data:

> The number of cards correctly identified by a sample of 16 subjects ($M$ = 5.75, $SD$ = 2.08) was not statistically significantly better than what would be expected for random guessing ($\mu = 5$, $\sigma = 2$), $z = 1.50$, $p = .067$, $d = 0.38$.

Notice that the phrasing of the results—"was not statistically significantly better"—indicates a 1-tailed test was used. Had this been a 2-tailed test, it could be written like this:

> The number of cards correctly identified by a sample of 16 subjects ($M$ = 5.75, $SD$ = 2.08) did not differ statistically significantly from what would be expected for random guessing ($\mu = 5$, $\sigma = 2$), $z = 1.50$, $p = .134$, $d = 0.38$.

This time, the phrasing of the results—"did not differ statistically significantly"—indicates a 2-tailed test was used.

## Problems

The following problems refer to the academics data set introduced earlier, for which $N = 40$. For the SAT variable, $M = 1210.50$ and $SD = 129.06$.

1. Suppose you want to test whether these SAT scores differ from the mean of all test-takers ($\mu = 1000$, $\sigma = 160$).
   a. State the statistical hypotheses ($H_0$ and $H_1$).
   b. Did you choose directional or a nondirectional hypotheses? Why?
   c. What $\alpha$ level would you use for this test? Why?
   d. What is the critical region for your test?
   e. What is the *z* value for these data? What does the *z* value mean?

f. What is the *p* value for this test? Use the unit normal table to find the proportion in the tail for the *z* value, and multiply this by 2 if you're doing a nondirectional (2-tailed) test. What does the *p* value mean?

g. Do these results this lead you to reject or retain $H_0$? Why?

h. What type of error—Type I or Type II—might you be making?

i. In a single sentence, report the results of this test in APA style.

2. Repeat parts (a) through (i), this time supposing that you want to test whether these SAT scores differ from the mean of all applicants to selective colleges nationwide ($\mu$ = 1100, $\sigma$ = 150). As you answer each part of the question, indicate whether your response is the same or different for #1 vs. #2, and explain why.

3. SPSS and other computer software that performs statistical analyses will sometimes provide a *p* value of .000. Keeping in mind what a *p* value represents, explain why it cannot equal 0. How should you report a *p* value that SPSS lists as .000 in APA style? (See the APA style guidelines summarized in an earlier chapter.)

4. A car manufacturer claims that a new hybrid model will get 50 MPG. Several magazines for consumers and car enthusiasts publish the results of their own independent tests, each performed under realistic driving conditions. Here are the MPG results they report:

   45, 48, 43, 52, 47, 47, 40

   Perform a statistical test of the manufacturer's claimed fuel economy, allowing a margin of error of $\pm$10% ($\sigma$ = 5) for acceptable variation under realistic driving conditions. Do parts (a) through (i), above.

**Problems 1 – 3 are due at the beginning of class.**

# 6. Effect Size

## Overview

There is an important difference between statistical significance and practical significance. Statistical significance deals with whether you can reject $H_0$. If so, you conclude that there is a systematic effect of some kind, that something other than sampling error is needed to explain the findings. However, a statistically significant effect is not necessarily practically significant, meaning that it may not be very important. A statistically significant but practically unimportant finding might be due to a flaw in the design of the study (e.g., failure to control for placebo effects, statistical regression, or other threats to internal validity by including a control group and assigning subjects to conditions randomly). This is why it's so important to consider threats to internal validity when interpreting results. Statistical significance alone doesn't necessarily lend support to a researcher's hypothesis.

Another way in which statistically significant results may not be practically significant is if they're too small to matter (e.g., a tiny difference across conditions that wouldn't be worth the time, effort, or expense of treatment). This is why you should always calculate a measure of **effect size** to help inform your judgment about whether the results are practically significant. This chapter shows how to calculate and interpret a common measure of effect size. Other measures will be introduced in later chapters.

## Measures of Effect Size

There are many different measures of effect size, most of which use a common scale to make it easy to apply rules of thumb and report effects as small, medium, or large. The choice of an appropriate measure depends on the research design and statistical analysis.

Whenever you want to compare the difference between two means, **Cohen's *d*** is a good choice of an effect size measure. The difference between the two means is the numerator, and to place this difference on a common scale you divide it by the standard deviation. This standardizes the mean difference in the same way that calculating a *z* score for an individual standardizes his or her score. It removes the unit of measurement in a particular study and places the score onto a standard scale with $\mu = 0$ and $\sigma = 1$.

To calculate *d*, you need to know which means to subtract in the numerator, and which standard deviation to place in the denominator. This depends on the type of design and analysis. We'll see several versions of *d* in the coming chapters. When you're doing a one sample *z* test, the formula for *d* looks like this:

$$d = (M - \mu) / \sigma$$

This is very much like the formula for a *z* score:

$$z = (X - \mu) / \sigma$$

The key difference is that a *z* score is calculated for an individual score ($X$), and *d* is calculated for a sample mean ($M$). Otherwise, they're interpreted the same way. A *z* score of

1.00, for example, indicates that a score falls one standard deviation above the mean. A *d* score of 1.00 indicates that a sample mean scores one standard deviation above the population mean.

Like a *z* score—but unlike a *z* test—Cohen's *d* does not take into account sample size. This is important because we want an estimate of the size of the effect to be independent of sample size. The goal is to estimate how large a difference exists between the means being compared. Collecting more data will improve the precision with which *d* estimates the true effect size in the population, but having a small sample will not bias *d* upward or downward. In short, the more data the better, but *N* is not part of the calculation of *d*.

To illustrate Cohen's *d*, recall the results for the precognition data from an earlier chapter:

> The number of cards correctly identified by a sample of 16 subjects ($M = 5.75$, $SD = 2.08$) was not statistically significantly better than what would be expected for random guessing ($\mu = 5$), $z = 1.50$, $p = .067$, $d = 0.38$.

Here's how the *d* value was calculated:

$$d = (M - \mu) / \sigma = (5.75 - 5) / 2 = 0.38$$

The *d* value tells us that the sample mean differed from the population mean by 0.38 standard deviations. Is this a large effect, a small one, or someplace in between?

## Interpreting Effect Size

Based on his extensive experience with the findings in many areas of psychological research, Cohen (1992)[15] suggested rules of thumb for what can be considered small, medium, and large effect sizes when they're measured using his *d* statistic:

0.20 = small
0.50 = medium
0.80 = large

Of course, the interpretation of any research result depends on the context. For example, even statistically small effects can be important if the outcome itself matters a great deal (e.g., risk of disease or death) or if the effect accumulates (e.g., small differences in athletic ability can add up to huge differences in performance over the course of a game, a season, or a career). Cohen intended his rules of thumb to be rough guidelines that provide for a common understanding, not hard-and-fast thresholds to apply rigidly.

The smallest value that *d* can take is 0.00—when there is no difference between the two means being compared—and anything less than 0.20 can be considered very small. Anything above 0.80 can be considered very large, and there is no largest value that *d* can take.[16] In practice, it's rare to find $d > 2.00$ or so in social and behavioral science.

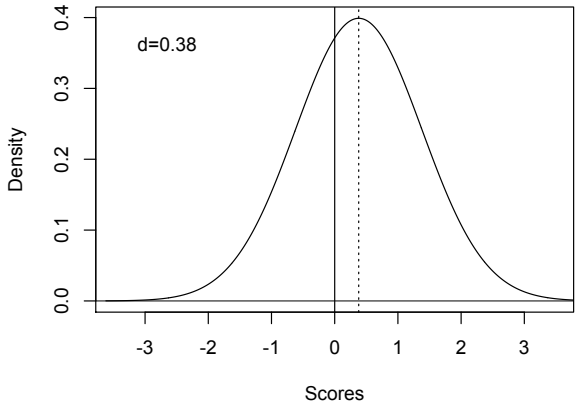
It's also important to note that *d* can be positive or negative. The sign indicates only the direction of the effect, which mean was larger. This is usually already obvious from looking at the data. The absolute value of *d* is what indicates the size of the effect.

---

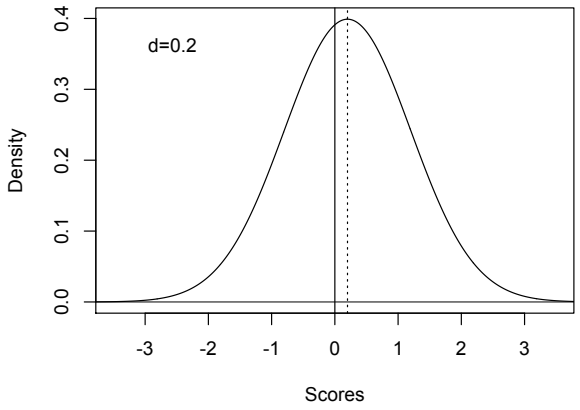
[15] Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

[16] In theory, the standard deviation could approach 0, in which case *d* could approach infinity.

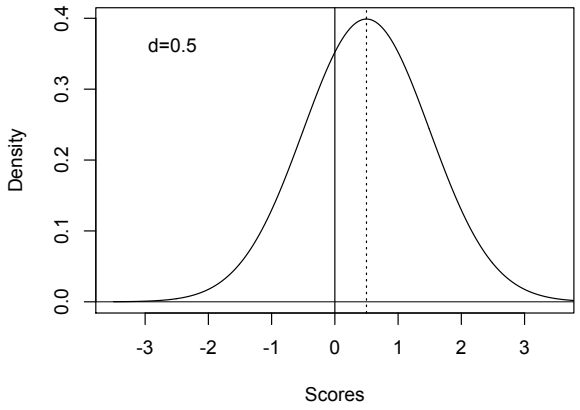
The value for the precognition data, $d = 0.38$, falls between a small and a medium effect. It can be helpful to show what this looks like. Imagine you have a normally distributed sample of scores, and their $M$ differs from $\mu$ by $d = 0.38$. Here's a density plot for that effect size, with $\mu = 0$ plotted as a solid line and $M = 0.38$ plotted as a dotted line:



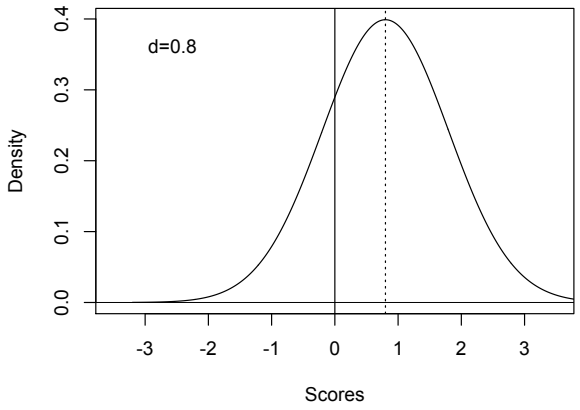

You can see the extent of the difference between the means, relative to the variability among all the scores. This is what Cohen would consider to be right in between a small and a medium effect. Here's a graph for a small effect:




Here's a graph for a medium effect:




Finally, here's a graph for a large effect:




You can see the effect size becoming larger as *d* increases. It's worth keeping in mind, though, just how much variation remains. Even for large effects, there are many individual exceptions to the general trend observed in a sample. When $d = 0.80$, for example, only 84.1% of the scores fall above μ, with the other 15.9% falling below μ. It would be foolish to presume that the trend—scores in the sample were generally above the population mean—applies to every individual.

Any new metric takes some time and experience to become familiar. For example, if you were to switch from measures of miles, pounds, and degrees Fahrenheit to the metric measures of kilometers, kilograms, and degrees Centigrade, everything would be pretty confusing for a while. At first, you'd convert metric values back into the old, familiar measures (e.g., each kilometer is about 0.62 miles, each kilogram about 2.20 pounds), but over time you'd become just as comfortable using the metric units themselves. The same applies to effect size measures. As you encounter more values of Cohen's *d*, you'll develop a sense for what it means to be a small, medium, or large effect. Because most of us are at least somewhat familiar with the magnitude of sex differences, they can provide some intuitive guideposts when expressed using *d*.

A review of many meta-analyses[17] (a research method described in the next section) found that for sex differences in cognitive abilities, communication, social and personality variables, and psychological well-being, about one-half of the studies reported small effects (i.e., $|d| \approx 0.20$) and another third of the studies reported virtually nonexistent effects (i.e., $|d| \approx 0$). In contrast, sex differences in adult height are very large: In the U.S., the mean difference of nearly 6" is about double the within-group *SD* of nearly 3" ($d \approx 2.00$).

## Meta-Analysis

The traditional way to synthesize the findings across a large number of studies is through a narrative literature review. Often, the reviewer tallies the number of studies that reported statistically significant results, as well as the number that did not, to see whether the evidence supports the existence of a certain effect. This can be very useful, but there are two major problems.

[17] Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581-592.

The first problem is that in individual studies, a real effect can be missed. Especially when investigators use small samples, the risk of such Type II errors can be large. This can bias the tally in a literature review, masking the existence of an effect.

The second problem is that results can appear inconsistent across studies due to ordinary sampling error. In other words, the apparent differences in results from study to study can mask an underlying similarity. Reviewers might be tempted to provide explanations for the apparent differences, stating them either as hypotheses to be tested further or conclusions supported by the review itself.

The availability of standardized measures of effect size, such as Cohen's *d*, helps to solve both of these problems through **meta-analysis**. A meta-analysis is a quantitative synthesis of research, which can supplement or replace a narrative review. The nuts and bolts of meta-analysis will not be covered here.[18] What follows is a conceptual overview of the method and its benefits.

To perform a meta-analysis, the first step is to calculate a common measure of effect size for each study. Cohen's *d* is a popular choice. Next, a weighted average of the effect sizes is calculated across all studies. The weights are based on sample size, which means that larger studies count more toward the overall average than do smaller studies. This weighted average is based on all available data, so it reflects the single best estimate of the size of the effect.

At this stage, the meta-analyst can perform a test of the statistical significance of this average effect size. If it differs from the null hypothesis value of 0, that supports the existence of an effect. Because the meta-analytic test is based on all the data in the research literature, it's highly sensitive and unlikely to lead to a Type II error. If the effect exists, it should be detected in a meta-analysis.

In addition, meta-analysis can be used to test for **moderators**, or variables that influence the size of the effect. Rather than subjectively judging whether there appear to be important differences in effects across studies, more objective statistical tests can be performed.

An illustration might help to clarify the benefits of performing a meta-analysis. For many decades, there was heated debate in the literature about IQ as a predictor of job performance. Dozens, perhaps hundreds, of studies had been conducted. Each sampled from a different type of job and measured performance in its own way. Various measures of IQ were used, too. Most of the studies were fairly small. It should come as no surprise that results appeared to vary substantially across studies. Different reviewers reached different conclusions, with many explanations proposed to explain the pattern of results.

Beginning in the 1980s, the method of meta-analysis was applied to these data. In short order, two conclusions emerged. First, across a wide range of jobs, IQ was one of the strongest predictors of performance. Much of the apparent inconsistency was due to normal sampling error, and Type II errors were common in individual studies. Second, there was one important trend in the data all along. The strength of the association between IQ and job performance was moderated by job complexity. IQ predicted performance for all kinds of jobs, but the association was stronger for more cognitively

---

[18] An excellent guide to meta-analysis is Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

demanding jobs. For example, the IQs of auto mechanics are a stronger predictor of their job performance than are the IQs of workers on assembly lines that manufacture auto parts. It takes considerably more knowledge and decision-making capacity for auto mechanics to diagnose and solve problems with complex mechanical and electrical systems than it does to perform the comparatively rote tasks on an assembly line. Whereas narrative reviews of this literature failed to achieve consensus, the meta-analytic conclusions are now widely accepted.[19]

## APA Style

The *Publication Manual* of the APA strongly recommends reporting effect sizes. Usually, this is done by appending an appropriate measure of effect size to the end of the results for each statistical test. As shown in an earlier chapter, using the precognition data, the value for Cohen's *d* can be listed after the results for the statistical test:

The number of cards correctly identified by a sample of 16 subjects ($M = 5.75$, $SD = 2.08$) was not statistically significantly better than what would be expected for random guessing ($\mu = 5$, $\sigma = 2$), $z = 1.50$, $p = .067$, $d = 0.38$.

## Problems

1. Explain the difference between statistical significance and practical significance. Provide your own example of a situation in which these might differ.
2. Using the SAT scores from the academics data introduced in an earlier chapter ($M = 1210.50$), what is the size of the effect for a test of whether these SAT scores differ from the mean of all test-takers ($\mu = 1000$, $\sigma = 160$)? According to Cohen's rules of thumb, how would you describe this effect size?
3. Using the same data, what is the size of the effect for a test of whether these SAT scores differ from the mean of all applicants to selective colleges nationwide ($\mu = 1100$, $\sigma = 150$)? According to Cohen's rules of thumb, how would you describe this effect size?
4. Using the data for the hybrid car's fuel economy (from a problem in an earlier chapter; $M = 46.00$), what is the size of the effect for a test of whether the manufacturer's claimed fuel economy is true within a tolerable margin of error ($\mu = 50$, $\sigma = 5$)? According to Cohen's rules of thumb, how would you describe this effect size?

* * *

[19] An outstanding review is provided by Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

The following problems refer to the results for eight studies evaluating treatments for depression. In each study, a sample of individuals seeking treatment was administered 12 sessions of therapy. Though depression was measured using several different scales, lower scores always represent better outcomes (fewer depressive signs or symptoms). The values of μ and σ represent the expected level of depression without treatment.

| **Study #** | ***N*** | ***M*** | **μ** | **σ** | ***z*** | ***p*** | ***H*₀** | ***d*** | ***N* × *d*** |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 22 | 27.3 | 30 | 8 | | | | | |
| 2 | 23 | 18.1 | 22 | 6 | | | | | |
| 3 | 17 | 26.5 | 30 | 8 | | | | | |
| 4 | 14 | 16.6 | 22 | 6 | | | | | |
| 5 | 17 | 15.8 | 22 | 6 | | | | | |
| 6 | 30 | 18.7 | 22 | 6 | | | | | |
| 7 | 15 | 47.5 | 50 | 10 | | | | | |
| 8 | 20 | 48.2 | 50 | 10 | | | | | |

5. Calculate *z* for each study and write it in the "*z*" column. Use the following formula:

   $z = (M - \mu) / (\sigma / \text{sqrt}(N))$

6. Determine the *p* value for a 2-tailed *z* test with $\alpha = .05$ for each study and write it in the "*p*" column. Use the unit normal table. Look up a *z* value to find the proportion in the tail (the value listed in the table), and multiply by 2 for a 2-tailed *p* value.

7. For each study, determine whether you would reject or retain $H_0$ by comparing the *p* value to your α level (.05). Write "Reject" or "Retain" in the "$H_0$" column.

8. Based on the results for all 8 studies, what general conclusions can you reach about these treatments for depression?

Problems 5 – 8 are what would normally be done in individual studies plus a traditional, narrative style of literature review. Problems 9 – 12 resemble what would be done in a meta-analysis; some of the steps are simplified, but the general idea is to synthesize research using effect sizes.

9. Calculate Cohen's *d* for each study and write it in the "*d*" column. Use the following formula:

   $d = (M - \mu) / \sigma$

10. Calculate a weighted average value of *d*. Use these three steps:

    a. Multiply $N \times d$ for each study.

    b. Sum the 8 values of $N \times d$.

    c. Divide this sum by the total *N* for all 8 studies.

11. How do your effect size calculations help you to refine the conclusions you reached in #8?

12. Suppose that studies #1, 3, 7, and 8 used interpersonal therapy, and studies #2, 4, 5, and 6 used cognitive-behavioral therapy. Repeat the steps in #10 to calculate a weighted average value of *d* for each kind of therapy. What conclusions can you draw from these findings?

* * *

13. Suppose that a meta-analyst finds that the added risk of a crash while using a cell phone is statistically significant, but the size of the effect is small. Would you consider this to be practically significant? Why or why not?

14. Research shows that men are more likely than women to negotiate for higher pay, both when they are first hired and periodically throughout their working careers. The initial difference in pay is approximately 3-5%, but that grows over time. Would you consider this to be practically significant? Why or why not?

**Problems 1 – 10 are due at the beginning of class.**

# 7. Statistical Power

## Overview

**Statistical power** is the probability of correctly rejecting a false $H_0$. Put more simply, it's the probability of detecting an effect that does exist. You can think of statistical power as analogous to the probability that a smoke detector will alert you when an actual fire occurs. Historically, researchers have focused on protecting against Type I errors, or false alarms. For example, low $\alpha$ levels have been used to reduce the chance of rejecting a true $H_0$, and thereby making a Type I error. Because precautions against Type I errors often involve a trade-off with Type II errors, investigators frequently perform studies with weak statistical power.

Cohen (1962) calculated estimates of statistical power for studies published in a leading psychology journal and found that the average power was only .50.[20] This means that even when a researcher's hypotheses was correct, that there was a systematic effect, the chance of detecting it statistically was equivalent to the toss of a coin. Researchers weren't giving their hypotheses a strong chance of being supported. This mediocre power was observed for studies published in an excellent journal. The power of studies in the much more numerous and less selective journals might have been even lower. After about 25 years in which methodologists tried to raise awareness of this serious problem in research design, two investigators repeated Cohen's power-estimation study.[21] They found that the average statistical power of articles published in journals comparable to those that Cohen had studied was still only about .50.

This chapter examines the factors that influence statistical power, the benefits of performing a power analysis when planning research, and ways to maximize power.

## Factors That Affect Statistical Power

Though we will not delve into the details of how statistical power is calculated for every type of research design and statistical analysis, it is helpful to understand that there are always three factors that affect power.

### Sample Size

The first factor that affects statistical power is sample size. Larger samples increase power because they reduce sampling error and provide more precise estimates of population parameters. Whereas genuine effects can easily be masked by sampling error in small samples, this is less likely to happen with increases in sample size. This is one of many reasons why it's a smart idea to collect as much data as possible.

---

[20] Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

[21] Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

### Decision Threshold

The second factor that affects statistical power is the decision threshold. This is determined by the choice of an $\alpha$ level, plus the choice of whether to perform a 1-tailed or 2-tailed test. A higher $\alpha$ level increases the size of the critical region, which makes it easier to reject $H_0$. Presuming one has predicted the correct direction of an effect, using a 1-tailed test achieves the same thing, enlarging the critical region where an effect is expected. This gain in statistical power, however, comes at the expense of a corresponding risk of making more Type I errors. In other words, by raising $\alpha$ or using a 1-tailed test, you increase power if your hypothesis is correct (and $H_0$ is false) but you increase the chance of a false alarm if your hypothesis is mistaken (and $H_0$ is true). In addition to keeping themselves honest in data analysis, the trade-off with Type I errors is another reason why researchers seldom use 1-tailed tests or raise the $\alpha$ level beyond the conventional value of .05.

### Effect Size

The third factor that affects statistical power is effect size. It is easier to detect a large effect than a small one, which means that power is influenced to some extent by what you choose to study. If you're studying sex differences in height, you'll have a pretty easy time rejecting $H_0$ and correctly concluding that men are taller than women because the effect size is quite large. In contrast, if you're studying sex differences in cognitive abilities, you'll have a much more difficult time detecting these very small effects. Though there are some ways to maximize the size of the effect measured in research, for the most part this is out of our hands. The reality is that some effects are small, others are large.

To illustrate the operation of all three factors, recall the precognition study introduced earlier. Subjects tried to identify the shape on each of 25 Zener cards. Given the sample size ($N$ = 16), decision threshold (1-tailed test with $\alpha$ = .05), and effect size ($d$ = 0.38), statistical power can be calculated to be .323.[22] That means that even if a precognition effect of this magnitude exists, a study of this size, with the data analyzed in this way, holds only a 32% chance of correctly rejecting $H_0$. That's not very good. A rough rule of thumb that has emerged in the literature is that a power of .80 is considered pretty good. Naturally, higher values are even better. Soon we'll see how to maximize power, but first let's review the benefits of estimating it in the first place.

## Benefits of Power Analysis

Because there are only three factors that affect statistical power, it's not very hard to estimate it. We won't deal with the specifics because they depend on the statistical test that's used, which in turn depends on the design of the study. The investigators cited earlier in this chapter calculated power for published studies to demonstrate that it's often weak in psychological science. Performing a power analysis at the planning stages is not quite as simple, but it can be extremely informative.

The challenge in estimating power while planning a study is that you can't know in advance how large an effect is. If you knew the true effect size, you probably wouldn't need

[22] This text won't show how to perform power calculations, which can be done most easily via any number of statistical test-specific calculators found online.

to do the study in the first place. However, you can make an educated guess about the likely effect size, or you can even consider a plausible range of values (e.g., worst- and best-case scenarios) to estimate a range of power levels. Prior research on related topics often provides clues about effect size.

The table in Appendix B, adapted from Cohen (1992), shows the sample size required to achieve a predetermined statistical power for many common research designs and measures of effect size. When you examine the sample size requirements in the table, you might be surprised to learn how large a study you would need to perform to have pretty good statistical power. This is especially true if you are studying a phenomenon that corresponds to a small effect. Consulting a table like this is the most rudimentary type of power analysis, but even this can be quite useful. There are several questions that you can address through power analysis at the planning stage of research.

### Does My Study Have a Good Chance of Yielding Informative Results?

If you're willing to make a rough estimate of the size of the effect you're hypothesizing, and you know how much data you'll be able to collect, you can estimate statistical power. The level of power corresponds to the likelihood that your study will yield informative results. If your hypothesis is correct, you might get lucky and be able to reject $H_0$. However, you might be unlucky and fail to reject $H_0$. That's not a very informative result because of the inherent ambiguity. There's no way to tell if your hypothesis is mistaken or you were the victim of an underpowered study, with sampling error masking the effect you expected to see. The higher your statistical power, the more likely it is that luck will be on your side. There's no sense taking a poor gamble when so much of your time and other resources will be on the line. There are always better things to do than underpowered research.

### How Large a Sample Will I Need?

Of the three factors that affect power, you usually have the most control over sample size. Performing a power analysis can help you make an informed decision about how much data to collect. Alternatively, if you find that you will be unable to collect as much data as you'd need to achieve an acceptable level of statistical power, you might not want to do this study after all. If a study is doomed by low power, it's better to know in advance rather than wasting the time and other resources required to go through with it.

### How Small an Effect Will I Be Able to Detect?

If there are limits on how much data you can collect, you can use power analysis to determine how small an effect will be detectable with an acceptable level of statistical power. This, too, can be helpful in deciding whether the study is worth doing.

### Is It Reasonable to Fund This Research?

Increasingly, granting agencies require investigators to perform power analyses to demonstrate that if their hypotheses are correct, their planned research will find statistical support for them. Considering how many well-qualified applicants are competing for the limited funds available for research grants, it makes very little sense to fund research with low power.

## How to Maximize Statistical Power

To review and extend the discussion of statistical power, here are ways that it can be maximized. Some of these should be fairly obvious, but others are a bit more subtle.

### Collect as Much Data as Possible

The more data, the less sampling error, and the stronger the statistical power. The usual way to collect more data is to increase sample size, but that's not the only possibility. Sometimes you can increase the number of trials in an experiment. In the precognition study, for example, you could do both: Recruit more than 16 subjects, and have each predict the shape on more than 25 Zener cards. Why not run through the deck multiple times? That provides a more rigorous test, and it would increase power by providing a more sensitive measure of performance for each subject. Whether by recruiting more subjects or testing each one more often, it becomes easier to identify a systematic effect when the background noise of chance is minimized by gathering more data.

### Use the Right Decision Threshold

Whenever you can justify it, you'll maximize power by creating as large a critical region as you can. This can be accomplished not only by using as large an $\alpha$ level as possible, but also by performing a 1-tailed test. Both of these ways to increase power come at the cost of increasing the chance of making Type I errors, though. In the precognition study, a 1-tailed test is justifiable, which increases the power of that analysis relative to the usual 2-tailed test. There is usually no good reason to increase $\alpha$ above .05, though, so that's about the best that can be done with the precognition data to maximize power.

### Choose Your Subject Matter Wisely

Investigators are free to choose what they want to study. Whether by good fortune or guided by educated guesses, those who target effects that are larger will have greater statistical power.

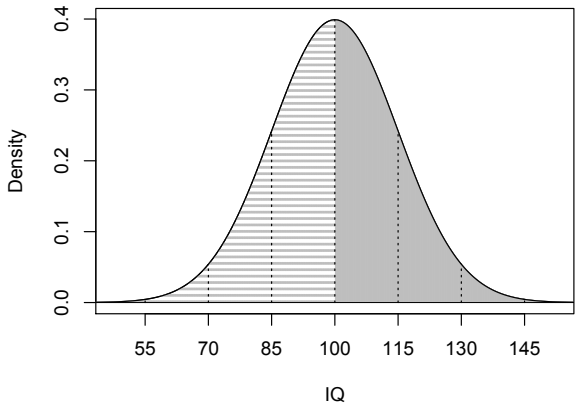
### Use the Most Reliable Measures

All variables are subject to some measurement error, however slight, but some measures are much more reliable than others. Measurement error acts in much the same way as sampling error, adding random noise to the data that serves to mask genuine effects. Choosing the most reliable measures will reduce measurement error and maximize statistical power. For example, the more Zener cards whose shape a subject is asked to identify, the more reliable a measure of precognitive ability is provided. This is how increasing the number of trials improves power, by increasing the reliability of measurement.

### Do Not Split Cases Into Groups

An unfortunately common practice in research is to divide scores on a quantitative measure into two groups, often according to whether they fall above or below a certain value (e.g., the mean or median). This simplifies the data, creating two groups that can be compared rather than allowing subjects to vary along the full continuum of scores. It can make it easier to analyze the data or to communicate the results. However, it's a bad idea because it throws away a great deal of information and can greatly reduce statistical power.

To understand the problems with splitting cases into groups, suppose that you have a normal distribution of IQ scores with $M = 100$ and $SD = 15$. If you split this at the mean (or the median) of 100, that creates two groups, low-IQ and high-IQ. In the density plot shown below, the groups are shaded differently:




The first problem with this split is that scores that are very far apart are treated as identical. For example, individuals with IQ scores of 55 and 99 would both be lumped together in the low-IQ group, and individuals with IQ scores of 101 and 145 would both be lumped together in the high-IQ group. A great deal of potentially meaningful variation within each group is lost. This is analogous to using a less reliable measure.

The second problem with this split is that many scores that are very close together are treated as different. For example, individuals with IQ scores of 99 and 101 would be classified into different groups despite being remarkably similar to one another. This will happen quite often because there are so many scores clustered around the dividing line of IQ = 100.

The combination of treating very different scores as identical and treating very similar scores as different has pernicious effects on statistical power. To illustrate the extent that this can weaken the analysis, Cohen (1983) examined what happens if you have two normally distributed variables and you want to test their correlation. Splitting one of them at the mean can reduce power by one-third or more, and splitting both of them at the mean can reduce power by about two-thirds.[23] This is the equivalent of throwing away large amounts of data, acting as though you hadn't gone to the trouble of testing many of your subjects. Weakening statistical power in this way is a remarkably foolish thing to do.

## Problems

1. What are the three factors that influence statistical power? In your own words, how is each one of these factors related to power?
2. What are the benefits of estimating statistical power when planning research? Provide your own example of how this might be helpful in a specific instance.
3. Using the academics data introduced in a problem in Chapter 5, which of the following two tests would have higher statistical power, and why?

[23] Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

a. A test of whether the SAT scores in the sample differ from the mean of all test-takers ($\mu = 1000$, $\sigma = 160$).

b. A test of whether the SAT scores in the sample differ from the mean of all applicants to selective colleges nationwide ($\mu = 1100$, $\sigma = 150$)?

4. Recall the data for the hybrid car's fuel economy, introduced in a problem in Chapter 5. What could be done to increase statistical power? (There is one answer that should be very easy to discover, and at least one more that's less obvious—see if you can figure out both.)

5. Zeke develops a mnemonic device that can help students learn vocabulary words when studying a foreign language. Suppose the effect on performance is, in reality, medium in size ($d = 0.50$). If Zeke does a study that compares two equal-sized groups—an experimental group trained to use the new technique and a control group—and analyzes the data using a 2-tailed $t$ test with $\alpha = .05$, how many people need to participate to have an 80% chance of statistically detecting the effect? In other words, what $N$ does Zeke need to attain what's usually considered to be acceptable statistical power? Take your best guess, do not use any tools to ensure that you answer this correctly.

6. Repeat #5, this time assuming that the effect size is large ($d = 0.80$). How many people need to participate for statistical power to reach .80?

7. Repeat #5 one last time, this time assuming that the effect size is small ($d = 0.20$). How many people need to participate for statistical power to reach .80?

8. Why is there a 0 before the decimal place for Cohen's $d$ (e.g., $d = 0.80$) but not for statistical power (e.g., power = .80)?

9. Use Appendix B, specifically the upper section that lists Cohen's $d$ as a measure of effect size for comparing two groups, to find the correct answer to #5. How does this compare with your best guess?

10. Repeat #9 for a large effect size ($d = 0.80$), this time comparing your answer to #6.

11. Repeat #9 for a small effect size ($d = 0.20$), this time comparing your answer to #7.

12. Zeke classifies each of 35 students as either high GPA (above 3.00) or low GPA (at or below 3.00) as well as either high SAT (above 1000) or low SAT (at or below 1000). He performs a statistical test following the usual conventions and finds that there is no statistically significant relationship between these variables. State any three distinct ways that Zeke could improve the statistical power with which he addresses this research question.

**Problems 1 – 11 are due at the beginning of class.**

# 8. One Sample *t* Test

## Overview

The one sample *z* test may be the simplest inferential statistic, but it's seldom used. The reason is that it requires knowledge of σ, the population standard deviation, a parameter that is usually unknown. Fortunately, we can use *SD*, the sample standard deviation, as an estimate of σ to perform a one sample *t* test. This chapter explores the similarities and differences between the *z* and *t* tests.

## Using *SD* to Estimate σ

Recall that the formulas used to perform a one sample *z* test are:

$$\sigma_M = \sigma / \text{sqrt}(N)$$

$$z = (M - \mu) / \sigma_M$$

If we replace σ with *SD*, we get the following formulas to perform a one sample *t* test:

$$SD_M = SD / \text{sqrt}(N)$$

$$t = (M - \mu) / SD_M$$

Using *SD* as an estimate of σ opens up a wide range of research possibilities. Whereas σ is seldom known, we can always calculate *SD* for a set of data. For example, a problem in an earlier chapter involved testing whether a new hybrid car actually got 50 MPG, as claimed by its manufacturer. It was easy to establish the appropriate statistical hypotheses:

$H_0$: $\mu = 50$
$H_1$: $\mu \neq 50$

Performing the *z* test, however, required knowledge of σ. For the sake of the problem, we defined a tolerable margin of error as ±10%, meaning that σ = 5 (which is 10% of μ = 50). This was a very arbitrary decision. A better approach would be to ask whether the sample mean differs from μ = 50 by more than we would expect by sampling error, or chance. The *t* test allows us to do this. Rather than setting an arbitrary value for σ, we use the value of *SD* calculated from the data as an estimate of σ.

## Performing the *t* Test

In the case of the new hybrid car, several magazines reported the fuel economy it achieved in their test drives: 45, 48, 43, 52, 47, 47, and 40 MPG. Once you've calculated the *M* and *SD* of these $N = 7$ scores ($M = 46.00000$, $SD = 3.82971$), the value of *t* is easy to calculate in two steps:

$$SD_M = SD / \text{sqrt}(N) = 3.82971 / \text{sqrt}(7) = 1.44749$$

The standard error ($SD_M$) tells us that we would expect the $M$ for a sample of 7 scores to vary by about ±1.45 points from μ by chance.

$$t = (M - \mu) / SD_M = (46.00000 - 50) / 1.44749 = -2.76$$

The $t$ value is interpreted just like a $z$ value: It's the ratio of how much the means actually differed to how much we expect them to differ by chance. In this case, it tells us that the $M$ for our data differed from μ by 2.76 times as much as we'd expect.

Before we determine whether this is a statistically significant difference, notice that the values of $M$, $SD$, and $SD_M$ were not rounded to two decimals (i.e., what we typically report in APA style) when calculating $t$. It's always a good idea to retain extra decimal places at intermediate stages of a calculation. If you round off too early, that rounding error can be compounded in later calculations and give a final answer that's incorrect. When you use a computer to perform data analysis, it retains a very large number of decimal places for all calculations. When doing data analysis by hand, a good rule of thumb is to retain two or three extra decimal places until you have your final answer. For example, as shown above you can keep five decimal places if you'll be rounding to two decimals at the end.

## Critical Region

With the $z$ test, we consulted the unit normal table to establish the critical region. With the $t$ test, we can no longer use the same table. Whereas there is a single $z$ distribution, there is actually a family of $t$ distributions that differ according to the sample size of the data used to calculate the $t$ value.

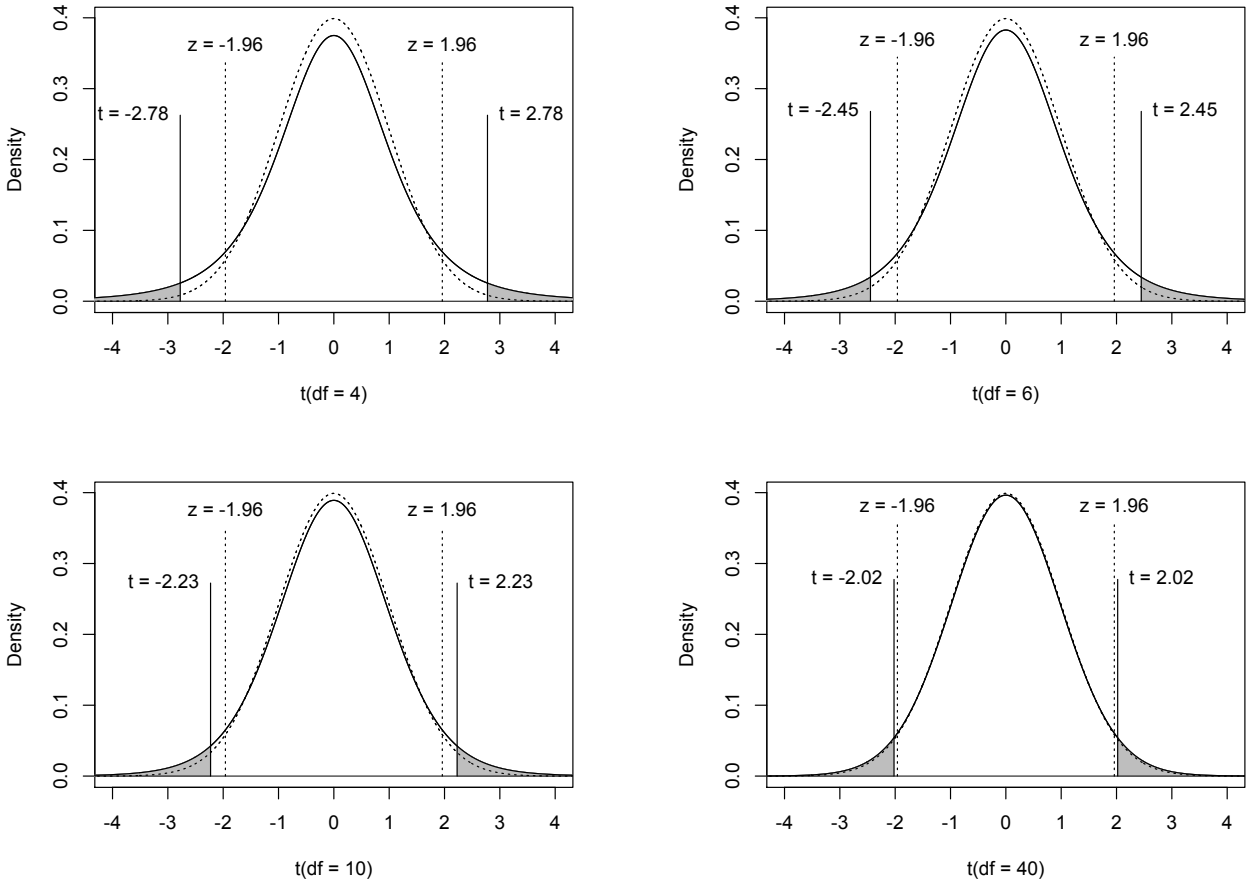
Why does sample size matter for $t$, but not for $z$? Because we're now using a statistic ($SD$) calculated from data to estimate a population parameter (σ). $SD$ is an unbiased estimate of σ, but the fact that it's an estimate introduces an additional source of sampling error. In a very small sample, for example, the $SD$ might be quite far from σ. In a very large sample, $SD$ will usually be quite close to σ. Thus, the amount of added sampling error depends on the size of the data set used to calculate $t$.

The family of $t$ distributions differs by how far it diverges from the $z$ distribution. Hypothetically, with an infinitely large sample size, $t$ is identical to $z$. That's because an infinitely large sample size eliminates all sampling error in estimating σ from $SD$. As sample size decreases, sampling error increases, so the $t$ distribution diverges further from the $z$ distribution. We index the size of the sample using **degrees of freedom**, abbreviated as $df$. Every statistical test that will be introduced from this point forward has its own expression for $df$, and it always depends on how many sample statistics are being used to estimate population parameters. For a one sample $t$ test, $df = N - 1$. We're using one statistic ($SD$) to estimate a parameter (σ), so we "give up" one degree of freedom.

The $t$ table (see Appendix A) allows us to determine the critical region for a $t$ test. We need to know three things to use the table:

1. Is this a 2-tailed (nondirectional) or a 1-tailed (directional) test?
2. What is the α level? The $t$ table in the appendix allows choices of α = .05, .01, or .001.
3. What is the $df$? The $t$ table in the appendix lists $df$ from 1 to 30, then a few additional values (40, 60, 120), and then the hypothetical limit at infinite $df$.

The four graphs below show *t* distributions with *df* = 4, 6, 10, and 40. Each graph plots the critical region for a 2-tailed *t* test with $\alpha = .05$. As a point of reference, the *z* distribution, and its critical value, is plotted using dotted lines. In each case, the *t* distribution is a little shorter than the *z* distribution in the middle and thicker in the tails:




You can see that a much larger value of *t* than of *z* is required to reject $H_0$ when the sample is small, but with even modest sample size the critical region for *t* becomes very similar to what it would be for *z*. In the hypothetical limiting case of infinite *df*, the *z* and *t* distributions are identical.[24]

In the case of the new hybrid car, we'd use the *t* table to find the critical region for a 2-tailed test with $\alpha = .05$ and $df = N - 1 = 7 - 1 = 6$. This provides a value of 2.447. Because we're doing a 2-tailed test, this refers to both tails of the sampling distribution. Our critical region includes the left tail, $t < -2.447$, as well as the right tail, $t > 2.447$. A simpler way to express this is $|t| > 2.447$.

The *t* value calculated from the data, -2.76, falls within the critical region, so we would reject $H_0$ and tentatively accept $H_1$. The observed $M = 46.00$ MPG is statistically significantly different from the value of $\mu = 50$ MPG claimed by the manufacturer.

[24] The bottom row of the *t* table provides a critical region identical to a *z* distribution. Thus, using the bottom row of the *t* table is a useful shortcut when doing a *z* test.

## Effect Size

We can use Cohen's *d* as the measure of effect size, just as we did for the one sample *z* test. The only difference is that we calculate this using *SD*, rather than σ, as the standard deviation in the denominator:

$$d = (M - \mu) / SD$$

The rules of thumb for interpreting *d* (0.20 = small, 0.50 = medium, and 0.80 = large) remain the same. For these data, the effect is large:

$$d = (46.00000 - 50) / 3.82971 = -1.04$$

Once again, notice that extra decimal places were used for the *M* and *SD* so that the final value of *d* would be correct when rounded to two decimals.

## Using SPSS

To perform a one sample *t* test in SPSS, you first enter your data into a single variable (column), here labeled "MPG":

| | MPG |
|---|---|
| 1 | 45.00 |
| 2 | 48.00 |
| 3 | 43.00 |
| 4 | 52.00 |
| 5 | 47.00 |
| 6 | 47.00 |
| 7 | 40.00 |

Next, you use the following command:

```
t-test vars = mpg
/testval = 50
```

Note that you have to provide the variable to be tested (here, "MPG") as well as the value of μ (which SPSS calls the "testval"; here, the test value of μ = 50). The output appears in two tables. The first table, labeled "One-Sample Statistics", provides the *M* and *SD*:

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| MPG | 7 | 46.0000 | 3.82971 | 1.44749 |

The second table, labeled "One-Sample Test", provides the *t* value, *df*, and the *p* value:

**One-Sample Test**

| | Test Value = 50 | | | | | |
|---|---|---|---|---|---|---|
| | t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference | |
| | | | | | Lower | Upper |
| MPG | -2.763 | 6 | .033 | -4.00000 | -7.5419 | -.4581 |

Note that the *p* value is labeled "Sig. (2-tailed)". In SPSS, "Sig." stands for "significance", as in statistical significance. SPSS assumes you'd like a 2-tailed test. If you want the *p* value

for a 1-tailed test, just divide the *p* value shown in the output by 2 (e.g., for a 1-tailed test of these data the *p* value would be .016).

## APA Style

You can report the results of any *t* test in a single sentence. Here's what that might look like for these data:

The fuel economy attained in test drives of a new hybrid car ($M$ = 46.00, $SD$ = 3.83) differed statistically significantly from the manufacturer's claim ($\mu$ = 50), $t(6) = -2.76$, $p$ = .033, $d$ = -1.04.

Notice that the *df* for the *t* test are provided in parentheses after *t*. Also notice that the sentence was phrased in a way that indicates this was a 2-tailed test. Had this been a 1-tailed test, the sentence (stripped of statistical results) might have been phrased like this: "The fuel economy attained in test drives of a new hybrid car was statistically significantly less than the manufacturer's claim." If your sentence states a direction, that implies a directional $H_0$ and a 1-tailed test. Normally, we use a nondirectional $H_0$ and a 2-tailed test, so the statistical results should be reported without reference to a direction. Naturally, elsewhere in the research report your interpretation of the results should take into account the direction of the effect.

## Problems

An investigator wants to test whether Alcoholics Anonymous (AA) is an effective treatment for alcoholism. Thirty people who choose to attend AA meetings are assigned sponsors (recovering alcoholics who attend AA meetings regularly) at their 1st meeting. The sponsor asks how much the subject drinks in a typical week, and records the total number of drinks (beer, wine, and liquor). The average number of drinks is treated as the population mean ($\mu$ = 20.00) against which the success of AA will be judged. When two months have passed since the 1st meeting, each subject reports to his or her AA sponsor how many drinks were consumed in the past week. $N$ = 16 of the original subjects are still attending AA meetings at this point, and the number of drinks they report that they consumed in the past week are:

13, 9, 14, 31, 22, 11, 9, 8, 29, 8, 18, 10, 9, 12, 9, 20

1. What is the researcher's hypothesis?
2. What are the statistical hypotheses ($H_0$ and $H_1$)?
3. Did you choose to use a 2-tailed or a 1-tailed test, and why?
4. Should this test be performed using $\alpha$ = .05 or $\alpha$ = .01, and why?
5. Why should a *t* test be performed rather than a *z* test?

6. What is *df* for this *t* test?
7. What is the critical region for this *t* test? (Use the *t* table in Appendix A to identify the critical region.)
8. Calculate the value of *t* for this sample of data. To help, note that $M = 14.50000$ and $SD = 7.42967$.
9. What is your statistical decision: Would you reject or retain $H_0$?
10. What kind of error—Type I or Type II—might you be making?
11. What is the size of the effect, using Cohen's *d*? According to the usual rules of thumb, how would you describe this?
12. Report the results of this test in APA style.
13. How would you interpret these results? Consider strengths and weaknesses of the research methods used in this study. There are at least three different interpretations that are consistent with the results.
14. How could you improve the design of this study to improve its internal validity?

* * *

The chair of a small department administers an optimism test to seniors graduating with this major to see whether there are changes in optimism from year to year. Higher scores on the test indicate greater levels of optimism. Last year's class had a mean score of $\mu = 15$. This year's graduating seniors scored as follows:

7, 12, 11, 15, 7, 8, 15, 9, 6

15. What is the researcher's hypothesis?
16. What are the statistical hypotheses ($H_0$ and $H_1$)?
17. Did you choose to use a 2-tailed or a 1-tailed test, and why?
18. Should this test be performed using $\alpha = .05$ or $\alpha = .01$, and why?
19. Why should a *t* test be performed rather than a *z* test?
20. What is *df* for this *t* test?
21. What is the critical region for this *t* test? (Use the *t* table in Appendix A to identify the critical region.)
22. Calculate the value of *t* for this sample of data. To help, note that $M = 10.00000$ and $SD = 3.42783$.
23. What is your statistical decision: Would you reject or retain $H_0$?
24. What kind of error—Type I or Type II—might you be making?
25. What is the size of the effect, using Cohen's *d*? According to the usual rules of thumb, how would you describe this?
26. Report the results of this test in APA style.

* * *

27. Use SPSS to analyze the fuel economy data from this chapter and verify that you get the same results and reach the same conclusions. Notice that you have to calculate Cohen's *d* by hand because SPSS does not do this for you. Check that you get the correct value for *d* by calculating that yourself.
28. Use SPSS to analyze the Alcoholics Anonymous data from the first series of problems and verify that you get the same results and reach the same conclusions as when you did the calculations by hand.
29. Use SPSS to analyze the optimism data from the second series of problems and verify that you get the same results and reach the same conclusions as when you did the calculations by hand.

**Problems 1 – 14 are due at the beginning of class.**

# 9. Related Samples *t* Test

## Overview

The one sample *t* test is more versatile than the one sample *z* test because we can use *SD* as an estimate of σ. However, we're still limited to analyzing data for a single sample. Researchers often want to compare results across conditions. This chapter shows how the *t* test can be extended to within-subjects designs to test for differences across two conditions. This is called the **related samples *t* test**, and it's also known as a **paired samples *t* test** or even simply a **paired *t* test**. In the next chapter, we'll examine a final variant of the *t* test that can be used for between-subjects designs.

## Using Difference Scores

Many research designs involve a comparison across two conditions. For example, a **pretest-posttest design** assesses the same subjects on two occasions (e.g., before and after treatment), as would a longitudinal study with measurements taken at two time points (e.g., attitudes measured at ages 20 and 40). Alternatively, the conditions can be two variables measured at the same point in time (e.g., scores on SAT Math and Verbal sections). In a **matching design**, pairs of subjects are matched on one or more variables and then assigned (ideally at random) to separate conditions (e.g., treatment and control). For purposes of data analysis, subjects matched in a pair can be treated as the same person.[25]

In each of these instances, a statistical comparison can be performed by calculating **difference scores**. For each subject, a difference score is calculated by subtracting the score for one condition from the score for the other condition (e.g., posttest – pretest, attitude at age 40 – attitude at age 20, SAT Verbal – SAT Math, treatment – control):

$$D = Y_1 - Y_2$$

Here, $Y_1$ and $Y_2$ refer to an individual's scores in the two conditions being compared. It makes no difference which condition is subtracted from which, but it is critical that the subtraction is performed in the same way for every subject. For example, you can use $Y_1$ = SAT Verbal and $Y_2$ = SAT Math or you can use $Y_1$ = SAT Math and $Y_2$ = SAT Verbal, but you have to pick one coding or the other to use for all subjects.

Suppose you wondered whether psychology majors tend to score higher on the Math or Verbal section of the SAT. Here are illustrative scores for 10 psychology majors, with difference scores calculated as SAT Math – Verbal:

[25] As you might imagine, this can be controversial. Only if subjects are matched very closely on all relevant variables is the calculation of difference scores and use of the related samples *t* test justified (and it's often called a **matched samples *t* test**. To the extent that the matching falls short of this ideal, it can be argued that calculating difference scores is inappropriate and the independent groups *t* test should be used instead.

| Math | Verbal | *D* |
|---|---|---|
| 540 | 590 | -50 |
| 630 | 620 | 10 |
| 590 | 650 | -60 |
| 530 | 610 | -80 |
| 490 | 580 | -90 |
| 660 | 720 | -60 |
| 590 | 660 | -70 |
| 490 | 560 | -70 |
| 650 | 610 | 40 |
| 680 | 750 | -70 |

Notice that some students tended to score relatively high on both sections (e.g., the student in the last row: 680 + 750 = 1430), others relatively low on both sections (e.g., the student in the first row: 540 + 590 = 1130). Individual differences like these are irrelevant to the question of whether, on average, students score higher on the Math or Verbal section. Difference scores essentially treat each subject as his or her own control, removing individual differences in overall cognitive ability to reveal whether or not there are systematic differences across conditions in the study (math and verbal abilities). Despite their individual differences in overall performance, the difference scores for the students in the first and last rows are highly similar (-50 and -70).

By removing individual differences, the difference scores reveal the consistency with which psychology majors scored higher on the Verbal than the Math section of the SAT. Within-subjects designs usually provide greater statistical power than between-subjects designs. We'll see examples of how big a boost this can be in the next chapter.

## Performing the *t* Test

Once you've calculated difference scores, all you need to do is subject them to a one sample *t* test. If there is no difference across conditions, the mean population difference score ($\mu_D$) will be 0. Thus, the statistical hypotheses for a related samples *t* test are:

$H_0$: $\mu_D = 0$
$H_1$: $\mu_D \neq 0$

These hypotheses are nondirectional, which is the norm. You could make them directional if there is sufficient justification.

The only difference in performing the *t* test involves notation. All terms are subscripted with a capital D to indicate that they refer to difference scores.

$$SD_{MD} = SD_D \,/\, \text{sqrt}(N)$$

$$t = (M_D - \mu_D) \,/\, SD_{MD}$$

Because we're testing $\mu_D = 0$, the *t* formula simplifies:

$$t = M_D \,/\, SD_{MD}$$

For the SAT data listed above, the calculations would look like this:

$$SD_{MD} = SD_D \,/\, \text{sqrt}(N) = 41.63332 \,/\, \text{sqrt}(10) = 13.16561$$

The standard error ($SD_{MD}$) tells us that we would expect the mean difference score ($M_D$) for a sample of 10 subjects to vary by about ±13.17 points from $\mu_D$ by chance.

$t = M_D / SD_{MD} = -50.00000 / 13.16561 = -3.80$

The $t$ value tells us that the $M_D$ for our data differed from $\mu_D$ by 3.80 times as much as we'd expect.

All that's left is to determine whether this falls in the critical region. Using the $t$ table (see Appendix A), we find that for a 2-tailed test with $\alpha = .05$ and $df = N - 1 = 10 - 1 = 9$, the critical $t$ value is 2.262. Because this is a 2-tailed test, the critical region includes the left tail ($t < -2.262$) and the right tail ($t > 2.262$). This is expressed most simply as $|t| > 2.262$. The absolute value of $t$, 3.80, exceeds 2.262, so we'd reject $H_0$ and tentatively accept $H_1$. There is a statistically significant difference between these students' SAT Verbal and Math scores.

## Effect Size

Recall that Cohen's $d$ is calculated as the difference between two means divided by the standard deviation. For one sample $z$ or $t$ tests, the relevant means were $M$ and $\mu$, and the relevant standard deviation was either $\sigma$ or $SD$. For related samples $t$ tests, however, the relevant means are those for the two conditions being compared, symbolized as $M_1$ and $M_2$.

The relevant standard deviation must be calculated as a type of average, though, because each condition has its own standard deviation, symbolized as $SD_1$ and $SD_2$. The way that these are averaged is to square each one to convert it back to a variance, add them up, divide by 2, and then take the square root to convert back to a standard deviation. This is called a **pooled standard deviation**, symbolized as $SD_p$ and calculated as follows:

$SD_p = \text{sqrt}((SD_1^2 + SD_2^2) / 2)$

Whenever you calculate $SD_p$ it's worth checking to make sure that it's in between $SD_1$ and $SD_2$. Because this is a type of average, it has to fall between the two values being averaged. It's easy to make calculation mistakes with this formula, and this quick check will catch most of them.

Cohen's $d$ is then calculated as the mean difference divided by the standard deviation in the usual way:

$d = (M_1 - M_2) / SD_p$

For the SAT data, the calculations look like this:

$SD_p = \text{sqrt}((69.96031^2 + 60.96447^2) / 2) = 65.61673$

$d = (585.00000 - 635.00000) / 65.61673 = -0.76$

Note that it makes no difference which condition you treat as 1 and which as 2. The only difference would be the sign of $d$, and that's unimportant for interpreting the size of the effect. In this case, we treated Math as condition 1 and Verbal as condition 2. Had we reversed these assignments, we'd have calculated $d = 0.76$ rather than $d = -0.76$. Either way, we can see from the relevant means that these psychology majors scored higher on the Verbal than the Math section of the SAT, and this is a large effect ($d$ is close to 0.80).

## Using SPSS

To perform a related samples *t* test in SPSS, you first enter your data into separate variables (columns), here labeled "Math" and "Verbal":

| | Math | Verbal |
|---|---|---|
| 1 | 540.00 | 590.00 |
| 2 | 630.00 | 620.00 |
| 3 | 590.00 | 650.00 |
| 4 | 530.00 | 610.00 |
| 5 | 490.00 | 580.00 |
| 6 | 660.00 | 720.00 |
| 7 | 590.00 | 660.00 |
| 8 | 490.00 | 560.00 |
| 9 | 650.00 | 610.00 |
| 10 | 680.00 | 750.00 |

Next, you use the following command:

t-test pairs = math verbal

Note that you have to provide the variables representing the two conditions to be compared (here, "Math" and "Verbal"). The output you need appears in two tables, though there is another table that you should ignore. The first table, labeled "Paired Samples Statistics", provides the *M* and *SD* for each condition:

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | SAT Math | 585.0000 | 10 | 69.96031 | 22.12339 |
| | SAT Verbal | 635.0000 | 10 | 60.96447 | 19.27866 |

The other table you need, labeled "Paired Samples Test", provides the *t* value, *df*, and the *p* value (labeled as "Sig. (2-tailed)", which you'd divide by 2 if you want a 1-tailed test):

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | SAT Math - SAT Verbal | -50.00000 | 41.63332 | 13.16561 | -79.78268 | -20.21732 | -3.798 | 9 | .004 |

There is another table that will appear between these, labeled "Paired Samples Correlations". This is not shown here. The correlation between scores in the two conditions is an entirely separate statistical analysis, and you should ignore this table. In particular, be careful not to report the *p* value in this table because it does not apply to the *t* test.

## APA Style

You can report the results of any *t* test in a single sentence. Here's what that might look like for these data:

Psychology majors scored statistically significantly differently on the Math (*M* = 585.00, *SD* = 69.96) and Verbal (*M* = 635.00, *SD* = 60.96) sections of the SAT, *t*(9) = -3.80, *p* = .004, *d*

= -0.76.

Notice that the phrasing indicates this was a 2-tailed test. Providing the means for both conditions clearly indicates that scores were higher on the Verbal than the Math section.

## Problems

The following problems refer to a study of two groups of men who attend Alcoholics Anonymous (AA) meetings. Members of group 1 (volunteers) choose to join AA and attend at least three meetings, and members of group 2 (court-ordered) are convicted of DUI and ordered by a court to attend at least three meetings. Prior to attending AA, subjects are matched on their age, race, and income level to form 12 pairs. After the third meeting, each subject is asked by his AA sponsor how many drinks he consumed each day in the past week. The data are shown below, with the number of drinks for both members of each pair in the same row.

| Volunteer | Court-Order |
|---|---|
| 16 | 23 |
| 8 | 7 |
| 9 | 16 |
| 5 | 14 |
| 10 | 8 |
| 14 | 17 |
| 13 | 22 |
| 3 | 0 |
| 5 | 9 |
| 19 | 16 |
| 8 | 13 |
| 9 | 17 |

1. What is the researcher's hypothesis?
2. What are the statistical hypotheses ($H_0$ and $H_1$)?
3. Did you choose to use a 2-tailed or a 1-tailed test, and why?
4. Should this test be performed using $\alpha = .05$ or $\alpha = .01$, and why?
5. Why should a related samples *t* test be performed, rather than a one-sample *z* or *t* test?
6. What is *df* for this *t* test?
7. What is the critical region for this *t* test? (Use the *t* table in Appendix A.)
8. Calculate the value of *t* for this sample of data. To help, note that $M_D = -3.58333$ and $SD_D = 4.69929$.
9. What is your statistical decision: Would you reject or retain $H_0$?
10. What kind of error—Type I or Type II—might you be making?
11. What is the size of the effect, using Cohen's *d*? To help, note that $M_1 = 9.91667$, $SD_1 = 4.77605$, $M_2 = 13.50000$, and $SD_2 = 6.57129$ (where subscripts of 1 = volunteer and subscripts of 2 = court-ordered). According to the usual rules of thumb, how would you describe this?
12. Report these results in APA style.
13. How would you interpret these results? Consider strengths and weaknesses of the research methods used in this study. There are at least three different interpretations that are consistent with the results.
14. How could you improve the design of this study to improve its internal validity?

* * *

The following problems refer to scores on two scales for the first 20 inmates in the parole data set introduced in an earlier chapter. LCSF1 is the Irresponsibility scale, and LCSF4 is the Social Rule Breaking scale.

| LCSF1 | LCSF4 |
|---|---|
| 0 | 1 |
| 1 | 2 |
| 0 | 0 |
| 1 | 0 |
| 0 | 0 |
| 1 | 0 |
| 2 | 4 |
| 1 | 1 |
| 4 | 4 |
| 0 | 0 |
| 1 | 0 |
| 3 | 0 |
| 1 | 0 |
| 3 | 1 |
| 3 | 3 |
| 4 | 4 |
| 3 | 1 |
| 0 | 0 |
| 1 | 1 |
| 4 | 0 |

15. What is the researcher's hypothesis?
16. What are the statistical hypotheses ($H_0$ and $H_1$)?
17. Did you choose to use a 2-tailed or a 1-tailed test, and why?
18. Should this test be performed using $\alpha = .05$ or $\alpha = .01$, and why?
19. Why should a related samples *t* test be performed, rather than a one-sample *z* or *t* test?
20. What is *df* for this *t* test?
21. What is the critical region for this *t* test? (Use the *t* table in Appendix A.)
22. Calculate the value of *t* for this sample of data. To help, note that $M_D = 0.55000$ and $SD_D = 1.39454$.
23. What is your statistical decision: Would you reject or retain $H_0$?
24. What kind of error—Type I or Type II—might you be making?
25. What is the size of the effect, using Cohen's *d*? To help, note that $M_1 = 1.65000$, $SD_1 = 1.46089$, $M_2 = 1.10000$, and $SD_2 = 1.48324$ (where subscripts of 1 = Irresponsibility and subscripts of 2 = Social Rule Breaking). According to the usual rules of thumb, how would you describe this?
26. Report these results in APA style.

* * *

27. Use SPSS to analyze the SAT data from this chapter and verify that you get the same results and reach the same conclusions. Notice that you have to calculate Cohen's *d* by hand because SPSS does not do this for you. Check that you get the correct value for *d* by calculating that yourself.
28. Use SPSS to analyze the Alcoholics Anonymous data from the first series of problems and verify that you get the same results and reach the same conclusions as when you did the calculations by hand.
29. Use SPSS to analyze the parole data from the second series of problems and verify that you get the same results and reach the same conclusions as when you did the calculations by hand.

**Problems 1 – 14 are due at the beginning of class.**

# 10. Independent Groups *t* Test

## Overview

The related samples *t* test built upon the one sample *t* test to allow the comparison of two conditions in a within-subjects design. The **independent groups *t* test** builds in another way to allow the comparison of two conditions in a between-subjects design. This chapter introduces this third and final kind of *t* test, which is also known as an **independent samples *t* test**, a **grouped *t* test**, or an **unpaired *t* test**.

## Pooling the *SD*s

Many research designs involve a comparison across two groups of subjects. Subjects might be randomly assigned to conditions (e.g., treatment and control groups), they might select their own conditions (e.g., attending a public or a private school), or they might simply belong to these conditions (e.g., men and women).

In each of these instances, the key to making a statistical comparison is to pool the within-group *SD*s for the groups. This enables us to estimate how much we would expect the groups' *M*s to differ by chance. This is done in two steps. Let's begin by seeing what this looks like for equal-size groups (i.e., $n_1 = n_2$), in which case we can just use $n$ to represent the size of each group. First, we pool the *SD*s as was shown in the last chapter:

$$SD_p = \text{sqrt}((SD_1^2 + SD_2^2) / 2)$$

Second, we calculate the standard error of the difference between the means:

$$SD_{M1\text{-}M2} = \text{sqrt}(2) \times SD_p / \text{sqrt}(n)$$

Recall that for a one sample *t* test, the standard error looked like this:

$$SD_M = SD / \text{sqrt}(N)$$

That's essentially the same thing that we have for two groups, with the only important difference being that the sampling error is twice as large for two groups as it is for one group.[26]

Fortunately, we need not have equal-size groups to perform valid statistical tests. These equations can be extended to accommodate groups that are unequal in size:

$$SD_p = \text{sqrt}((SD_1^2 \times df_1 + SD_2^2 \times df_2) / (df_1 + df_2))$$

In this formula, $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$ are used as weights for the groups' *SD*s when pooling them. The standard error formula can also deal with unequal group sizes:

$$SD_{M1\text{-}M2} = \text{sqrt}((SD_p^2 / n_1) + (SD_p^2 / n_2))$$

[26] The doubling in sampling error occurs when expressed as sampling variance. When expressed in *SD* units, the multiplier becomes sqrt(2) because *SD* is the square root of variance.

## Statistical Power

What is probably not obvious in these formulas is that, all else being equal, the standard error for an independent groups *t* test will be larger than the standard error for related samples *t* test. In other words, a larger difference between the means would be expected by chance for a between-subjects design than for a within-subjects design.

To illustrate this very important fact, let's revisit the data on psychology majors' SAT scores introduced in the last chapter. We'll use the same 20 test scores, but this time we'll pretend that the data come from separate groups: 10 students took the SAT Math test, and 10 different students took the SAT Verbal test.

For a related samples *t* test, we calculated $SD_{MD} = 13.16561$. This indicates that the expected difference between $M_D$ and $\mu_D$ is about 13.17 points. Let's see how large the standard error is if we treat the exact same scores as though they came from a between-subjects design. First, when we pooled the *SD*s, we found $SD_p = 65.61673$. The next step is to calculate the standard error. We can use the simpler version of the formula because the groups are equal in size:

$$SD_{M1\text{-}M2} = \text{sqrt}(2) \times SD_p \,/\, \text{sqrt}(n) = \text{sqrt}(2) \times 65.61673 \,/\, \text{sqrt}(10) = 29.34469$$

This is more than twice as large as the standard error for the related samples *t* test. Why the difference? Because the related samples *t* test removes individual differences through the calculation of difference scores. In this case, that means that the test gets rid of the fact that some individuals score high on both the math and the verbal tests, whereas others score low on both. With independent groups, we can't do this. The individual differences in overall cognitive ability are part of the normal sampling error and cannot be removed from the analysis. This makes it considerably more difficult to reject $H_0$. In other words, statistical power is much lower.

## Performing the *t* Test

For the independent groups *t* test, the statistical hypotheses very directly state the relationship between the two populations from which the groups' scores were drawn:

$H_0$: $\mu_1 = \mu_2$
$H_1$: $\mu_1 \neq \mu_2$

These hypotheses are nondirectional, which is the norm. You could make them directional if there is sufficient justification. We've already seen how to calculate the standard error, so all that's left is to calculate the *t* value:

$$t = ((M_1 - M_2) - (\mu_1 - \mu_2)) \,/\, SD_{M1\text{-}M2}$$

Because we're testing $H_0$: $\mu_1 = \mu_2$, this means that $\mu_1 - \mu_2 = 0$, and the formula simplifies:

$$t = (M_1 - M_2) \,/\, SD_{M1\text{-}M2}$$

For the SAT data, here's what we get:

$$t = (M_1 - M_2) \,/\, SD_{M1\text{-}M2} = (585.00000 - 635.00000) \,/\, 29.34469 = -1.70$$

To determine whether this falls in the critical region, we consult the *t* table (see Appendix A) for a 2-tailed test with $\alpha = .05$ and $df = N - 2 = 20 - 2 = 18$ and find the critical *t* value is 2.101. Because this is a 2-tailed test, the critical region includes the left tail ($t < -2.101$) and the right tail ($t > 2.101$). This is expressed most simply as $|t| > 2.101$. The absolute value of *t*, 1.70, does not exceed 2.101, so we'd retain $H_0$. There is no statistically significant difference between these students' SAT Verbal and Math scores.

Notice that this conclusion differs from what we found using the related samples *t* test with the same data. In that case, the *t* value of -3.80 was in the critical region, so $H_0$ was rejected. The *M*, *SD*, and *n* for each condition were identical. All that changed was the research design. For related samples, individual differences were removed and this enabled the test to detect a difference between the means. For independent groups, individual differences masked this difference between the means. This is the statistical reason why you should try to use within-subjects designs whenever possible.

## Effect Size

Cohen's *d* is calculated in the same way for an independent groups *t* test that it was for a related samples *t* test:

$$d = (M_1 - M_2) / SD_p$$

As shown earlier in this chapter, there are two versions of the formula for $SD_p$. Choosing between them depends only on whether the group sizes are equal. For the SAT data, in which the groups are equal in size, the calculations are as follows:

$$SD_p = \text{sqrt}((69.96031^2 + 60.96447^2) / 2) = 65.61673$$

$$d = (585.00000 - 635.00000) / 65.61673 = -0.76$$

Notice that this is identical to what was calculated in the last chapter. This is as it should be, for we used the same 20 scores. The size of the effect remained the same. Also note, once again, that it makes no difference which condition you treat as 1 and which as 2. The only difference would be the sign of *d*, and that's unimportant for interpreting the size of the effect.

## Using SPSS

To perform an independent groups *t* test in SPSS, you first enter your data into two separate variables (columns), here labeled "Test" (coded as 1 = Math, 2 = Verbal) and "SAT". Note that you have to create a variable that indicates group membership for each subject, and the dependent variable is placed in a separate column for all subjects.

| | Test | SAT |
|---|---|---|
| 1 | 1.00 | 540.00 |
| 2 | 1.00 | 630.00 |
| 3 | 1.00 | 590.00 |
| 4 | 1.00 | 530.00 |
| 5 | 1.00 | 490.00 |
| 6 | 1.00 | 660.00 |
| 7 | 1.00 | 590.00 |
| 8 | 1.00 | 490.00 |
| 9 | 1.00 | 650.00 |
| 10 | 1.00 | 680.00 |
| 11 | 2.00 | 590.00 |
| 12 | 2.00 | 620.00 |
| 13 | 2.00 | 650.00 |
| 14 | 2.00 | 610.00 |
| 15 | 2.00 | 580.00 |
| 16 | 2.00 | 720.00 |
| 17 | 2.00 | 660.00 |
| 18 | 2.00 | 560.00 |
| 19 | 2.00 | 610.00 |
| 20 | 2.00 | 750.00 |

Next, you use the following command:

```
t-test groups = test(1,2)
/vars = sat
```

Note that you have to provide the group membership variable (here, "Test") and the dependent variable (here, "SAT"). The output you need appears in two tables. The first table, labeled "Group Statistics", provides the *M* and *SD* for each condition:

**Group Statistics**

| | Section of Test | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| SAT Score | Math | 10 | 585.0000 | 69.96031 | 22.12339 |
| | Verbal | 10 | 635.0000 | 60.96447 | 19.27866 |

The second table, labeled "Independent Samples Test", provides the *t* value, *df*, and the *p* value (labeled as "Sig. (2-tailed)", which you'd divide by 2 if you want a 1-tailed test):

**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 |
| SAT Score | Equal variances assumed | .432 | .519 | -1.704 | 18 | .106 | -50.00000 | 29.34469 | -111.65092 | 11.65092 |
| | Equal variances not assumed | | | -1.704 | 17.669 | .106 | -50.00000 | 29.34469 | -111.73369 | 11.73369 |

When you're using this table, you must choose whether or not to assume the variances for the two groups are equal. For this course, we'll go ahead and make that assumption. That means you find the *t* value, *df*, and the *p* value in the top row (labeled "Equal variances assumed").

## APA Style

You can report the results of any *t* test in a single sentence. Here's what that might look like for these data:

Psychology majors did not score statistically significantly differently on the Math ($M$ = 585.00, $SD$ = 69.96) and Verbal ($M$ = 635.00, $SD$ = 60.96) sections of the SAT, $t(18) = -1.70$, $p$ = .106, $d$ = -0.76.

Notice that the phrasing indicates this was a 2-tailed test. Providing the means for both conditions allows readers to see the direction of the difference, even though it was not statistically significant in this case.

## Problems

The following problems refer to a study of two groups of men who attend Alcoholics Anonymous (AA) meetings. Members of group 1 (volunteers) choose to join AA and attend at least three meetings, and members of group 2 (court-ordered) are convicted of DUI and ordered by a court to attend at least three meetings. Prior to attending AA, subjects are matched on their age, race, and income level to form 12 pairs. After the third meeting, each subject is asked by his AA sponsor how many drinks he consumed each day in the past week. The data are shown below, with the number of drinks for both members of each pair in the same row.

You've already worked with these data in the last chapter. This time, treat them as though this is a between-subjects design that you'd analyze with an independent groups *t* test. Many, but not all, of your answers should be the same as last time.

| Volunteer | Court-Order |
|---|---|
| 16 | 23 |
| 8 | 7 |
| 9 | 16 |
| 5 | 14 |
| 10 | 8 |
| 14 | 17 |
| 13 | 22 |
| 3 | 0 |
| 5 | 9 |
| 19 | 16 |
| 8 | 13 |
| 9 | 17 |

1. What is the researcher's hypothesis?
2. What are the statistical hypotheses ($H_0$ and $H_1$)?
3. Did you choose to use a 2-tailed or a 1-tailed test, and why?
4. Should this test be performed using $\alpha = .05$ or $\alpha = .01$, and why?
5. Why should an independent groups *t* test be performed, rather than a related samples *t* test?
6. What is *df* for this *t* test?
7. What is the critical region for this *t* test? (Use the *t* table in Appendix A.)
8. Calculate the value of *t* for this sample of data. To help, note that $M_1 = 9.91667$, $SD_1 = 4.77605$, $M_2 = 13.50000$, and $SD_2 = 6.57129$ (where subscripts of 1 = volunteer and subscripts of 2 = court-ordered).
9. What is your statistical decision: Would you reject or retain $H_0$?

10. What kind of error—Type I or Type II—might you be making?
11. What is the size of the effect, using Cohen's *d*? According to the usual rules of thumb, how would you describe this?
12. Report these results in APA style.
13. Compare your results to what you found when you performed a related samples *t* test for these data. What are the similarities and differences?

* * *

The following problems refer to scores on two scales for inmates in the parole data set introduced in an earlier chapter. LCSF1 is the Irresponsibility scale, and LCSF4 is the Social Rule Breaking scale. Pretend that these data were obtained from two separate groups of 20 inmates apiece.

14. What is the researcher's hypothesis?
15. What are the statistical hypotheses ($H_0$ and $H_1$)?
16. Did you choose to use a 2-tailed or a 1-tailed test, and why?
17. Should this test be performed using $\alpha = .05$ or $\alpha = .01$, and why?
18. What is *df* for this *t* test?
19. What is the critical region for this *t* test? (Use the *t* table in Appendix A.)
20. Calculate the value of *t* for this sample of data. To help, note that $M_1 = 1.65000$, $SD_1 = 1.46089$, $M_2 = 1.10000$, and $SD_2 = 1.48324$ (where subscripts of 1 = Irresponsibility and subscripts of 2 = Social Rule Breaking).
21. What is your statistical decision: Would you reject or retain $H_0$?
22. What kind of error—Type I or Type II—might you be making?
23. What is the size of the effect, using Cohen's *d*? According to the usual rules of thumb, how would you describe this?
24. Report these results in APA style.
25. Compare your results to what you found when you performed a related samples *t* test for these data. What are the similarities and differences?

| LCSF1 | LCSF4 |
|---|---|
| 0 | 1 |
| 1 | 2 |
| 0 | 0 |
| 1 | 0 |
| 0 | 0 |
| 1 | 0 |
| 2 | 4 |
| 1 | 1 |
| 4 | 4 |
| 0 | 0 |
| 1 | 0 |
| 3 | 0 |
| 1 | 0 |
| 3 | 1 |
| 3 | 3 |
| 4 | 4 |
| 3 | 1 |
| 0 | 0 |
| 1 | 1 |
| 4 | 0 |

* * *

26. Use SPSS to analyze the SAT data from this chapter and verify that you get the same results and reach the same conclusions. Notice that you have to calculate Cohen's *d* by hand because SPSS does not do this for you. Check that you get the correct value for *d* by calculating that yourself.

27. Use SPSS to analyze the Alcoholics Anonymous data from the first series of problems and verify that you get the same results and reach the same conclusions as when you did the calculations by hand.
28. Use SPSS to analyze the parole data from the second series of problems and verify that you get the same results and reach the same conclusions as when you did the calculations by hand.

**Problems 1 – 13 are due at the beginning of class.**

# 11. Overview of ANOVA

## Overview

To this point, statistical analyses have been introduced that afford comparisons between a single sample and a population mean, between two related samples, or between two independent groups. This covers many of the most popular research designs, yet it remains fairly limited. This chapter will examine ways that the **analysis of variance** (ANOVA), a technique for breaking the variation in scores into its component parts, expands upon the *t* test. ANOVA models afford comparisons between more than two independent groups or more than two related samples, and they can also be used to test the effects of more than one independent variable. These extensions enable the use of a much wider array of research designs.

## Extending Beyond Two Conditions

Researchers often include more than two conditions in a study. For example, a clinical scientist might randomly assign subjects to one of three treatment groups, but the *t* test only compares two conditions. Thus, we'd have to test repeatedly to analyze all of the data from a study with more than two conditions. Our clinical scientist could perform three *t* tests to compare all treatments to one another: conditions 1 vs. 2, 1 vs. 3, and 2 vs. 3.

In addition to the fact that it would be tiresome to run multiple *t* tests, there's a more serious problem. In the event that there really are no differences across conditions, each test introduces another chance of making a Type I error (a false alarm). The chance of making at least one Type I error is known as the **experimentwise Type I error rate**; it's also sometimes referred to as the **familywise Type I error rate**. Here's a table that illustrates the magnitude of this problem for studies with varying numbers of conditions; tests are performed using $\alpha = .05$.

| Number of Conditions | Number of *t* Tests | Experimentwise Type I Error Rate |
|---|---|---|
| 2 | 1 | .050 |
| 3 | 3 | .143 |
| 4 | 6 | .265 |
| 5 | 10 | .401 |
| 6 | 15 | .537 |
| 7 | 21 | .659 |
| 8 | 28 | .726 |
| $k$ | $m = k \times (k - 1) / 2$ | $1 - (1 - \alpha)^m$ |

When there are only two conditions, the single *t* test carries an $\alpha = .05$ probability of making a Type I error. With three conditions, the use of three *t* tests increases the chance of making at least one Type I error to .143. By the time you reach six conditions, the experimentwise Type I error rate exceeds .50, meaning you're more likely than not to reach at least one mistaken statistical decision. The bottom row contains the general expression

for the experimentwise Type I error rate that applies for any number of conditions $k$ and any $\alpha$ level.

The problem of an experimentwise Type I error rate that can be much larger than $\alpha$ reveals the value of having a statistical test that can hold the Type I error rate down to the $\alpha$ level when simultaneously comparing all conditions to one another. That is what the $F$ test provided by an ANOVA does.[27] A single $F$ test is used to test the null hypothesis of no difference between any conditions' means, symbolized as follows:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1: \sim(\mu_1 = \mu_2 = \ldots = \mu_k)$$

The expression for $H_0$ indicates the equality of all population means from which the $k$ conditions' scores are drawn. The expression would include $k$ terms, one for each of the $k$ conditions. The expression for $H_1$ is simply the negation of the expression for $H_0$. The "~" symbol means "it is not the case that", and the expression in parentheses is $H_0$ itself. By definition, $H_0$ and $H_1$ are nondirectional. The means for all conditions are compared to see whether they differ from one another in any way—not in some specific way. If you have more specific predictions than this, the $F$ test itself is not the way to test them. Instead, you'll need to follow the $F$ test with another procedure.[28]

## Multiple Comparisons and Post-Hoc Tests

The $F$ test solves the problem of large experimentwise Type I error rates by performing a single test rather than a series of tests. That limits the chance of making a Type I error to the desired $\alpha$ level. This benefit comes at a cost, however. Whenever an $F$ test leads us to reject $H_0$, it is not clear which particular conditions differ from one another. Rejecting $H_0$ is vague. It tells us only that there is some difference between the conditions. To be more specific about which conditions differ requires that we make **multiple comparisons**, a series of tests between subsets of the conditions.

One common way of making multiple comparisons is to compare every condition to every other condition using a **post-hoc test**. This works in a two-step process. First, you calculate a threshold value for statistical significance. Second, you calculate the differences between the means for all pairs of conditions to determine which exceed the threshold.

For example, suppose that you have three treatment groups numbered 1, 2, and 3 with means of 10, 20, and 30, respectively. If the $F$ test allows you to reject $H_0$, this suggests that there is some difference among the conditions. You'd then calculate the threshold for a post-hoc test to see which conditions differ by enough to conclude that they're statistically significantly different. Suppose this threshold is 15 points. This would indicate that treatments 1 and 2 do not differ (20 – 10 = 10, which is less than the threshold of 15), nor do treatments 2 and 3 (30 – 20 = 10, again less than the threshold of 15). Treatments 1 and

---

[27] The test is named for its creator, Sir Ronald Fisher. Think what you will of ANOVA, the F stands for Fisher and nothing more.

[28] Alternatively, you can perform tests of planned contrasts. This is often the best approach, but its implementation goes beyond an introductory course in statistics and will not be discussed here. An excellent source on this subject is Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

3, however, do differ statistically significantly (30 – 10 = 20, which exceeds the threshold of 15). This is how a post-hoc test is used to make all pairwise comparisons.

There are many post-hoc tests available, and they differ in how liberal or conservative they are. These terms are used here in their statistical—rather than political—sense. Liberal statistical procedures (e.g., using $\alpha = .05$) make it easier to reject $H_0$ and conservative statistical procedures (e.g., using $\alpha = .01$) make it harder to reject $H_0$. A liberal procedure increases the chance of making a Type I error (false alarm) and decreases the chance of making a Type II error (miss), and a conservative procedure does the reverse. Neither approach is uniformly better or worse. Making a smart choice depends on the research context. For example, whereas in exploratory research one usually wants to avoid Type II errors by using liberal procedures, in hypothesis-testing research one usually wants to avoid Type I errors by using conservative procedures.

The post-hoc test that will be introduced and used exclusively in this text, **Tukey's *HSD***, is not particularly liberal or conservative.[29] It's more of a "middle of the road" technique. This doesn't make it the best choice, as you might want a more liberal post-hoc test (e.g., Fisher's *LSD*[30]) for exploratory research or a more conservative one (e.g., Sheffé's method) for hypothesis-testing research. Once you understand when and how to use Tukey's *HSD*, it would be extremely easy to use another post-hoc test because they operate by the same logic and steps. For example, a very conservative post-hoc test might have generated a threshold of value of 22 points for the comparison of three means introduced above, in which case none of the treatments would have differed statistically significantly from one another (no pair of means differed by more than 22 points). A very liberal post-hoc test might have generated a threshold value of 6 points, in which case all three treatments would have differed statistically significantly from one another (all pairs of means differed by more than 6 points).

To sum up the procedure for testing for differences between more than two conditions, you begin by using the *F* test and, if needed, you proceed to use a post-hoc test. The *F* test holds the rate of Type I errors to the chosen $\alpha$ level. If you retain $H_0$, you're done. You conclude that there are no statistically significant differences across conditions. If you reject $H_0$, you use a post-hoc test to see which conditions differ from one another. Just as you can select a higher or lower $\alpha$ level for *F* when testing $H_0$, you can select a post-hoc test that is more liberal or conservative in comparing conditions to one another.

## Partitioning the Variance

The formula for a *t* test reveals how it compares conditions to each other. The numerator is the difference between two means, and the denominator is the difference expected due to sampling error. If this ratio of mean difference to standard error is sufficiently large that it falls in the critical region, you reject $H_0$.

---

[29] Two notes on the name. First, its creator was John Tukey. He was not John Turkey, so it's not a Turkey Test. Second, *HSD* stands for "honestly significant difference". I'm not sure whether John Tukey thought other post-hoc tests were somehow dishonest, but that's the name he chose for this test and it stuck.

[30] Sir Ronald Fisher, of *F* test fame, introduced the *LSD* post-hoc test. The *LSD* stands for "least significant difference" and has nothing to do with psychedelic drugs. It's very much like performing all pairwise *t* tests.

The formula for an *F* test is different. We're not going to delve into the details of how *F* tests are calculated because it gets pretty tedious, and if you try it you might spend more time correcting arithmetic mistakes than learning about statistical concepts. What's most important is to understand the logic by which these tests operate.

An *F* test provides a ratio of **systematic variance** to **error variance**. Systematic variance refers to variation in scores due to differences across the conditions being studied. Error variance refers to unexplained variation in scores that remains within the conditions. It's important to note that the term "error variance" does not imply that any mistakes were made or that a study is biased in any way. Though we can and should take steps to minimize it, some amount of error variance is unavoidable in research (e.g., sampling error). We can't measure everything that might affect the dependent variable, so some of its variance will remain unexplained. We call this error variance to differentiate it from sources of systematic variance that can be explained by variables that are manipulated or measured in research. So here's the conceptual definition of the *F* ratio:

*F* = systematic variance / error variance

Let's flesh out the example of a clinical study with three treatment conditions. Suppose there are 50 subjects in each condition, for a total of 150 subjects. These 150 subjects' scores on the dependent variable (outcome following treatment) will vary, and this is the **total variance** in the study. For simplicity, let's use an arbitrary scale and say that there are 60 units of total variance on the dependent variable. The important question then becomes: How much of this total variance of 60 units is systematic, and how much is error?

At one extreme, it's possible that all of this is error variance. That would mean that there are no differences between groups, that the only sources of variation in scores are things like individual differences (i.e., subjects start out in better or worse shape than one another, and the treatment itself had no effect) and measurement error (i.e., nothing can be assessed with perfect reliability). In this case, the *F* ratio would take on its minimum possible value: *F* = systematic variance of 0 / error variance of 60 = 0.

At the other extreme, it's possible that all 60 units of total variance are systematic variance. That would mean that the only reason the 150 subjects' scores differ is because of the treatments they received, with no variation in scores remaining within treatment groups. In this case, the *F* ratio would take on its maximum possible value: *F* = systematic variance of 60 / error variance of 0 = ∞.

As you might have recognized already, neither of these extremes is likely to happen in practice. No matter how effective or ineffective the treatments may be, individuals will still differ to some extent within groups. There will always be some error variance. Likewise, there will almost always be at least some systematic variance, or even just the appearance of systematic variance due to sampling error. Because of this, the *F* ratio will be greater than 0 but less than ∞. In short, it will be some positive number.

Suppose we find that there is twice as much systematic variance as error variance, in other words that *F* = systematic variance of 40 / error variance of 20 = 2.00. Is this statistically significant? Whether *F* is sufficiently large to reject $H_0$ presents the usual problem of determining whether it falls in the critical region of the appropriate sampling distribution. Though tables of critical *F* values can be consulted, it's more convenient to compare the *p* value provided by a computer to one's chosen $\alpha$ level. As usual, we reject $H_0$ if $p < \alpha$ and retain $H_0$ otherwise.

## Degrees of Freedom

With a *t* test, there was a single value for degrees of freedom (*df*). This relates to sample size (i.e., it's $N - 1$ for a one sample or related samples *t* test, $N - 2$ for an independent groups *t* test). Because all *t* tests compare means for two conditions, this fact doesn't have to be specified.

With an *F* test, however, the number of conditions being compared does need to be specified. It can be only two,[31] or it can be much larger. Because of this, there are two *df* values for an *F* test. The first *df* value relates to the number of conditions being compared (specifically, it's $k - 1$, where $k$ is the number of conditions). The second *df* value is like the *df* for a *t* test in that it relates to sample size. For example, in a between-subjects design, it's $N - k$, where $N$ is the total sample size. In a study with 150 subjects assigned to 3 conditions, the *df* for the *F* test would be $k - 1 = 2$ and $N - k = 147$.

How to calculate *df* for all varieties of *F* tests we'll explore isn't important. In the chapters on ANOVA models that follow, we'll obtain the *df* from computer output rather than calculating them. It's important to understand why there are two *df* values for an *F* test, though. You can think about it this way: There really are two *df* values for a *t* test or an *F* test, with the first relating to the number of conditions compared and the second relating to sample size. With a *t* test, the first *df* value is always 1, so we don't bother to report it.

## Effect Size

When comparing more than two conditions, we can no longer use Cohen's *d* as our measure of effect size. Instead, we need a measure that can accommodate more than two means. The most popular measure used with *F* tests is $\eta^2$, which represents the proportion of variance in the dependent variable that can be explained by the independent variable.[32] Whereas *d* can be positive or negative in sign and can range from 0 to $\infty$ in absolute value, $\eta^2$ can only range from .00 to 1.00.

If $\eta^2 = .00$, then none of the variation in outcomes (0%) can be attributed to the independent variable. If $\eta^2 = 1.00$, then all of the variation in outcomes (100%) can be attributed to the independent variable. Naturally, you won't see either of these extreme values in practice. Cohen introduced the following rules of thumb for interpreting $\eta^2$:

.01 = small
.09 = medium
.25 = large

These values may be tough to remember because they're unequally spaced. We'll see in a later chapter that these values are related to our final measure of effect size, the correlation coefficient (*r*). It might be helpful to remember the rules of thumb for $\eta^2$ by

---

[31] With only two conditions, $F = t^2$. Though the formulas for *F* and *t* appear quite different, they are equivalent tests that will yield identical *p* values when there are only two conditions.

[32] The Greek letter η is pronounced "eta", so $\eta^2$ is "eta squared". In many ANOVA models, technically what we'll be using is partial $\eta^2$, where the term "partial" indicates that effects other than the one of interest have been statistically removed or "partialled out". For simplicity, this text will refer only to $\eta^2$ even when the measure is more fully expressed as partial $\eta^2$.

noting that they're the squared values of the rules of thumb for *r*, which are evenly spaced (.10 = small, .30 = medium, and .50 = large). Squaring these rule-of-thumb values for *r* yields $r^2$ = .01, .09, and .25. It turns out that $r^2$ is defined in the same way as $\eta^2$, the proportion of variance in the dependent variable that can be explained by the independent variable. Thus, the rules of thumb for interpreting $\eta^2$ are the same as those for interpreting $r^2$. And these are most easily remembered as the rules of thumb for *r* squared.

As will be the case for *df*, the formulas for $\eta^2$ will not be presented in the following chapters on ANOVA models. Instead, we'll obtain $\eta^2$ from computer output.

Just as you should always report the value of Cohen's *d* following the results of a *t* test, you should always report the value of $\eta^2$ following the results of an *F* test. If you then proceed to compare specific conditions using a post-hoc test, you can use *d* as a measure of effect size for some or all of these. This is optional but it can be helpful. For example, you might report the value(s) of *d* for one or more comparisons of special interest or importance, or you might report that all values of *d* were less than (or greater than) a particular value to summarize how small (or large) all of the differences were.

## Extending Beyond One Factor

The kind of ANOVA introduced above can be used for a between-subjects design, like the treatment study in which subjects were assigned to different conditions, or a within-subjects design in which the same subjects are measured in more than one condition. Just like with *t* tests, when the design is between-subjects we use an **independent groups ANOVA** and when the design is within-subjects we use a **related samples ANOVA**. These two ANOVA models will be described more fully in the next two chapters.

Suppose that our clinical scientist wanted to examine not only differences across three treatment conditions, but also differences across men and women. This would be a **factorial design** because it incorporates more than one **factor**, or independent variable. Factor A is treatment condition, and it has three **levels** (the three treatments). Factor B is gender, and it has two levels (men and women). This is an example of a 3 (treatment) × 2 (gender) factorial design, which therefore has a total of 3 × 2 = 6 **cells**, or conditions, in the full design.[33] This new terminology—factors, levels, and cells—is reserved for factorial designs. In single-factor designs, we refer more simply to the conditions in the study without labeling them as levels of a factor or cells in a design.

To analyze the data from a factorial design we use a **factorial ANOVA**. Whereas an independent groups ANOVA or a related samples ANOVA will provide a single *F* test, a factorial ANOVA will provide more than one *F* test. Some of these will test for what are called **main effects**, and some for **interaction effects**.

A main effect refers to differences across the levels of one factor, collapsing across the other factor(s) in the design. For example, to test for a main effect of treatment conditions, the clinical scientist would pool the results for men and women. Then, to test for a main effect of gender, results for the three treatments would be pooled. For each factor in the

[33] It's arbitrary which factor is labeled as A and which as B. This design could just as easily be described as a 2 (gender) × 3 (treatment) factorial ANOVA. Either way, it's a design with two factors, one with three levels and one with two levels, for a total of six cells.

design, a factorial ANOVA will provide an *F* test to determine whether that main effect is statistically significant.

An interaction effect is the joint influence of two or more factors. If the effect of treatment depends on gender (e.g., if the treatment is more effective for women than for men), this would be an interaction of factors A and B. The factorial ANOVA would provide an *F* test to determine whether this interaction effect is statistically significant. If there are more than two factors, there will be multiple *F* tests for potential interactions. For example, with three factors A, B, and C, a factorial ANOVA would test for interactions labeled as A × B, A × C, and B × C (referred to as two-way interactions because they involve two factors) as well as A × B × C (a three-way interaction).

In principle, there's no limit to how complex a factorial design we can accommodate with a corresponding ANOVA model. You can include two, three, four, or more factors, each of which can have two, three, four, or more levels. Also, each factor can be either between-subjects or within-subjects, and you can mix and match these types of factors in the design.

Though there's no theoretical limit to how complex a factorial design we can conceive, there are practical limits to what we can implement. Each between-subjects factor poses the difficulty of recruiting enough subjects to flesh out all the levels (separate groups) of that factor. Each within-subjects factor increases the burden on subjects who must be tested repeatedly in all the levels (conditions) of that factor. As the number of factors, levels, and cells in the design increases, these challenges multiply. There is a trade-off between the complexity of the design you might like (e.g., more complex designs can address more research questions and/or control for more variables of interest) and your ability to actually perform that study well (e.g., obtaining sufficient data within each cell of the design).

## Problems

1. If you have a study with more than two conditions, why is it better to perform a single *F* test rather than a series of *t* tests?
2. When you perform an *F* test to compare more than two conditions and you reject $H_0$, what's ambiguous about this conclusion? How is this ambiguity resolved?
3. Why do we need to report two *df* values for an *F* test, but only one *df* value for a *t* test?
4. Suppose you're reading a research paper and the authors report a negative *F* ratio. Why can you be certain that they've either calculated or reported this incorrectly?
5. According to Cohen's rules of thumb, which effect is larger: $d = 0.30$ or $\eta^2 = .30$? Following APA style, (a) why do we report *d* with a leading 0 but $\eta^2$ without one and (b) why is the letter $\eta$ from the statistic $\eta^2$ not italicized?

* * *

The following problems refer to a study of distracted driving. Using a driving simulator, each subject will be tested under three conditions: While maintaining a conversation on a cell phone, while maintaining a conversation with a passenger, and with no conversation taking place. Dr. Flurpple will recruit young adults (less than 20 years old) and older adults (between 40 and 60) to test for age differences. The dependent variable will be the number

of errors made while driving (e.g., speeding, tailgating, failure to stop, veering out of the driving lane, crashing into any object).

6. What are the factors in this study's design? Indicate whether each is a between-subjects or a within-subjects factor, and name its levels.
7. How many cells are there in the design?
8. How many *F* tests will the factorial ANOVA provide? What will each one test?

* * *

For each of the following problems, determine what statistical test should be performed and explain why this is an appropriate selection. The range of tests to consider includes the *z* test, one sample *t* test, related samples *t* test, independent groups *t* test, independent groups ANOVA, related samples ANOVA, and factorial ANOVA.

9. A demographer working for the U.S. Census Bureau wants to compare salaries for urban vs. rural areas. She gets a sample of psychologists, some who live in urban areas and some who live in rural areas. Do earnings differ across these areas?
10. First-year college students were surveyed about how much they liked their roommates at three points in time: within five minutes of meeting them, after the first week of classes, and at the end of the semester. Ratings were made on a 7-point Likert scale. Does degree of liking change over the course of the semester?
11. A clinical psychologist wondered whether adults with attention deficit hyperactivity disorder (ADHD) had reflexes that differed in speed from those of the general population. She located a test of reaction time that was normed on adults in the U.S. ($\mu$ = 200 msec). From treatment centers in her home state, a random sample of 141 adults diagnosed with ADHD were tested for reaction time ($M$ = 220, $SD$ = 27). Do adults with ADHD differ in reaction time from the general population?
12. Each child in the 4$^{th}$ grade at a large elementary school is classified by the teacher as predominantly right-handed, left-handed, or ambidextrous. The children's art teachers rate their artistic ability on a 10-point scale. Does artistic ability differ by handedness?
13. A nutritionist wanted to find out if coffee and tea, as served in restaurants, differed in caffeine content. She went to 30 restaurants, ordered coffee and tea in each one, and had the caffeine content of each beverage tested. Do these servings of coffee and tea differ in caffeine levels?
14. A developmental psychologist is interested in the study of aggression. She observes aggressive behavior on school playgrounds to test for gender differences in both physical and verbal aggression. Does the level of aggression differ by gender, by type of aggression, or both?
15. A scientific supply company has developed a new breed of lab rat, which it claims weighs the same as the classic white rat ($\mu$ = 485 grams, $\sigma$ = 50 grams). A researcher obtained a sample of 76 of the new breed of rats, weighed them, and found $M$ = 515 grams. Is the company's claim true?

* * *

The following problems refer to the following experiment. A total of 24 office workers were given the chance to pay $1 for a lottery ticket for a prize of $25. One half of all tickets came with a random number already assigned, the other half were blank such that their purchasers could choose and write their own ticket numbers. Subjects were randomly assigned to conditions. After all tickets were sold, the researcher approached each subject individually to buy back the ticket. He or she was told that someone else wanted to enter the lottery but there were no more tickets, so the researcher would pay what it takes to buy back this ticket to offer it to the newcomer. The researchers recorded how much each subject charged to sell back his or her ticket. Tickets were purchased from subjects at one of three different times: (1) immediately after originally selling the ticket, (2) the next day, or (3) just before drawing the winning lottery number at the end of the week. Subjects were randomly assigned to conditions.

16. What is the dependent variable? What is its scale of measurement?
17. What are the factors in this study's design? Indicate whether each is a between-subjects or a within-subjects factor, and name its levels.
18. How many cells are there in the design?
19. How many *F* tests will the factorial ANOVA provide? What will each one test?

* * *

For each of the following problems, determine what statistical test should be performed and explain why this is an appropriate selection. The range of tests to consider includes the *z* test, one sample *t* test, related samples *t* test, independent groups *t* test, independent groups ANOVA, related samples ANOVA, and factorial ANOVA.

20. In 1997, Nabisco came out with a clever advertising campaign, the Chips Ahoy Challenge. Nabisco guaranteed that there were more than 1,000 chocolate chips in every bag, and they challenged consumers to count. Suppose 25 people go to the trouble of counting the chips in one bag apiece. Do their findings statistically significantly refute Nabisco's claim?
21. A clinical psychologist wanted to compare three treatments for Generalized Anxiety Disorder (GAD). She put an ad in the local paper to find people with GAD. Based on severity of symptoms, she matched the volunteers for her study into triads and randomly assigned each of the matched cases to one of the three treatments. Outcomes were assessed individually by a clinician blind to treatment assignments. Are the treatments equally effective?
22. An investigator wonders whether the reduced mental alertness due to sleep deprivation can be counteracted by consuming caffeine. Three groups of volunteers are subjected to varying amounts of sleep deprivation (0 hours, 1 hour, or 2 hours). One-half of all volunteers is given a standardized dose of caffeine, the other half is not. Does mental alertness differ by sleep deprivation, caffeine intake, or both?
23. A developmental psychologist wondered if birth order had an impact on academic performance. She found families with two children and obtained the high school GPA of each child. Is there a difference in GPA between first-born and second-born children?

24. A behavioral economist wonders whether portion sizes influence weight change. Rather than performing a one-shot experiment in the laboratory, she arranges for dining halls on three college campuses to systematically vary their portion sizes for one full semester. One serves small portions, another serves medium-sized portions, and the third serves large portions. Students who regularly eat in these dining halls are asked to weigh themselves at the beginning and the end of the semester, and their change in weight is the dependent variable. Does portion size affect weight change?
25. Across U.S. cities, the average vacancy rate for apartments is $\mu = 10\%$ ($\sigma = 4.6\%$). An urban studies major obtained a sample of 15 rust-belt cities and found that the average vacancy rate was $M = 13.3\%$. Does the vacancy rate for these cities differ from the U.S. average?
26. An exercise physiologist classifies people—on the basis of their body mass index, heart rate, and lung capacity—as above or below average in terms of fitness. He then directs the same people to walk on a treadmill, individually, at an increasing speed until they can no longer walk. The speed when a person maxes out is the dependent variable. Is there a difference in maximum walking speed based on fitness level?

**Problems 1 – 15 are due at the beginning of class.**

# 12. Independent Groups ANOVA

## Overview

The independent groups ANOVA extends the independent groups *t* test to between-subjects research designs that include more than two conditions. This chapter describes the procedure for performing and reporting the results of the *F* test and, if necessary, a post-hoc test for multiple comparisons.

## The ANOVA Model and the *F* Test

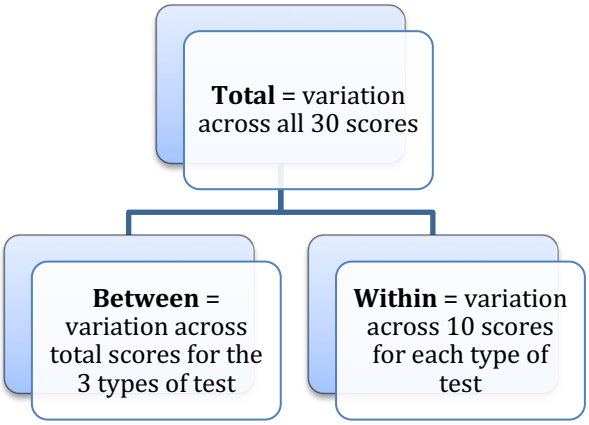
Consider a study designed to examine whether college students tend to score higher on verbal, quantitative, or spatial tests of cognitive ability, with 10 students randomly assigned to take each type of test. The 30 scores (3 conditions × 10 subjects apiece = 30) are shown below:

| Verbal | Quantitative | Spatial |
|---|---|---|
| 67 | 66 | 65 |
| 65 | 94 | 82 |
| 28 | 41 | 42 |
| 43 | 31 | 55 |
| 36 | 53 | 42 |
| 32 | 36 | 42 |
| 37 | 31 | 32 |
| 61 | 55 | 82 |
| 51 | 74 | 66 |
| 25 | 30 | 48 |

We could perform a series of three independent groups *t* tests to compare all conditions to one another (i.e., verbal vs. quantitative, verbal vs. spatial, and quantitative vs. spatial), but this would increase the experimentwise Type I error rate well above our desired $\alpha$ level. Instead, we can use an independent groups ANOVA to compare scores across all three conditions in a single *F* test that holds $\alpha$ to our desired level. The null and alternative hypotheses can be expressed as follows:

$H_0$: $\mu_1 = \mu_2 = \mu_3$
$H_1$: $\sim(\mu_1 = \mu_2 = \mu_3)$

Testing the null hypothesis involves calculating an *F* ratio. When subjects belong to independent groups, the total variance on the dependent variable can be split into two sources. The systematic source of variance is **between groups** and the error variance is what remains **within groups**. Between-groups variance is due to differences across conditions. In this case, that would indicate that subjects perform differently on the verbal, quantitative, and spatial tests. Within-groups variance is due to any other factors, such as individual differences (i.e., people differ in their overall levels of cognitive ability) and measurement error (i.e., tests of cognitive ability are not perfectly reliable). Schematically, here's how the independent groups ANOVA partitions the variance:




The *F* ratio for an independent groups ANOVA is calculated as the systematic variance (between groups) divided by the error variance (within groups). The critical region for the *F* test is based on the $\alpha$ level (usually .05) and the *df*. There are two *df* for any *F* test, one relating to the number of conditions and one relating to sample size. For an independent groups ANOVA, the *df* are calculated as $k - 1$ (where $k$ is the number of groups being compared) and $N - k$ (where $N$ is the number of subjects). In this case, the *df* would be $3 - 1 = 2$ and $30 - 3 = 27$. You can get the *df* from computer output.

You can consult a table of *F* values to determine whether the *F* ratio calculated from the data falls in the critical region, but it's easier to obtain the *p* value from computer output and compare this to the $\alpha$ level. As usual, if $p < \alpha$ you reject $H_0$, otherwise you retain $H_0$. For example, the *F* ratio for the cognitive ability data presented earlier is not statistically significant ($p = .410$, which is larger than $\alpha = .05$). We would retain $H_0$, and that's the end of the analysis.

## Multiple Comparisons

Whenever we reject $H_0$ and conclude there is some difference between conditions, this gives us license to make multiple comparisons to determine which conditions differ from one another. To do this, you can use a post-hoc test such as Tukey's *HSD*. The first step is to calculate a threshold value for statistical significance, and the second step is to make all pairwise comparisons of means to determine which differences exceed this threshold. Here, too, we'll rely on computer output to perform Tukey's *HSD* and indicate which conditions differ from one another statistically significantly.

## Effect Size

The measure of effect size for an independent groups ANOVA is $\eta^2$, which indicates the proportion of variance in the dependent variable that can be explained by the independent variable. This can be obtained from computer output. Cohen's rules of thumb for interpreting the size of $\eta^2$ are that .01 = small, .09 = medium, and .25 = large. For the cognitive ability data, $\eta^2 = .06$, which falls between a small and medium effect size.

If you like, you can also report one or more values of Cohen's *d* to indicate the size of pairwise comparisons. This measure would be calculated in the same way as for an

independent groups *t* test: The numerator is the difference between *M*s for two groups and the denominator is the pooled *SD* for those groups (weighted by *df* if groups are of unequal sizes).

## Using SPSS

To perform an independent groups ANOVA in SPSS, you first enter your data into two separate variables (columns), here labeled "Test" (coded as 1 = verbal, 2 = quantitative, 3 = spatial) and "Score" (each subject's test score). Note that you have to create a variable that indicates group membership for each subject, and the dependent variable is placed in a separate column for all subjects. The full data set didn't fit onto the screen, but here's what the beginning of the data file looks like:

| Test | Score |
|---|---|
| 1.00 | 67.00 |
| 1.00 | 65.00 |
| 1.00 | 28.00 |
| 1.00 | 43.00 |
| 1.00 | 36.00 |
| 1.00 | 32.00 |
| 1.00 | 37.00 |
| 1.00 | 61.00 |
| 1.00 | 51.00 |
| 1.00 | 25.00 |
| 2.00 | 66.00 |
| 2.00 | 94.00 |
| 2.00 | 41.00 |
| 2.00 | 31.00 |
| 2.00 | 53.00 |
| 2.00 | 36.00 |
| 2.00 | 31.00 |
| 2.00 | 55.00 |
| 2.00 | 74.00 |
| 2.00 | 30.00 |
| 3.00 | 65.00 |
| 3.00 | 82.00 |

Next, you use the following command:

```
unianova score by test
/posthoc test (tukey)
/print desc etasq
```

On the first line, you provide the dependent variable (here, "Score") and group membership variable (here, "Test"). On the second line, you indicate the group membership variable again. The third line requests descriptive statistics and $\eta^2$ as an effect size measure, and you don't need to change this line at all.

SPSS will provide many tables of output, but you can ignore all but three of them. I recommend deleting the ones you don't need because some of them look similar to those

you do need and it's easy to mistakenly read and report output from the wrong table. The first table, labeled "Descriptive Statistics", provides the *M* and *SD* for each condition:

**Descriptive Statistics**

Dependent Variable: Score on Test

| Type of Test | Mean | Std. Deviation | N |
|---|---|---|---|
| Verbal | 44.5000 | 15.56527 | 10 |
| Quantitative | 51.1000 | 21.57390 | 10 |
| Spatial | 55.6000 | 17.51317 | 10 |
| Total | 50.4000 | 18.33895 | 30 |

The second table, labeled "Tests of Between-Subjects Effects", provides the *F* value, *df*, *p* value (labeled as "Sig."), and $\eta^2$ (labeled as "Partial Eta Squared"). Use the middle row, labeled with your group membership variable (here, "Test"), to find the *F* value (here, 0.92), the first *df* value (here, 2), *p* the value (here, .410), and $\eta^2$ (here, .06). Use the next row, labeled "Error", to find the second *df* value (here, 27).

**Tests of Between-Subjects Effects**

Dependent Variable: Score on Test

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 623.400[a] | 2 | 311.700 | .922 | .410 | .064 |
| Intercept | 76204.800 | 1 | 76204.800 | 225.364 | .000 | .893 |
| Test | 623.400 | 2 | 311.700 | .922 | .410 | .064 |
| Error | 9129.800 | 27 | 338.141 | | | |
| Total | 85958.000 | 30 | | | | |
| Corrected Total | 9753.200 | 29 | | | | |

The third table, labeled "Homogeneous Subsets", provides the results of the Tukey's *HSD* post-hoc test. The way the table works is that any conditions whose means appear in the same subset do not differ statistically significantly. In this case, all conditions' means are in a single subset, hence there are no significant differences between conditions.

**Homogeneous Subsets**

**Score on Test**

Tukey HSD[a,b]

| Type of Test | N | Subset 1 |
|---|---|---|
| Verbal | 10 | 44.5000 |
| Quantitative | 10 | 51.1000 |
| Spatial | 10 | 55.6000 |
| Sig. | | .381 |

If there are differences, there will be two or more subsets in this table. To illustrate, I modified the cognitive ability data to introduce significant differences. Below are the homogeneous subsets results:

**Homogeneous Subsets**

Score on Test

Tukey HSD[a,b]

| Type of Test | N | Subset 1 | Subset 2 |
|---|---|---|---|
| Verbal | 10 | 35.5000 | |
| Quantitative | 10 | 48.1000 | 48.1000 |
| Spatial | 10 | | 61.6000 |
| Sig. | | .123 | .093 |

For these modified data, the results show no significant difference between the verbal and quantitative tests; their means appear together in subset 1. Likewise, there was no significant difference between the quantitative and spatial tests; their means appear together in subset 2. There was, however, a significant difference between verbal and spatial tests; their means never appear in the same subset.

## APA Style

When you retain $H_0$, you can report the results of an independent groups ANOVA in a single sentence. You can include the *M* and *SD* for each condition if you like, but that's considered optional for results that are not statistically significant. Here's what the report might look like for the original cognitive ability data:

Groups of 10 college students apiece took verbal, quantitative, and spatial tests of cognitive ability, and there was no statistically significant difference in performance across conditions, $F(2, 27) = 0.92$, $p = .410$, $\eta^2 = .06$.

When you reject $H_0$, you begin by reporting the results of the *F* test in a single sentence and then follow this with the post-hoc test results. Note that you should not only include the *M* and *SD* for each condition, but also specify the $\alpha$ level (usually .05) and procedure used to make multiple comparisons (we'll be using Tukey's *HSD* post-hoc tests). When you report multiple comparisons, make sure you fully review which conditions differed significantly from one another and which did not. Here's what the report might look like for the cognitive ability data after I modified the scores to introduce significant differences:

Groups of 10 college students apiece took verbal ($M = 35.50$, $SD = 7.46$), quantitative ($M = 48.10$, $SD = 16.41$), and spatial ($M = 61.60$, $SD = 15.81$) tests of cognitive ability, and there was a statistically significant difference in performance across conditions, $F(2, 27) = 8.89$, $p = .001$, $\eta^2 = .40$. A post-hoc comparison of means using Tukey's *HSD* with $\alpha = .05$ revealed

that scores were significantly higher on the spatial test than on the verbal test. Scores on the quantitative test were not significantly different from those on either of the other tests.

## Problems

Each of 24 subjects is randomly assigned to consume either 0, 2, or 4 oz. of alcohol and then take a test on a driving simulator. The dependent variable is the number of errors made while driving (e.g., speeding, tailgating, failure to stop, veering out of the driving lane, crashing into any object). Here are the data:

0 oz. condition: 1, 5, 3, 8, 4, 6, 2, 7

2 oz. condition: 3, 6, 2, 10, 7, 5, 9, 4

4 oz. condition: 6, 8, 4, 13, 9, 5, 10, 11

These data were entered into SPSS like this:

| Alcohol | Errors |
|---|---|
| 1.00 | 1.00 |
| 1.00 | 5.00 |
| 1.00 | 3.00 |
| 1.00 | 8.00 |
| 1.00 | 4.00 |
| 1.00 | 6.00 |
| 1.00 | 2.00 |
| 1.00 | 7.00 |
| 2.00 | 3.00 |
| 2.00 | 6.00 |
| 2.00 | 2.00 |
| 2.00 | 10.00 |
| 2.00 | 7.00 |
| 2.00 | 5.00 |
| 2.00 | 9.00 |
| 2.00 | 4.00 |
| 3.00 | 6.00 |
| 3.00 | 8.00 |
| 3.00 | 4.00 |
| 3.00 | 13.00 |
| 3.00 | 9.00 |
| 3.00 | 5.00 |
| 3.00 | 10.00 |
| 3.00 | 11.00 |

An independent groups ANOVA was performed using the following command:

```
unianova errors by alcohol
/posthoc alcohol (tukey)
/print desc etasq
```

The three tables of output you'd need to examine are shown below:

**Descriptive Statistics**

Dependent Variable: Number of Driving Errors

| Alcohol Condition | Mean | Std. Deviation | N |
|---|---|---|---|
| 0 oz. | 4.5000 | 2.44949 | 8 |
| 2 oz. | 5.7500 | 2.81577 | 8 |
| 4 oz. | 8.2500 | 3.10530 | 8 |
| Total | 6.1667 | 3.11611 | 24 |

**Tests of Between-Subjects Effects**

Dependent Variable: Number of Driving Errors

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 58.333[a] | 2 | 29.167 | 3.712 | .042 | .261 |
| Intercept | 912.667 | 1 | 912.667 | 116.158 | .000 | .847 |
| Alcohol | 58.333 | 2 | 29.167 | 3.712 | .042 | .261 |
| Error | 165.000 | 21 | 7.857 | | | |
| Total | 1136.000 | 24 | | | | |
| Corrected Total | 223.333 | 23 | | | | |

**Homogeneous Subsets**

**Number of Driving Errors**

Tukey HSD[a,b]

| Alcohol Condition | N | Subset 1 | Subset 2 |
|---|---|---|---|
| 0 oz. | 8 | 4.5000 | |
| 2 oz. | 8 | 5.7500 | 5.7500 |
| 4 oz. | 8 | | 8.2500 |
| Sig. | | .651 | .199 |

1. What is the researcher's hypothesis?
2. Why would you perform an ANOVA rather than a series of *t* tests to analyze these data?
3. What are the statistical hypotheses ($H_0$ and $H_1$)?
4. Why don't you need to decide whether to perform a 2-tailed or a 1-tailed test?
5. What are the values of *F*, *df* (there are two *df* values), *p*, and $\eta^2$? Use the SPSS output to find these.
6. What is your statistical decision: Would you reject or retain $H_0$?
7. What is the size of the effect, using $\eta^2$? According to the usual rules of thumb, how would you describe this?
8. Which, if any, pairs of conditions differ statistically significantly from one another? How can you tell?
9. Report the results in APA style. Include the *F* test and, if necessary, post-hoc test results.

* * *

Using the parole data introduced earlier, we can test whether there are differences in scores on the Lifetime Criminality Screening Form (LCSF) across education levels. Subjects were classified into three levels based on how much schooling they completed: less than high school, some high school, high school diploma or further. The data were entered into SPSS (the data file is too large to show here), and the three tables of output you'd need to examine are shown below:

**Descriptive Statistics**

Dependent Variable: Lifestyle Criminality Screening Form

| Educational level | Mean | Std. Deviation | N |
|---|---|---|---|
| < HS | 8.89 | 2.804 | 9 |
| Some HS | 8.00 | 2.625 | 64 |
| HS diploma | 4.83 | 2.940 | 41 |
| Total | 6.93 | 3.164 | 114 |

**Tests of Between-Subjects Effects**

Dependent Variable: Lifestyle Criminality Screening Form

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 288.745[a] | 2 | 144.372 | 19.017 | .000 | .255 |
| Intercept | 3121.086 | 1 | 3121.086 | 411.111 | .000 | .787 |
| educ3 | 288.745 | 2 | 144.372 | 19.017 | .000 | .255 |
| Error | 842.694 | 111 | 7.592 | | | |
| Total | 6606.000 | 114 | | | | |
| Corrected Total | 1131.439 | 113 | | | | |

**Homogeneous Subsets**

**Lifestyle Criminality Screening Form**

Tukey HSD[a,b,c]

| Educational level | N | Subset 1 | Subset 2 |
|---|---|---|---|
| HS diploma | 41 | 4.83 | |
| Some HS | 64 | | 8.00 |
| < HS | 9 | | 8.89 |
| Sig. | | 1.000 | .568 |

10. What is the researcher's hypothesis?
11. What are the statistical hypotheses ($H_0$ and $H_1$)?
12. What are the values of $F$, $df$ (there are two $df$ values), $p$, and $\eta^2$? Use the SPSS output to find these.
13. What is your statistical decision: Would you reject or retain $H_0$?
14. What is the size of the effect, using $\eta^2$? According to the usual rules of thumb, how would you describe this?
15. Which, if any, pairs of conditions differ statistically significantly from one another? How can you tell?
16. Report the results in APA style. Include the $F$ test and, if necessary, post-hoc test results.

* * *

17. Using SPSS, enter the cognitive ability data from this chapter. Follow the instructions in the text for how to organize the data file and enter the command to perform an independent groups ANOVA. Check that your results match what's shown in the text.
18. Using SPSS, enter the data from the first set of problems (on the influence of alcohol on driving performance). Follow the instructions in the text for how to organize the data file and enter the command to run an independent groups ANOVA. Check that your results match what you found earlier.

**Problems 1 – 9 are due at the beginning of class.**

# 13. Related Samples ANOVA

## Overview

The related samples ANOVA extends the related samples *t* test to within-subjects research designs that include more than two conditions. This chapter describes the procedure for performing and reporting the results of the *F* test and, if necessary, a post-hoc test for multiple comparisons.

## The ANOVA Model and the *F* Test

Let's revisit the study from the last chapter. Rather than having 30 students complete either a verbal, quantitative, or spatial test of cognitive ability, let's suppose that 10 students complete all three tests. This is a much better research design because it allows us to remove individual differences from the analysis, which will greatly increase statistical power. To illustrate how this works, we'll use the same 30 test scores. In the summary of the data shown below, the means shown in the right margin provide a measure of the overall cognitive ability of each of the 10 students:

**Original Scores**

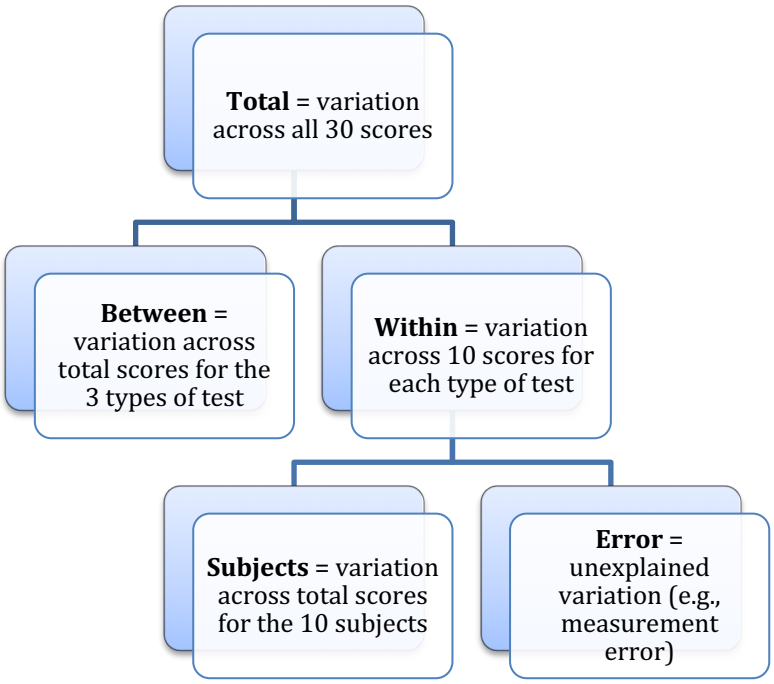
| | Verbal | Quantitative | Spatial | M |
|---|---|---|---|---|
| | 67 | 66 | 65 | 66.00 |
| | 65 | 94 | 82 | 80.33 |
| | 28 | 41 | 42 | 37.00 |
| | 43 | 31 | 55 | 43.00 |
| | 36 | 53 | 42 | 43.67 |
| | 32 | 36 | 42 | 36.67 |
| | 37 | 31 | 32 | 33.33 |
| | 61 | 55 | 82 | 66.00 |
| | 51 | 74 | 66 | 63.67 |
| | 25 | 30 | 48 | 34.33 |
| M | 44.50 | 51.10 | 55.60 | |
| SD | 15.57 | 21.57 | 17.51 | |

We could perform a series of three related samples *t* tests to compare all conditions to one another (i.e., verbal vs. quantitative, verbal vs. spatial, and quantitative vs. spatial), but this would increase the experimentwise Type I error rate well above our desired $\alpha$ level. Instead, we can use a related samples ANOVA to compare scores across all three conditions in a single *F* test that holds $\alpha$ to our desired level. The null and alternative hypotheses can be expressed as follows:

$H_0$: $\mu_1 = \mu_2 = \mu_3$
$H_1$: $\sim(\mu_1 = \mu_2 = \mu_3)$

Testing the null hypothesis involves calculating an *F* ratio. When subjects are tested repeatedly, the total variance on the dependent variable can be split into several sources. The first division is between the systematic source of variance, **between conditions**, and the variance is what remains **within conditions**. Between-conditions variance is due to

differences across conditions. In this case, that would indicate that subjects perform differently on the verbal, quantitative, and spatial tests. Within-conditions variance can be divided further into two components, **subjects** and **error**. Subjects variance refers to individual differences (i.e., people differ in their overall levels of cognitive ability), and error refers to any remaining sources of unexplained variance, such as measurement error (i.e., tests of cognitive ability are not perfectly reliable). Schematically, here's how the related samples ANOVA partitions the variance:




The first level, splitting total variance into between and within, is the same as for an independent groups ANOVA. The second level, splitting the variance within conditions into subjects and error, is only possible for a related samples ANOVA. Because we now have a measure of individual differences—subjects' average scores across conditions—we can remove this from the error variance. In the case of the cognitive ability data, here's a way to represent what it means to remove individual differences from the analysis:

**Individual Differences Removed**

| | Verbal | Quantitative | Spatial | M |
|---|---|---|---|---|
| | 1.00 | 0.00 | -1.00 | 0.00 |
| | -15.33 | 13.67 | 1.67 | 0.00 |
| | -9.00 | 4.00 | 5.00 | 0.00 |
| | 0.00 | -12.00 | 12.00 | 0.00 |
| | -7.67 | 9.33 | -1.67 | 0.00 |
| | -4.67 | -0.67 | 5.33 | 0.00 |
| | 3.67 | -2.33 | -1.33 | 0.00 |
| | -5.00 | -11.00 | 16.00 | 0.00 |
| | -12.67 | 10.33 | 2.33 | 0.00 |
| | -9.33 | -4.33 | 13.67 | 0.00 |
| M | -5.90 | 0.70 | 5.20 | |
| SD | 6.11 | 8.70 | 6.53 | |

This summary of the data was constructed by subtracting each subject's mean score (the value shown in the right margin in the original data) from his or her scores on each of the three tests. This removes individual differences in overall cognitive ability. Each subject now has a mean score of 0, so within each row positive scores are areas of relative strength and negative scores are areas of relative weakness. When the point of the research is to examine differences across conditions, it's a great help to remove as much variation within conditions as possible.

Notice that when you examine means across conditions, the differences remain the same. For example, in the original data the difference between the verbal and spatial conditions is 55.60 – 44.50 = 11.10, and in the transformed data the difference between these conditions is 5.20 – (-5.90) = 11.10. Removing individual differences has no effect on the focal point of the study: Differences across conditions.

What shrinks dramatically, however, is the variability within each condition. In the original data, the *SD*s were 15.57, 21.57, and 17.51 for the three conditions, and the total variance (sum of the squared *SD*s) is 1014.42. In the transformed data, the *SD*s are 6.11, 8.70, and 6.53, for a total variance of only 155.62. This means that about 85% of all the within-condition variance was attributable to individual differences, and only 15% remains as error variance. This demonstrates the statistical advantage of using a within-subjects design and a related samples analysis rather than a between-subjects design and an independent groups analysis. The related samples analysis can achieve much greater statistical power by removing individual differences from the error variance.

The *F* ratio for a related samples ANOVA is calculated as the systematic variance (between conditions) divided by the error variance (error). The critical region for the *F* test is based on the $\alpha$ level (usually .05) and the *df*. For a related samples ANOVA, the *df* are calculated as $k - 1$ (where $k$ is the number of conditions being compared) and $(N - 1) \times (k - 1)$ (where $N$ is the number of subjects). In this case, the *df* would be $3 - 1 = 2$ and $(10 - 1) \times (3 - 1) = 18$. You can get the *df* from computer output.

You can consult a table of *F* values to determine whether the *F* ratio calculated from the data falls in the critical region, but it's easier to obtain the *p* value from computer output and compare this to the $\alpha$ level. As usual, if $p < \alpha$ you reject $H_0$, otherwise you retain $H_0$. For example, the *F* ratio for the cognitive ability data presented earlier is statistically significant ($p = .036$, which is less than $\alpha = .05$). We would reject $H_0$.

It's worth emphasizing that these cognitive ability data are the same scores analyzed in the last chapter. When treated as though they'd come from a between-subjects design, the independent groups ANOVA yielded $F = 0.92$ and $p = .410$. When treated as though they'd come from a within-subjects design, the related samples ANOVA yielded $F = 4.01$ and $p = .036$. This substantial difference is due to the removal of individual differences from the error variance. As a consequence of this, what was formerly a nonsignificant finding has become statistically significant.

## Multiple Comparisons

Whenever we reject $H_0$ and conclude there is some difference between conditions, this gives us license to make multiple comparisons to determine which conditions differ from

one another. To do this, you can use a post-hoc test such as Tukey's *HSD*. Unfortunately, SPSS will not calculate this for you. Therefore, the procedure will be shown here.

The first step is to calculate a threshold value for statistical significance:

$$HSD = q \times \text{sqrt}(MS_{\text{error}} / N)$$

The value of $q$ can be found in a table, such as the one provided in Appendix A. You need to know $k$, the number of conditions being compared, and the $df$ related to sample size (referred to as $df_{\text{error}}$). The values of $MS_{\text{error}}$ and $N$ can be found in the SPSS output. Specifically, $MS_{\text{error}}$ appears in the column labeled "Mean Square" and the row labeled "Error" in the table labeled "Tests of Within-Subjects Effects", and $N$ appears in the table labeled "Descriptive Statistics". Using the table in Appendix A and borrowing from the output provided below in the section on using SPSS, here's what this calculation looks like for the cognitive ability data:

$$HSD = 3.61 \times \text{sqrt}(77.811 / 10) = 10.070$$

The second step is to make all pairwise comparisons of means to determine which differences exceed this threshold:

Verbal vs. quantitative: 51.10 – 44.50 = 6.60, which is < *HSD*, so it's not significant
Quantitative vs. spatial: 55.60 – 51.10 = 4.50, which is < *HSD*, so it's not significant
Verbal vs. spatial: 55.60 – 44.50 = 11.10, which is > *HSD*, so it's significant

## Effect Size

The measure of effect size for a related samples ANOVA is $\eta^2$, which indicates the proportion of variance in the dependent variable that can be explained by the independent variable. This can be obtained from computer output. Cohen's rules of thumb for interpreting the size of $\eta^2$ are that .01 = small, .09 = medium, and .25 = large. For the cognitive ability data, $\eta^2 = .31$, which is a large effect size. You may recall that when the same data were analyzed using an independent groups ANOVA, $\eta^2 = .06$. The effect size increased dramatically after removing individual differences from the error variance.

If you like, you can also report one or more values of Cohen's *d* to indicate the size of pairwise comparisons. This measure would be calculated in the same way as for a related samples *t* test: The numerator is the difference between *M*s for two conditions and the denominator is the pooled *SD* for those conditions.

## Using SPSS

To perform a related samples ANOVA test in SPSS, you first enter your data into separate variables (columns) representing each condition in the study, here labeled "Verbal", "Quantitative", and "Spatial". Here are the data:

| Verbal | Quantitati... | Spatial |
|---|---|---|
| 67.00 | 66.00 | 65.00 |
| 65.00 | 94.00 | 82.00 |
| 28.00 | 41.00 | 42.00 |
| 43.00 | 31.00 | 55.00 |
| 36.00 | 53.00 | 42.00 |
| 32.00 | 36.00 | 42.00 |
| 37.00 | 31.00 | 32.00 |
| 61.00 | 55.00 | 82.00 |
| 51.00 | 74.00 | 66.00 |
| 25.00 | 30.00 | 48.00 |

Next, you use the following command:

```
glm verbal quantitative spatial
/wsfactor tests (3)
/print desc etasq
/wsdesign
```

On the first line, you the variables representing all the conditions to compare (here, "Verbal", "Quantitative", and "Spatial"). On the second line, you provide a label for your independent variable (here, "Tests") and indicate in parentheses how many conditions there are (here, 3). The third line requests descriptive statistics and $\eta^2$ as an effect size measure, and you don't need to change this line at all. The fourth line indicates the design is within-subjects, and you shouldn't change this, either.

SPSS will provide many tables of output, but you can ignore all but two of them. I recommend deleting the ones you don't need because some of them look similar to those you do need and it's easy to mistakenly read and report output from the wrong table. The first table, labeled "Descriptive Statistics", provides the *M* and *SD* for each condition:

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Verbal | 44.5000 | 15.56527 | 10 |
| Quantitative | 51.1000 | 21.57390 | 10 |
| Spatial | 55.6000 | 17.51317 | 10 |

The second table, labeled "Tests of Within-Subjects Effects", provides the *F* value, *df*, *p* value (labeled as "Sig."), and $\eta^2$ (labeled as "Partial Eta Squared"). Use the top row of the table to find the *F* value (here, 4.01), the first *df* value (here, 2), *p* the value (here, .036), and $\eta^2$ (here, .31). Use the first row in the section labeled "Error" to find the second *df* value (here, 18):

**Tests of Within-Subjects Effects**

Measure: MEASURE_1

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| tests | Sphericity Assumed | 623.400 | 2 | 311.700 | 4.006 | .036 | .308 |
| | Greenhouse-Geisser | 623.400 | 1.646 | 378.638 | 4.006 | .047 | .308 |
| | Huynh-Feldt | 623.400 | 1.967 | 316.933 | 4.006 | .037 | .308 |
| | Lower-bound | 623.400 | 1.000 | 623.400 | 4.006 | .076 | .308 |
| Error(tests) | Sphericity Assumed | 1400.600 | 18 | 77.811 | | | |
| | Greenhouse-Geisser | 1400.600 | 14.818 | 94.521 | | | |
| | Huynh-Feldt | 1400.600 | 17.703 | 79.118 | | | |
| | Lower-bound | 1400.600 | 9.000 | 155.622 | | | |

As noted above, SPSS does not perform post-hoc tests for a related samples ANOVA. If you want to use the Tukey's *HSD* procedure, you'll need to do this by hand, as illustrated earlier.

## APA Style

When you retain $H_0$, you can report the results of a related samples ANOVA in a single sentence. You can include the *M* and *SD* for each condition if you like, but that's considered optional for results that are not statistically significant. The last chapter shows what that kind of a report would look like.

When you reject $H_0$, you begin by reporting the results of the *F* test in a single sentence and then follow this with the post-hoc test results. Note that you should not only include the *M* and *SD* for each condition, but also specify the $\alpha$ level (usually .05) and procedure used to make multiple comparisons (we'll be using Tukey's *HSD* post-hoc tests). When you report on multiple comparisons, make sure you fully review which conditions differed significantly from one another and which did not. Here's what the report might look like for the cognitive ability data:

Each of 10 college students took three tests of cognitive ability, and there was a statistically significant difference in performance across tests, $F(2, 18) = 4.01$, $p = .036$, $\eta^2 = .31$. A post-hoc comparison of means using Tukey's *HSD* with $\alpha = .05$ revealed that scores were significantly higher on the spatial test ($M = 55.60$, $SD = 17.51$) than on the verbal test ($M = 44.50$, $SD = 15.57$). Scores on the quantitative test ($M = 51.10$, $SD = 21.57$) were not significantly different from those on either of the other tests.

## Problems

Each of 8 subjects is tested on a driving simulator under three conditions, namely after consuming either 0, 2, or 4 oz. of alcohol. The order of conditions is counterbalanced. The dependent variable is the number of errors made while driving (e.g., speeding, tailgating, failure to stop, veering out of the driving lane, crashing into any object). Here are the data:

| 0 oz. | 2 oz. | 4 oz. |
|---|---|---|
| 1 | 3 | 6 |
| 5 | 6 | 8 |
| 3 | 2 | 4 |
| 8 | 10 | 13 |
| 4 | 7 | 9 |
| 6 | 5 | 5 |
| 2 | 9 | 10 |
| 7 | 4 | 11 |

These data were entered into SPSS like this:

| alc0 | alc2 | alc4 |
|---|---|---|
| 1.00 | 3.00 | 6.00 |
| 5.00 | 6.00 | 8.00 |
| 3.00 | 2.00 | 4.00 |
| 8.00 | 10.00 | 13.00 |
| 4.00 | 7.00 | 9.00 |
| 6.00 | 5.00 | 5.00 |
| 2.00 | 9.00 | 10.00 |
| 7.00 | 4.00 | 11.00 |

A related samples ANOVA was performed using the following command:

```
glm alc0 alc2 alc4
/wsfactor alcohol (3)
/print desc etasq
/wsdesign
```

The two tables of output you'd need to examine are shown below:

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Driving errors with 0 oz. alcohol | 4.5000 | 2.44949 | 8 |
| Driving errors with 2 oz. alcohol | 5.7500 | 2.81577 | 8 |
| Driving errors with 4 oz. alcohol | 8.2500 | 3.10530 | 8 |

**Tests of Within-Subjects Effects**

Measure: MEASURE_1

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| alcohol | Sphericity Assumed | 58.333 | 2 | 29.167 | 8.221 | .004 | .540 |
| | Greenhouse-Geisser | 58.333 | 1.700 | 34.307 | 8.221 | .007 | .540 |
| | Huynh-Feldt | 58.333 | 2.000 | 29.167 | 8.221 | .004 | .540 |
| | Lower-bound | 58.333 | 1.000 | 58.333 | 8.221 | .024 | .540 |
| Error(alcohol) | Sphericity Assumed | 49.667 | 14 | 3.548 | | | |
| | Greenhouse-Geisser | 49.667 | 11.902 | 4.173 | | | |
| | Huynh-Feldt | 49.667 | 14.000 | 3.548 | | | |
| | Lower-bound | 49.667 | 7.000 | 7.095 | | | |

1. What is the researcher's hypothesis?
2. Why would you perform an ANOVA rather than a series of *t* tests to analyze these data?
3. Why would you perform a related samples ANOVA rather than an independent groups ANOVA?
4. What are the statistical hypotheses ($H_0$ and $H_1$)?
5. Why don't you need to decide whether to perform a 2-tailed or a 1-tailed test?
6. What are the values of *F*, *df* (there are two *df* values), *p*, and $\eta^2$? Use the SPSS output to find these.
7. What is your statistical decision: Would you reject or retain $H_0$?
8. What is the size of the effect, using $\eta^2$? According to the usual rules of thumb, how would you describe this?
9. Which, if any, pairs of conditions differ statistically significantly from one another? Perform Tukey's *HSD* using $\alpha = .05$ to answer this question.
10. Report the results in APA style. Include the *F* test and, if necessary, post-hoc test results.
11. How do the results of this analysis compare to those you found when analyzed the same data using an independent groups ANOVA? (Compare and contrast your findings with those from the previous problem set.)

* * *

Using the parole data introduced earlier, we can test whether there are differences in scores on three of the Lifetime Criminality Screening Form (LCSF) subscales: Irresponsibility, Interpersonal Intrusiveness, and Social Rule Breaking. The data were entered into SPSS, and the two tables of output you'd need to examine are shown below:

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| LCSF Irresponsibility | 1.92 | 1.263 | 114 |
| LCSF Interpersonal Intrusiveness | .71 | .929 | 114 |
| LCSF Social Rule Breaking | 1.39 | 1.328 | 114 |

**Tests of Within-Subjects Effects**

Measure: MEASURE_1

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| scale | Sphericity Assumed | 84.000 | 2 | 42.000 | 48.264 | .000 | .299 |
| | Greenhouse-Geisser | 84.000 | 1.959 | 42.880 | 48.264 | .000 | .299 |
| | Huynh-Feldt | 84.000 | 1.993 | 42.144 | 48.264 | .000 | .299 |
| | Lower-bound | 84.000 | 1.000 | 84.000 | 48.264 | .000 | .299 |
| Error(scale) | Sphericity Assumed | 196.667 | 226 | .870 | | | |
| | Greenhouse-Geisser | 196.667 | 221.363 | .888 | | | |
| | Huynh-Feldt | 196.667 | 225.227 | .873 | | | |
| | Lower-bound | 196.667 | 113.000 | 1.740 | | | |

12. What is the researcher's hypothesis?
13. What are the statistical hypotheses ($H_0$ and $H_1$)?
14. What are the values of $F$, $df$ (there are two $df$ values), $p$, and $\eta^2$? Use the SPSS output to find these.
15. What is your statistical decision: Would you reject or retain $H_0$?
16. What is the size of the effect, using $\eta^2$? According to the usual rules of thumb, how would you describe this?
17. Which, if any, pairs of conditions differ statistically significantly from one another? Perform Tukey's *HSD* using $\alpha = .05$ to answer this question.
18. Report the results in APA style. Include the $F$ test and, if necessary, post-hoc test results.

* * *

19. Using SPSS, enter the cognitive ability data from this chapter. Follow the instructions in the text for how to organize the data file and enter the command to perform a related samples ANOVA. Check that your output matches what's shown in the text.
20. Using SPSS, enter the data from the first series of problems (on the influence of alcohol on driving performance). Follow the instructions in the text for how to organize the data file and enter the command to run a related samples ANOVA. Check that your results match what you found earlier.

**Problems 1 – 11 are due at the beginning of class.**

# 14. Factorial ANOVA

## Overview

The previous two chapters extended the *t* tests for independent groups and related samples to research designs with more than two conditions. The final way that we'll extend the comparison of means across conditions is with factorial ANOVA models. By including more than one independent variable, or factor, in the analysis, a factorial ANOVA allows us to analyze the data from an even greater range of research designs. Not only can each factor vary along two or more levels, but also it can be either between-subjects or within-subjects, and these types of factors can be mixed in a single study.

Purely for simplicity, this chapter will examine only factorial ANOVAs with two between-subjects factors. Such an analysis will provide a test for a main effect of each factor as well as a test for the interaction between the two factors. Once you understand the key distinction between main effects and interactions, generalizing this knowledge to the case of ANOVA models with three or more factors, as well as with one or more within-subjects factors, is not too difficult.

## The ANOVA Model and the *F* Tests

To illustrate the use of factorial ANOVAs, consider the following experiment.[34] A total of 24 office workers were given the chance to pay $1 for a lottery ticket for a prize of $25. One half of all tickets came with a random number already assigned, the other half were blank such that their purchasers could choose and write their own ticket numbers. This independent variable was manipulated by random assignment to conditions. In what follows, this factor will be described simply as "ticket", with its two levels being "random number" and "choice of number".

After all tickets were sold, the researcher approached each subject individually to buy back the ticket. He or she was told that someone else wanted to enter the lottery but there were no more tickets, so the researcher would pay what it takes to buy back this ticket to offer it to the newcomer. The dependent variable in this experiment was how much each subject charged to sell back his or her ticket, described as "price".

Tickets were purchased from subjects at one of three different times: (1) immediately after originally selling the ticket, (2) the next day, or (3) just before drawing the winning lottery number at the end of the week. This independent variable was also manipulated by random assignment to conditions. In what follows, this factor will be described as "time", with its three levels being "immediate", "next day", and "before drawing".

[34] This fictional study is based on work done by Ellen Langer in the 1970s on the "illusion of control" phenomenon by which people misunderstand an outcome determined purely by chance (e.g., a lottery) as one that can be influenced with some skill (e.g., choosing numbers provides better odds of winning).

In sum, this is a 2 (ticket) × 3 (time) design. There are a total of 6 cells in the fully between-subjects design, with $n = 4$ scores within each cell. If we label ticket as factor A and time as factor B,[35] the factorial ANOVA will provide the following three *F* tests:

1. Main effect for ticket (A). Ignoring time, does price differ across random number and choice of number conditions?
2. Main effect for time (B). Ignoring ticket, does price differ across immediate, next day, and before drawing conditions?
3. Interaction between ticket and time (A × B). Is price influenced by the combination of ticket type and time of resale?

We can use a factorial ANOVA to obtain these three *F* tests. For simplicity, we can express the null and alternative hypotheses as the absence ($H_0$) or presence ($H_1$) of an effect. In other words, the three null hypotheses would be (1) no main effect for factor A, (2) no main effect for factor B, and (3) no interaction between factors A and B.

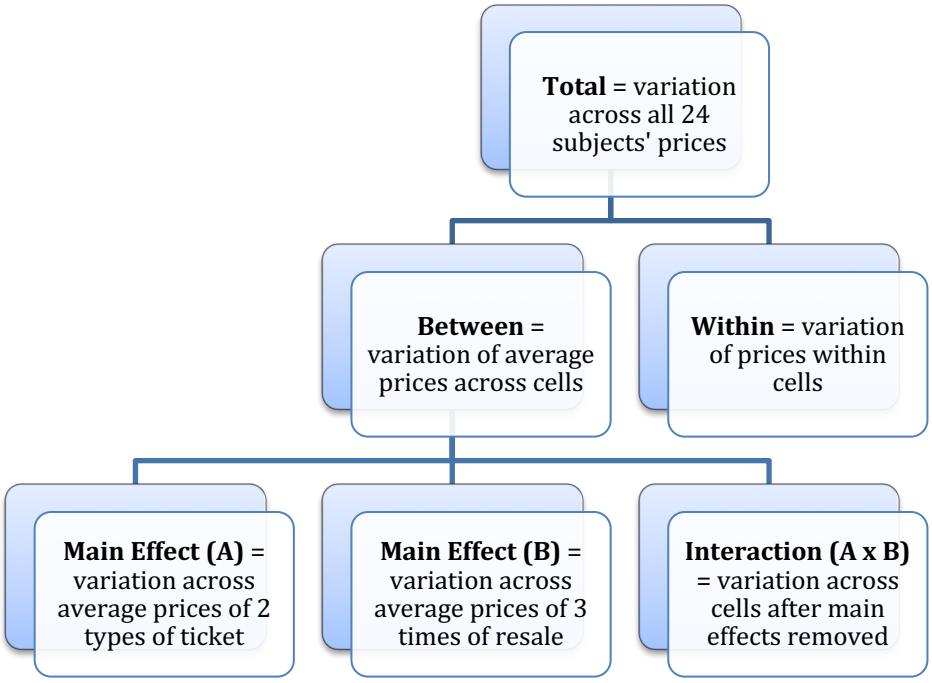
It's critical to understand that these are three independent tests. In other words, there are a total of eight possible outcomes of this analysis. The following table summarizes every possibility:

| | Main Effect (A) | Main Effect (B) | Interaction (A × B) |
|---|---|---|---|
| 1 | No | No | No |
| 2 | Yes | No | No |
| 3 | No | Yes | No |
| 4 | Yes | Yes | No |
| 5 | No | No | Yes |
| 6 | Yes | No | Yes |
| 7 | No | Yes | Yes |
| 8 | Yes | Yes | Yes |

Thus, interpreting and reporting the results of a factorial ANOVA requires that three effects be considered. A careful approach is required to ensure that all three effects are described accurately.

Testing each null hypothesis involves calculating an *F* ratio. As with other ANOVA models, the total variance on the dependent variable can be split into several sources. The first division is between systematic sources of variance, **between groups**, and the error variance that remains **within groups**. This split is the same as the for an independent groups ANOVA. Between-groups variance is due to differences across conditions (whether main effects or an interaction), and within-groups variance is due to unexplained sources of variation within groups, such as individual differences and measurement error. The second division splits the between groups variance into three sources: **main effect for factor A**, **main effect for factor B**, and **interaction between factors A and B**. Schematically, here's how the factorial ANOVA partitions the variance for the lottery ticket study:

[35] It would make no difference for anything discussed in this chapter if we reverse this and label time as factor A and ticket as factor B.




For the three effects tested in this factorial ANOVA model, the numerator corresponds to a systematic source of variance—main effect for A, main effect for B, or interaction between A and B. The error (within-groups) variance serves as the denominator for each of these *F* ratios. The critical region for each *F* test is based on the $\alpha$ level (usually .05) and the *df*. There are two *df* for any *F* test, one relating to the number of conditions and one relating to sample size. The *df* relating to sample size is calculated as $N - (a \times b)$ for all three *F* tests (where *N* is the number of subjects, *a* is the number of levels for factor A, and *b* is the number of levels for factor B). For main effect A, the first *df* is calculated as $a - 1$. For main effect B, the first *df* is calculated as $b - 1$. For the interaction, the first *df* is calculated as $(a - 1) \times (b - 1)$. You can get the *df* from computer output.

You can consult a table of *F* values to determine whether each *F* ratio falls in the critical region, but it's easier to obtain the *p* values from computer output and compare these to the $\alpha$ level. As usual, if $p < \alpha$ you reject $H_0$, otherwise you retain $H_0$. For example, the *F* ratios, *df*, and *p* values for the lottery ticket study are as follows:

1. Main effect for ticket: $F(1, 18) = 10.08$, $p = .005$
2. Main effect for time: $F(2, 18) = 29.31$, $p < .001$
3. Interaction between ticket and time: $F(2, 18) = 4.65$, $p = .024$

In this case, all three null hypotheses would be rejected because each *p* value is less than $\alpha = .05$.

## Effect Size

An appropriate measure of effect size to accompany the results of any *F* test is $\eta^2$. The usual rules of thumb (.01 = small, .09 = medium, .25 = large) apply when this is used for the test of a main effect or an interaction in a factorial ANOVA. For the lottery ticket study, $\eta^2 =$

.36 for the ticket main effect, .76 for the time main effect, and .34 for the interaction between ticket and time. Each of these is a very large effect.[36]

## Procedural Overview

Because a factorial ANOVA provides multiple *F* tests, it's important to proceed carefully and systematically when performing, interpreting, and reporting the results. Here's a general procedure to follow.

First, begin with one of the tests for a main effect. If that *F* test is statistically significant, examine the means to interpret the results. Computer output will provide the relevant means, but you have to make sure that you're using the **marginal means** rather than the **cell means**. Cell means are calculated using all scores within each cell of the design, and marginal means are calculated by collapsing across cells. Here's a table of means for the lottery ticket study:

| | | Time (Factor B) | | | |
|---|---|---|---|---|---|
| | | Immediate | Next Day | Before Drawing | |
| Ticket (Factor A) | Random Number | $3.75 | $3.38 | $5.00 | $4.04 |
| | Choice of Number | $4.88 | $3.12 | $6.88 | $4.96 |
| | | $4.31 | $3.25 | $5.94 | |

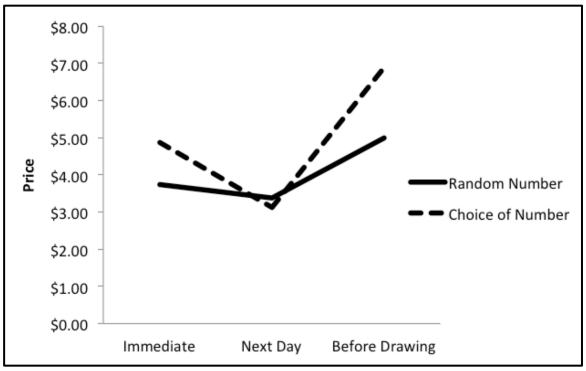
The cells means are the 6 values corresponding to the cells in the design, indicated with dark borders. The marginal means are the values outside the 2 × 3 table, indicated with light borders in the margins. In this case, factor A is ticket, and the *F* ratio is statistically significant. Because there are only two levels, no post-hoc test is required to compare the means. The marginal means are $4.04 and $4.96, and they show us that subjects charged more to sell back tickets when they had chosen the number than when they were assigned a random number. This is a classic finding, demonstrating an "illusion of control". Choosing your own number doesn't objectively affect the probability of winning a lottery, but subjectively it seems to; people value tickets more when they choose their own numbers. This completes the first step, evaluating one of the main effects.

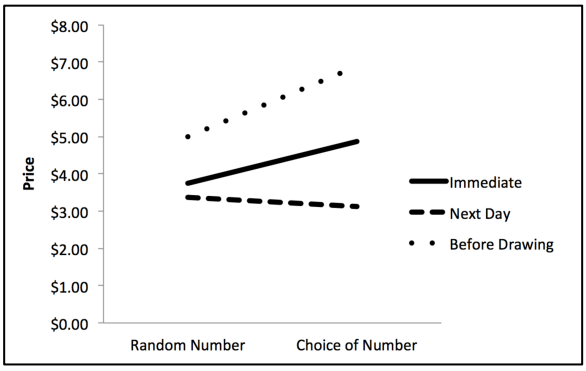
Second, repeat this procedure for the other main effect. In this case, factor B is time, and the *F* ratio is statistically significant. Because there are more than two levels, a post-hoc test is required to compare the means. Using Tukey's *HSD* with $\alpha = .05$ reveals that all three of the conditions differ significantly from one another. People charged a modest amount to resell a ticket immediately ($4.31), a lower amount the next day ($3.25), and a much higher amount before the drawing ($5.94). The perceived value of the ticket varied substantially over time.

---

[36] The reason why these are unrealistically large effects is that I created a very small sample of data (*N* = 24 for a study with 6 cells, meaning *n* = 4 within each cell) but wanted to illustrate how to detect and interpret statistically significant effects. The only way for effects to be statistically significant with such a tiny sample size is if they are very large.

Third, after evaluating both of the tests for main effects, move on to the test for an interaction. This is a higher-level effect and should always be considered after the lower-level main effects.[37] In this case, the interaction between ticket and time is statistically significant. To interpret an interaction, it's helpful to plot a line graph using the cell means. You construct this graph such that the dependent variable (here, price) is on the *y* axis, the levels of one of the factors appear along the *x* axis, and the levels of the other factor are plotted using separate lines. It makes no difference which factor you place on the *x* axis and which is plotted with separate lines, but it is important to label your graph fully so that the factors can be distinguished. Once you have the axes labeled, you carefully plot each of the cell means to form the lines. For the lottery ticket study, the line graph could be plotted as either of these two versions:




Though these look different, they reveal the same pattern: The effect of one factor depends on the levels of the other. That's what an interaction effect is all about. When you graph the cell means, an interaction effect is present if the resulting lines are not parallel. Of course, random sampling error will cause the lines' slopes to differ a bit, just by chance. An *F* test for an interaction effect determines whether the slopes differ enough to be statistically significant.

In this case, the *F* test was statistically significant. What we see in both versions of the line graph is that the time of resale made less difference in price when tickets had random numbers on them than when subjects had chosen their own numbers. For the graph on the left, the solid line (for random number) remains flatter than the dashed line (for choice of number). For the graph on the right, the three lines begin at similar values on the left (for choice of number) but diverge as they move to the right (for random number). Whichever graph you examine, you'd arrive at the same conclusion: Time of resale made more of a difference for one type of ticket (choice of number) than the other (random number).

As a final reminder, this is just one possible pattern of results. In this analysis, all three effects—both main effects and the interaction—were statistically significant and needed to be described. There are seven other possible patterns of results in which some or all of the three effects are not statistically significant. Factorial ANOVA results differ from one study to the next much more so than *z* or *t* test results. That's why it's so important to follow a step-by-step procedure—progressing through the three separate *F* tests in an orderly, careful way—when doing a factorial ANOVA.

---

[37] Similarly, if you have more than two factors in the design, you'd begin with main effects and then deal with interactions in increasing order of complexity (i.e., 2-way interactions before 3-way interactions).

## Using SPSS

To perform a factorial ANOVA with two between-subjects factors on SPSS, you first enter your data into three separate variables (columns), here labeled "Ticket" (coded as 1 = random number, 2 = choice of number), "Time" (coded as 1 = immediate, 2 = next day, 3 = before drawing), and "Price". Note that you have to create two variables that indicate group membership for each subject, one for each factor in the design, and the dependent variable is placed in a separate column for all subjects. The full data set didn't fit onto the screen; here's what all but the final row of the data file looks like:

| Ticket | Time | Price |
|---|---|---|
| 1.00 | 1.00 | 4.50 |
| 1.00 | 1.00 | 2.50 |
| 1.00 | 1.00 | 3.75 |
| 1.00 | 1.00 | 4.25 |
| 2.00 | 1.00 | 3.75 |
| 2.00 | 1.00 | 5.25 |
| 2.00 | 1.00 | 5.00 |
| 2.00 | 1.00 | 5.50 |
| 1.00 | 2.00 | 3.00 |
| 1.00 | 2.00 | 4.25 |
| 1.00 | 2.00 | 3.75 |
| 1.00 | 2.00 | 2.50 |
| 2.00 | 2.00 | 3.75 |
| 2.00 | 2.00 | 2.75 |
| 2.00 | 2.00 | 3.50 |
| 2.00 | 2.00 | 2.50 |
| 1.00 | 3.00 | 5.50 |
| 1.00 | 3.00 | 5.75 |
| 1.00 | 3.00 | 4.50 |
| 1.00 | 3.00 | 4.25 |
| 2.00 | 3.00 | 6.50 |
| 2.00 | 3.00 | 7.00 |
| 2.00 | 3.00 | 7.25 |

Next, you use the following command:

```
unianova price by ticket time
/posthoc ticket time (tukey)
/print desc etasq
```

This is the same kind of command used for an independent groups ANOVA. The only difference here is that rather than listing a single group membership variable on the first and second lines, you list two of them (here, "Ticket" and "Time"). You can list these group membership variables in either order. The second line requests Tukey's *HSD* as a post-hoc test for both factors, which you may or may not need. SPSS will give you a warning (don't be alarmed by it!) to indicate if post-hoc tests are not performed for one or both factors in the event that they vary across fewer than three levels. The third line requests descriptive statistics and $\eta^2$ as an effect size measure, and you don't need to change this line at all.

SPSS will provide many tables of output, but you can ignore some of them. I recommend deleting the ones you don't need because some of them look similar to those you do need and it's easy to mistakenly read and report output from the wrong table. The first table

you'll need is labeled "Descriptive Statistics", and it provides both the cell means and the marginal means (with their corresponding standard deviations) for interpreting any statistically significant main effects or interactions:

**Descriptive Statistics**

Dependent Variable: Resale Price

| Type of Ticket | Time of Resale | Mean | Std. Deviation | N |
|---|---|---|---|---|
| Random Number | Immediate | 3.7500 | .88976 | 4 |
| | Next Day | 3.3750 | .77728 | 4 |
| | Before Drawing | 5.0000 | .73598 | 4 |
| | Total | 4.0417 | 1.02710 | 12 |
| Choice of Number | Immediate | 4.8750 | .77728 | 4 |
| | Next Day | 3.1250 | .59512 | 4 |
| | Before Drawing | 6.8750 | .32275 | 4 |
| | Total | 4.9583 | 1.68831 | 12 |
| Total | Immediate | 4.3125 | .97970 | 8 |
| | Next Day | 3.2500 | .65465 | 8 |
| | Before Drawing | 5.9375 | 1.13192 | 8 |
| | Total | 4.5000 | 1.44463 | 24 |

In this first table, cell means are identified by looking for the rows that correspond to combinations of levels on both factors (e.g., "Random Number" and "Immediate" is a cell mean). Marginal means are identified by looking for rows that contain "Total" in one of the first two columns (e.g., "Random Number" and "Total" is a marginal mean). The bottom row of the table ("Total" and "Total") can be ignored. You might find it helpful to double-check that you can read this table to produce the one containing only the cell and marginal means shown earlier in this chapter.

The second table you'll need, labeled "Tests of Between-Subjects Effects", provides the *F* value, *df*, *p* value (labeled as "Sig."), and $\eta^2$ (labeled as "Partial Eta Squared") for each test:

**Tests of Between-Subjects Effects**

Dependent Variable: Resale Price

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 39.000[a] | 5 | 7.800 | 15.600 | .000 | .813 |
| Intercept | 486.000 | 1 | 486.000 | 972.000 | .000 | .982 |
| Ticket | 5.042 | 1 | 5.042 | 10.083 | .005 | .359 |
| Time | 29.313 | 2 | 14.656 | 29.313 | .000 | .765 |
| Ticket * Time | 4.646 | 2 | 2.323 | 4.646 | .024 | .340 |
| Error | 9.000 | 18 | .500 | | | |
| Total | 534.000 | 24 | | | | |
| Corrected Total | 48.000 | 23 | | | | |

Use the rows labeled with each of your group membership variables (here, "Ticket" and "Time") to find the tests of main effects, and the row labeled with both group membership variables (here, "Ticket * Time") to find the test of the interaction. Note that the first *df* value for each test is provided on the row indicated above, and the second *df* for all three tests is provided on the row labeled "Error". Here's a summary of the results for these three *F* tests that you can use to double-check that understand where to find all the values:

- Main effect for ticket: $F(1, 18) = 10.08$, $p = .005$, $\eta^2 = .36$
- Main effect for time: $F(2, 18) = 29.31$, $p < .001$, $\eta^2 = .76$
- Interaction between ticket and time: $F(2, 18) = 4.65$, $p = .024$, $\eta^2 = .34$

If you need post-hoc test results, they will be provided in tables labeled "Homogeneous Subsets". In this case, there is a statistically significant main effect for time, which varies

across more than two levels, so you'd need the results of a post-hoc test to determine which conditions differed from one another. The table is shown below, and the results indicate that all three conditions differed significantly from one another. Instructions for how to read and interpret these results are the same as for an independent groups ANOVA.

**Homogeneous Subsets**

**Resale Price**

Tukey HSD[a,b]

| Time of Resale | N | Subset 1 | Subset 2 | Subset 3 |
|---|---|---|---|---|
| Next Day | 8 | 3.2500 | | |
| Immediate | 8 | | 4.3125 | |
| Before Drawing | 8 | | | 5.9375 |
| Sig. | | 1.000 | 1.000 | 1.000 |

## APA Style

As the procedural overview explained, you should begin by describing the results of one test for a main effect, then proceed to another test for a main effect, and finally to the test for an interaction. If an effect is not statistically significant, that's easy to state in a single sentence. If an effect is statistically significant, you need to describe the pattern of results. For a main effect, this entails reporting the *M*s and *SD*s for each level and, if there are more than two levels, reporting the results of a post-hoc test to indicate which levels differed significantly from one another. For an interaction effect, you'd need to explain how the effect of one factor varies across levels of the other. It can be very helpful to provide a graph to illustrate an interaction, but you still have to explain the pattern of results in your text. Here's what the report might look like for the lottery ticket study:

There was a statistically significant main effect for type of ticket, $F(1, 18) = 10.08$, $p = .005$, $\eta^2 = .36$. Those who chose their own ticket number charged more to sell it back ($M = 4.96$, $SD = 1.69$) than those who were assigned a random ticket number ($M = 4.04$, $SD = 1.03$). There was a statistically significant main effect for time of resale, $F(2, 18) = 29.31$, $p < .001$, $\eta^2 = .76$. A post-hoc comparison of means using Tukey's *HSD* with $\alpha = .05$ revealed that prices differed significantly across all three conditions (immediate $M = 4.31$, $SD = 0.98$; next day $M = 3.25$, $SD = 0.65$; before drawing $M = 5.94$, $SD = 1.13$). There was a statistically significant interaction between type of ticket and time of resale, $F(2, 18) = 4.65$, $p = .024$, $\eta^2 = .34$. Variation in price across time of resale was much greater when subjects chose their own ticket numbers than when they were given random ticket numbers.

Notice that the report proceeds from one main effect to another, and then to the interaction. In an actual research report, it would be a good idea to provide (and cite) a line graph that illustrates the interaction effect; this was shown earlier and not repeated here.

## Problems

Below is SPSS output for a factorial ANOVA examining the annual income (in thousands of dollars) for 160 doctors. Factor A is area of medical practice (pediatrician, general practitioner, and surgeon), and Factor B is gender (female, male).

**Descriptive Statistics**

Dependent Variable: Annual Income (Thousands $)

| Area of Medical Practice | Gender | Mean | Std. Deviation | N |
|---|---|---|---|---|
| Pediatrician | Female | 216.3122 | 60.91474 | 45 |
| | Male | 184.7422 | 58.42838 | 17 |
| | Total | 207.6559 | 61.42919 | 62 |
| General Practitioner | Female | 278.8243 | 70.90998 | 18 |
| | Male | 253.3685 | 46.70289 | 14 |
| | Total | 267.6874 | 61.94115 | 32 |
| Surgeon | Female | 346.4965 | 62.72981 | 17 |
| | Male | 309.5529 | 65.50522 | 49 |
| | Total | 319.0687 | 66.35001 | 66 |
| Total | Female | 258.0416 | 81.97363 | 80 |
| | Male | 273.1984 | 78.93471 | 80 |
| | Total | 265.6200 | 80.57451 | 160 |

**Tests of Between-Subjects Effects**

Dependent Variable: Annual Income (Thousands $)

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 431619.03[a] | 5 | 86323.805 | 22.133 | .000 | .418 |
| Intercept | 8792905.77 | 1 | 8792905.77 | 2254.408 | .000 | .936 |
| Area | 405715.640 | 2 | 202857.820 | 52.011 | .000 | .403 |
| Gender | 30739.432 | 1 | 30739.432 | 7.881 | .006 | .049 |
| Area * Gender | 647.050 | 2 | 323.525 | .083 | .920 | .001 |
| Error | 600648.867 | 154 | 3900.317 | | | |
| Total | 12320903.9 | 160 | | | | |
| Corrected Total | 1032267.89 | 159 | | | | |

**Homogeneous Subsets**

**Annual Income (Thousands $)**

Tukey HSD[a,b,c]

| Area of Medical Practice | N | Subset 1 | Subset 2 | Subset 3 |
|---|---|---|---|---|
| Pediatrician | 62 | 207.6559 | | |
| General Practitioner | 32 | | 267.6874 | |
| Surgeon | 66 | | | 319.0687 |
| Sig. | | 1.000 | 1.000 | 1.000 |

1. Using the descriptive statistics, create your own table that includes only the cell means and marginal means. Organize the table like the one shown in the chapter. Begin by creating a table with rows for the levels of factor A and columns for the levels of factor B. Within this table, identify and list the cell means. Then, identify the marginal means that correspond to totals for each row and each column.

2. Using your table from #1, plot a line graph of the cell means. Make sure that you label the axes and the lines clearly.
3. Why was a factorial ANOVA performed to analyze these data rather than an independent groups ANOVA or a related samples ANOVA?
4. What are the factors in this study? For each factor, list its levels.
5. How many *F* tests were performed in this ANOVA? What did each one test?
6. For area of medical practice, what are the values of *F*, *df*, *p*, and $\eta^2$? Is this a statistically significant effect? How would you characterize the size of the effect?
7. Begin with the descriptive statistics to interpret the results for this factor. Why do you also need to consult the results of a post-hoc test? Report the results in APA style. (This should take just two sentences.)
8. For gender, what are the values of *F*, *df*, *p*, and $\eta^2$? Is this a statistically significant effect? How would you characterize the size of the effect?
9. Examine the descriptive statistics to interpret the results for this factor. Why don't you need a post-hoc test? Report the results in APA style. (This should take just one or two sentences.)
10. For the interaction between area of medical practice and gender, what are the values of *F*, *df*, *p*, and $\eta^2$? Is this a statistically significant effect? How would you characterize the size of the effect?
11. Report the results for the test of an interaction effect in APA style. (This should take just one sentence.)
12. Revisit the conclusions you reached for the main effect for gender, and then look carefully at your table and graph. Do you see an apparent contradiction in the findings? How can this be resolved? (Hint: Consider the cell sizes in the descriptive statistics.)

* * *

For this series of problems, Dr. Flurpple compares testing procedures. Students in four sections of a statistics course take a common final exam under different conditions; each section is told in advance what its conditions will be. Two sections of the class take it as an open-book exam, whereas the other two sections do not. In addition, one of the open-book sections is given an untimed test, whereas the other is given a time limit; the same goes for the two closed-book sections. The average scores on the exam for the four sections are as follows: open-book, untimed = 90; open-book, timed = 50; closed-book, untimed = 70; closed-book, timed = 60.

13. What is the dependent variable?
14. What are the factors in the design? For each, state whether it's a between- or within-subjects factor and list its levels
15. Construct a table showing the cell means and marginal means. (For simplicity, assume equal cell sizes when calculating marginal means.)
16. Using your table of means, plot and fully label a line graph.

17. Based on the table in #15, does it appear that there are any main effects? (Don't worry about statistical significance; if you observe any difference between means, assume it's large enough to be statistically significant.)
18. Based on the graph in #16, does it appear that there is an interaction? (Again, set aside the question of statistical significance; if the lines are not parallel, assume the effect is large enough to be statistically significant.)
19. In plain English (not APA style), write an interpretation of the results.

* * *

For this series of problems, Dr. Flurpple wonders whether student achievement can be increased through challenging out-of-class assignments. In addition, Dr. F wonders whether challenges may differentially affect the achievement of low and high aptitude students. To test these hypotheses, students in a statistics course complete both highly challenging assignments (which required answers to conceptual questions) and relatively easy assignments (which required conceptually simple solutions to problems); all students complete both assignments. Early in the semester, each student's aptitude is measured using a brief test of logical reasoning and students are then classified into high and low aptitude groups. The average scores for the high aptitude students are 90 on the challenging assignment and 80 on the easy assignment; for the low aptitude students, averages are 70 on the challenging assignment and 80 on the easy assignment.

20. What is the dependent variable?
21. What are the factors in the design? For each, state whether it's a between- or within-subjects factor and list its levels
22. Construct a table showing the cell means and marginal means. (For simplicity, assume equal cell sizes when calculating marginal means.)
23. Using your table of means, plot and fully label a line graph.
24. Based on the table in #22, does it appear that there are any main effects? (Don't worry about statistical significance; if you observe any difference between means, assume it's large enough to be statistically significant.)
25. Based on the graph in #23, does it appear that there is an interaction? (Again, set aside the question of statistical significance; if the lines are not parallel, assume the effect is large enough to be statistically significant.)
26. In plain English (not APA style), write an interpretation of the results.

* * *

27. Using SPSS, enter the lottery ticket data from this chapter. The screen shot omitted the final row, which contained the values 2, 3, and 6.75. Follow the instructions in the text for how to organize the data file and enter the command to perform a factorial ANOVA. Check that your output matches what's shown in the text.

**Problems 1 – 12 are due at the beginning of class.**

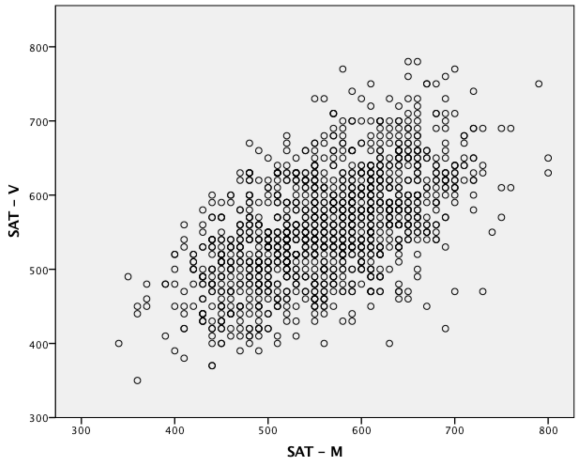
# 15. Scatterplots and Correlation

## Overview

All of the data-analytic procedures introduced so far have compared means across conditions. Correlational analysis is different: It assesses the strength of the relationship between two variables. In this chapter, we'll examine the standard type of correlation coefficient as well as a few variations that are used fairly often.

## Scatterplot

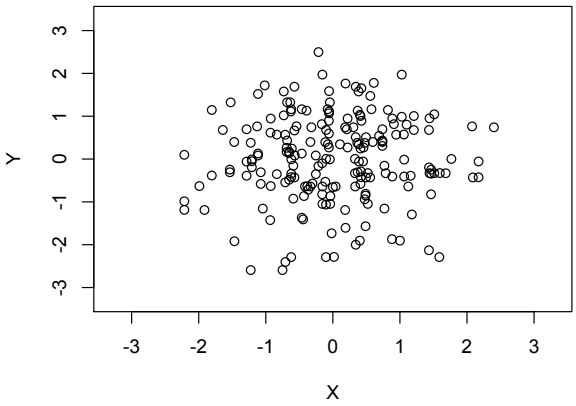
Before calculating a correlation coefficient, it's important to inspect a graph that shows the relationship between scores on the two variables. Constructing a scatterplot is simple. All you need to do is label the *x* and *y* axes according to the *X* and *Y* variables in your analysis, and then plot one data point for each case in the data set. The location of each point is determined by that case's scores on *X* and *Y*. Here's an example of a scatterplot between the SAT Math and Verbal scores for 1,313 students at a private college:



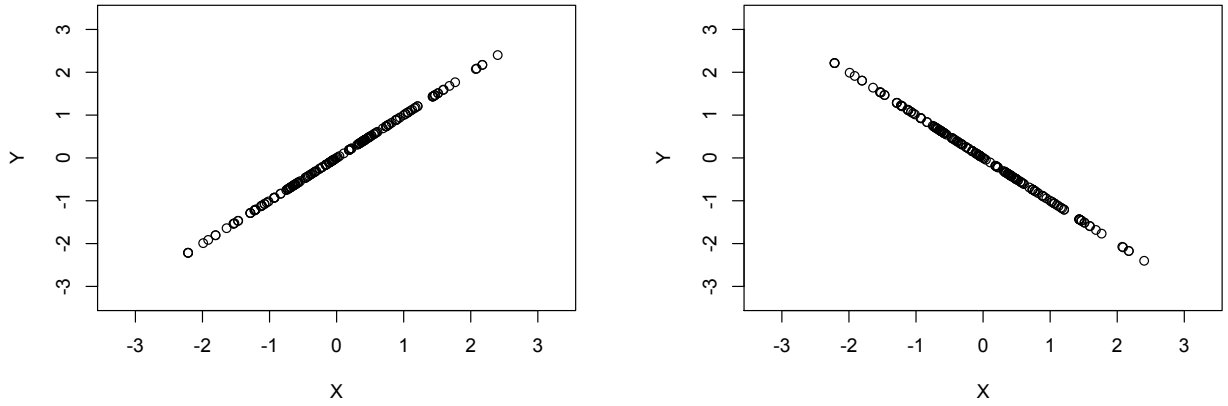

Even though there are many tied scores (e.g., multiple students with identical SAT Math and Verbal scores), this scatterplot still clearly shows us that as SAT Math scores increase, so do SAT Verbal scores. Not only is the association easy to see, but it's also apparent that a straight line could be fit to these data fairly well.

## Correlation Coefficient

Whereas the scatterplot displays the relationship between two variables, the correlation coefficient summarizes this in a single number. Correlations can range from .00 to 1.00 in absolute value. The size of the correlation is an index of how well a scatterplot can be fit by a straight line. To the extent that the points cluster tightly around a line, the correlation will be large. At one extreme, a correlation of .00 means there's no association between the variables, that they're scattered around fairly randomly throughout a cloud:




At the other extreme, a correlation of 1.00 (or -1.00) means there's a perfect relationship. In other words, all of the data points would lie directly on a line. Here are scatterplots depicting correlations of 1.00 (left) and -1.00 (right):

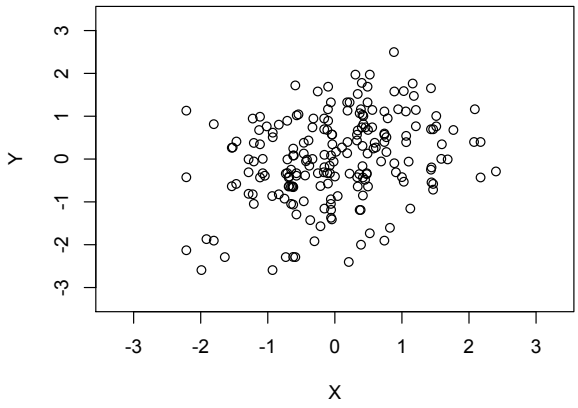



As you can see, if the sign of the correlation is positive, scores on *Y* increase along with scores on *X*. If the sign is negative, scores on *Y* decrease as scores on *X* increase; this is also known as an inverse relationship. Note that which variable you treat as *X*, and which as *Y*, will not affect the sign or the size of a correlation. The scatterplot will look different if you swap the variables, but the direction and strength of the relationship remain the same.

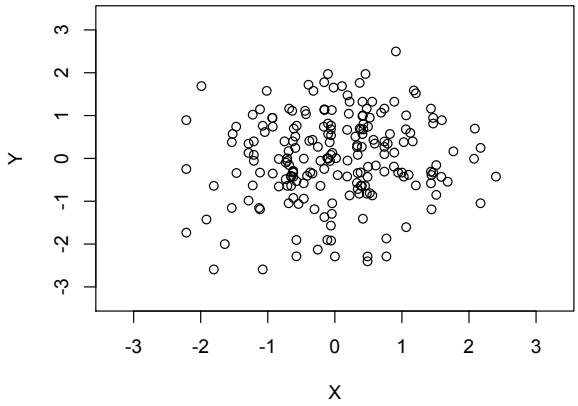
For real data, you seldom observe a correlation as small as .00 or as large as 1.00. Instead, you usually get a value someplace in between. For the SAT data plotted above, the correlation is .56. Unlike most other statistics (e.g., *z*, *t*, or *F*), the correlation coefficient is its own measure of effect size. Cohen's rules of thumb are as follows:

.10 = small
.30 = medium
.50 = large

Thus, the SAT sections correlate with one another at a level that would be described as a large effect. Next, here's a scatterplot illustrating a medium effect (.30). The points cluster somewhat near a line, but there remains a lot of variability around that line:




Finally, here's a scatterplot illustrating a small effect (.10), which is hard to distinguish visually from a correlation of .00:




Examples of what are considered to be small, medium, and large correlations are provided in part so that you can visualize the strength of the relationship between two variables when only the correlation coefficient is provided. Another reason for showing these is to emphasize that even a so-called "large effect" is very far from a perfect correlation. If scores on *X* were used to predict scores on *Y*, for example, these predictions would be much better than chance-level guessing but far from perfectly accurate.

## Coefficient of Determination

The correlation coefficient is often squared to express the strength of the relationship between variables. The squared correlation is called the **coefficient of determination**, and it represents the same thing as $\eta^2$, the effect size measure used with *F* tests: The proportion of variance in one variable that can be explained by the other variable. If you square the rules of thumb for the correlation, you get the rules of thumb for interpreting the size of the coefficient of determination (or $\eta^2$):

$.10^2 = .01$ = small
$.30^2 = .09$ = medium
$.50^2 = .25$ = large

For example, the correlation between SAT sections of .56 yields a coefficient of determination of $.56^2 = .31$. This means that scores on one section explain .31 (or 31%) of

the variation in scores on the other section, which of course also means that the other .69 (69%) of the variation is unexplained, or due to other factors. That's another way of keeping the magnitude of effects in perspective: Even with a correlation this large, a majority of the variation in *Y* remains unexplained by variation in *X*.

## Types of Correlation

There are many types of correlation coefficient, each designed for use with different kinds of data. The four most frequently used types of correlation are described here.

### Pearson Product-Moment Correlation

By far, the most common type of correlation coefficient is the **Pearson product-moment correlation**. This is used whenever both *X* and *Y* are measured using interval or ratio scales, and it's symbolized as *r* in APA style. Because this type of correlation is so popular, its full name is seldom used. You can safely assume that someone means the Pearson product-moment correlation coefficient unless they specify otherwise. The correlation between SAT sections is an example of this kind.

### Spearman Rank-Order Correlation

When the *X* and *Y* variables are measured using ordinal scales (i.e., ranked data), you'd use a **Spearman rank-order correlation**. This is symbolized as $r_S$ in APA style; the subscript of a capital "S" indicates it's a Spearman correlation. Whether the data were collected as ranks or quantitative data were subsequently converted to ranks, the Spearman correlation is used to assess the strength of relationship between *X* and *Y*. For example, if you record the order that students complete an exam (1 = first, 2 = second, …) and their ranked scores on the exam (1 = highest score, 2 = next highest, …), you'd use $r_S$ to assess the relationship between these variables.

### Point-Biserial Correlation

When one variable is measured using an interval or ratio scale and the other is dichotomous—meaning that it can only take two values (e.g., correct/incorrect, high/low, true/false, male/female)—you'd use a **point-biserial correlation**. The two values must be coded numerically, but the choice of codes will not affect the size of the correlation.[38] This is symbolized as $r_{pb}$ in APA style; the subscript of lowercase "pb" indicates it's a point-biserial correlation.

Though this might not be obvious at first glance, using a point-biserial correlation to analyze data is equivalent to using an independent groups *t* test. For example, asking the question of whether gender (a dichotomous variable) correlates with self-esteem (a quantitative variable) poses the same fundamental question as asking whether there is a difference in self-esteem by gender. The two statistics ($r_{pb}$ and *t*) have the same *df* ($N - 2$) and there's a 1:1 relationship between their values: $r_{pb} = \text{sqrt}(t^2 / (t^2 + df))$. You can choose either of these analyses and you'll obtain the same *p* value.

For example, an independent groups *t* test performed in the parole data shows that there is a statistically significant difference in Lifestyle Criminality Screening Form (LCSF)

[38] The sign of the point-biserial correlation will change if you reverse the coding, but its size will not change.

scores between those who have been arrested and those who have not: $t(112) = -2.08$, $p = .040$. A correlational analysis leads to the same conclusion, that LCSF scores are correlated with arrest status: $r_{pb}(112) = .19$, $p = .040$. As shown above, the correlation can be calculated directly from the $t$ test results: $r_{pb} = \text{sqrt}((-2.08)^2 / (-2.08)^2 + 112) = .19$.

Whether to use $r_{pb}$ or $t$ is a matter of preference or, in some cases, consistency with other analyses in the study. For example, if you've already done a series of correlations for other variables, it might be simplest for your audience if you use $r_{pb}$ rather than $t$.

### Phi Coefficient

When both variables are dichotomous, you'd use a **phi coefficient**. This is symbolized using the Greek letter ϕ in APA style.[39] For example, if you're correlating gender with self-esteem, but the self-esteem scores have been categorized as high vs. low, you'd use ϕ.

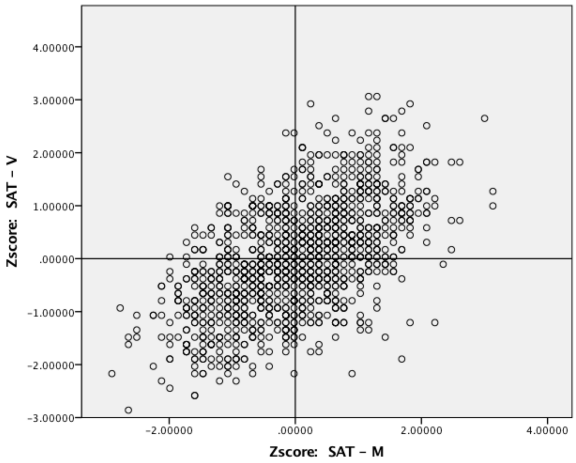
## Calculating Correlations and Testing Hypotheses

There's a formula that can be used to calculate all of the correlations listed above, and there are specialized formulas that can be used to simplify calculations for some of them. Back when this had to be done by hand, such shortcuts were invaluable. Nowadays, you'll use a computer to calculate correlations for you, in which case the many formulas need not concern you. To help you understand, conceptually, what is going on, here's one version of a formula that can be used to calculate a correlation:

$$r = \Sigma(z_x \times z_y) / N$$

To use this formula, you first standardize the $X$ and $Y$ variables by converting them to $z$ scores; that's what $z_x$ and $z_y$ represent. Then, for each of the $N$ cases, you multiple the $z$ scores for the $X$ and $Y$ variables. The average of these products is the correlation.

To see how this gives us something very useful, let's revisit the first scatterplot shown in this chapter, this time standardizing both SAT scores and adding reference lines at $z$ scores of 0 on both axes:




[39] Yes, life would be simpler if this was also called a correlation and symbolized with $r$ plus a subscript, but that's just not how it is.

Notice that most of the data points fall in either the upper-right or the lower-left quadrants. This means that for most points, the *z* scores are either both positive values or both negative values. When we multiply two positive—or two negative—*z* scores, we get positive products. There are relatively few points for which multiplying the *z* scores will yield a negative product (i.e., few points with positive *z* for Math and negative *z* for Verbal, or vice versa). Thus, the average of the products will be a strong positive value. In this case, that average comes to $r = .56$. The sign captures the upward trend in the scatterplot, the fact that SAT Verbal scores tend to increase along with SAT Math scores. The size captures the extent to which the points tend to cluster around a best-fitting line, meaning that the points fall into two diagonally aligned quadrants. Here, the association is fairly strong.

When using a correlation to test the relationship between two variables, the null hypothesis represents no association and the alternative hypothesis represents the opposite:

$H_0: \rho = 0$
$H_1: \rho \neq 0$

The Greek letter ρ (rho) represents the population correlation, and the null value is always 0. These are nondirectional hypotheses, and one can test directional hypotheses if desired (e.g., $H_0: \rho \leq 0$ and $H_1: \rho > 0$).

The *df* for a correlation is $N - 2$, the same as for an independent groups *t* test because there are two sample statistics used to estimate population parameters.[40] You can consult a table of critical values to determine whether an observed correlation falls in the critical region, but it's easier to obtain the *p* value from computer output and compare this to the α level. As usual, if $p < \alpha$ you reject $H_0$, otherwise you retain $H_0$. For example, for the SAT scores plotted above, there is a statistically significant correlation: $r(198) = .56$, $p < .001$.

## Using SPSS

To generate a scatterplot and calculate a correlation on SPSS, you first enter your data into two separate variables (columns). To serve as an illustration, two variables from the parole data set are used. The variables are "LCSF" and "Educ" (years of education). The full data set didn't fit onto the screen, but here's the beginning:

[40] Specifically, the sample statistics are the variance of *X* and the variance of *Y*, each of which is used to estimate its population variance. This is not apparent when you view the formula for calculating a correlation from *z* scores, but that's only because the variances were standardized to 1 when creating *z* scores and therefore drop out of the correlation formula.

| lcsf | educ |
|---|---|
| 3 | 12 |
| 6 | 11 |
| 2 | 18 |
| 3 | 12 |
| 1 | 12 |
| 4 | 10 |
| 11 | 9 |
| 5 | 9 |
| 13 | 10 |
| 3 | 12 |
| 3 | 12 |
| 6 | 12 |
| 4 | 9 |
| 10 | 9 |
| 10 | 6 |
| 11 | 10 |
| 11 | 12 |
| 3 | 16 |
| 6 | 11 |
| 8 | 9 |
| 10 | 9 |
| 7 | 10 |

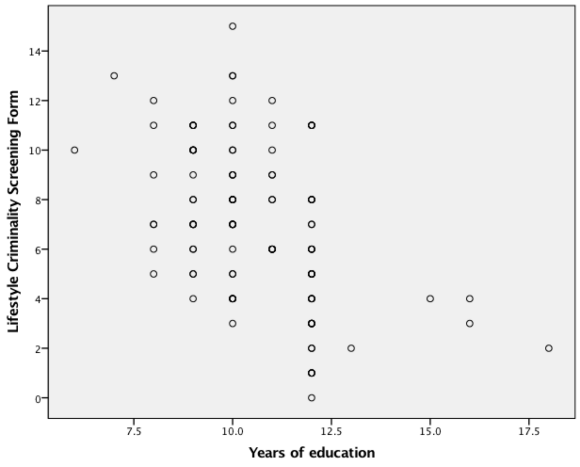
Next, you use the following commands:

```
graph
/scatterplot(bivar) = educ with lcsf

corr vars = educ lcsf
```

The "graph" command will generate a scatterplot. Specify your *X* variable (here, "Educ") and then your *Y* variable (here, "LCSF"), separated by "with". The "corr" command will calculate a correlation coefficient between the *X* and *Y* variables that you specify.

The graph command will produce a scatterplot:




Though there are tied scores, this plot shows an inverse relationship: Higher LCSF scores are associated with fewer years of education.

The correlation command will produce a correlation matrix:

**Correlations**

| | | Years of education | Lifestyle Criminality Screening Form |
|---|---|---|---|
| Years of education | Pearson Correlation | 1 | -.480 |
| | Sig. (2-tailed) | | .000 |
| | N | 114 | 114 |
| Lifestyle Criminality Screening Form | Pearson Correlation | -.480 | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 114 | 114 |

Even when you list only two variables on the correlation command, SPSS produces a matrix of results. Cells along the diagonal represent the correlation of a variable with itself, which you can ignore. Cells above and below the diagonal are mirror images, so you can also ignore either the top or the bottom of the matrix. In this case, there is only a single cell with results that you need. First, SPSS provides the *r* value (labeled as "Pearson Correlation"). Note that even though it's labeled as a Pearson correlation, you can use the same command on SPSS to calculate any of the other correlations described earlier ($r_S$, $r_{pb}$, or $\phi$). The output will still be listed as "Pearson Correlation", but if you provide two ranked variables it will be $r_S$ (and likewise for $r_{pb}$ or $\phi$). Second, SPSS provides the *p* value (labeled as "Sig. (2-tailed)"). If you want to perform a 1-tailed test, simply divide this value by 2 to get the correct *p* value. Third, SPSS provides *N*. Note that you need to calculate $df = N - 2$ to report this in APA style.

## APA Style

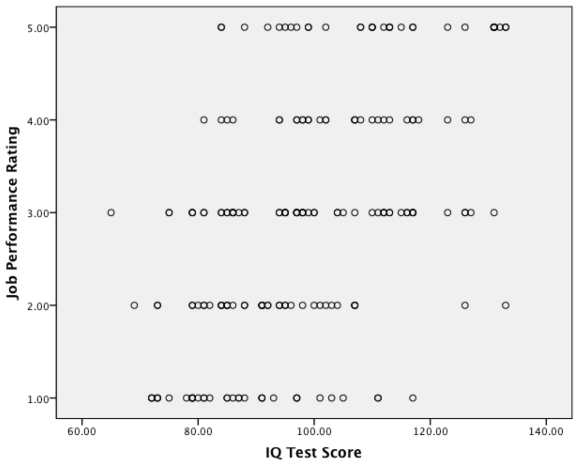
Scatterplots are seldom presented in research reports. As discussed in the next chapter, these are tools for you to check for influences on the correlation that you'd want to know about. Correlations themselves are very simple to report in a single sentence in APA style. Whereas a measure of effect size is appended to many kinds of statistical results (e.g., a *z* or *t* test is followed by a *d* value, an *F* test is followed by an $\eta^2$ value), a correlation coefficient is its own measure of effect size. Here's what the results would look like for the correlation between LCSF scores and education:

Years of education correlated statistically significantly with scores on the Lifestyle Criminality Screening Form (LCSF), $r(112) = -.48$, $p < .001$.

Notice that this was phrased in a way that indicates a 2-tailed test. The correlation coefficient indicates the direction of the observed effect—that education is inversely related to LCSF score—but the phrasing is neutral. Note also that the decision to treat one variable as *X* and the other as *Y* is arbitrary in correlational analyses. This means that you can reverse the *X* and *Y* variables with no change in the correlational results.

## Problems

Below is a scatterplot for 200 employees' scores on an IQ test (for which $\mu = 100$, $\sigma = 15$) and ratings of their job performance on a 5-point scale.




1. If you were to calculate a correlation coefficient, would you expect its sign to be positive or negative? Why?
2. If you were to calculate a correlation coefficient, approximately how large would you expect it to be? Just take your best guess, keeping in mind the possible range of values for any correlation.
3. Below is the SPSS output for a correlation analysis. Report the results in APA style.

**Correlations**

| | | IQ Test Score | Job Performance Rating |
|---|---|---|---|
| IQ Test Score | Pearson Correlation | 1 | .510 |
| | Sig. (2-tailed) | | .000 |
| | N | 200 | 200 |
| Job Performance Rating | Pearson Correlation | .510 | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 200 | 200 |

4. Suppose that both of these variables were converted to ranks. The highest IQ score would become a 1 (the highest rank), the second-highest IQ would become a 2, and so forth. Job performance would still vary along just 5 values because there are so many tied scores; the highest value (5) would become a 1 (the highest rank), the second-highest (4) would become a 2, and so forth. When the correlation is calculated (see table below), what type of correlation coefficient does this become? Report these new results in APA style.

**Correlations**

| | | Rank of IQ | Rank of Performance |
|---|---|---|---|
| Rank of IQ | Pearson Correlation | 1 | .513 |
| | Sig. (2-tailed) | | .000 |
| | N | 200 | 200 |
| Rank of Performance | Pearson Correlation | .513 | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 200 | 200 |

5. Suppose that IQ scores were split at the median and recoded as 1 = low, 2 = high, with job performance remaining on its original 5-point scale. When the correlation is recalculated (see table below), what type of correlation coefficient does this become? Report these new results in APA style.

**Correlations**

| | | IQ Dichotomy | Job Performance Rating |
|---|---|---|---|
| IQ Dichotomy | Pearson Correlation | 1 | .423 |
| | Sig. (2-tailed) | | .000 |
| | N | 200 | 200 |
| Job Performance Rating | Pearson Correlation | .423 | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 200 | 200 |

6. Suppose that both IQ scores and job performance ratings were split at their median values and recoded as 1 = low, 2 = high. When the correlation is recalculated (see table below), what type of correlation coefficient does this become? Report these new results in APA style.

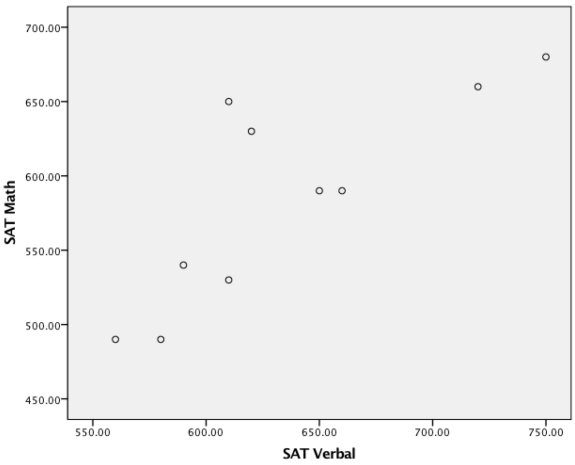
**Correlations**

| | | IQ Dichotomy | Job Performance Dichotomy |
|---|---|---|---|
| IQ Dichotomy | Pearson Correlation | 1 | .410 |
| | Sig. (2-tailed) | | .000 |
| | N | 200 | 200 |
| Job Performance Dichotomy | Pearson Correlation | .410 | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 200 | 200 |

7. What is the coefficient of determination for the original correlation (shown in problem #3)? What does this number represent?
8. What is the coefficient of determination for the correlation when IQ was dichotomized (shown in problem #5)? What does this number represent?
9. Why is the value for #8 so much smaller than the value for #7?

* * *

10. Below are the scatterplot and correlation analysis for the very small sample of SAT data from the chapter on the related samples *t* test. Report the results in APA style.




**Correlations**

| | | SAT Verbal | SAT Math |
|---|---|---|---|
| SAT Verbal | Pearson Correlation | 1 | .806 |
| | Sig. (2-tailed) | | .005 |
| | N | 10 | 10 |
| SAT Math | Pearson Correlation | .806 | 1 |
| | Sig. (2-tailed) | .005 | |
| | N | 10 | 10 |

11. Enter the SAT data (shown below) into SPSS. Follow the instructions in the text for how to organize the data file and enter the commands to generate a scatterplot and calculate a correlation coefficient. Check that your output matches what's shown above.

| Math | Verbal |
|---|---|
| 540 | 590 |
| 630 | 620 |
| 590 | 650 |
| 530 | 610 |
| 490 | 580 |
| 660 | 720 |
| 590 | 660 |
| 490 | 560 |
| 650 | 610 |
| 680 | 750 |

**Problems 1 – 9 are due at the beginning of class.**

# 16. Factors Influencing Correlation

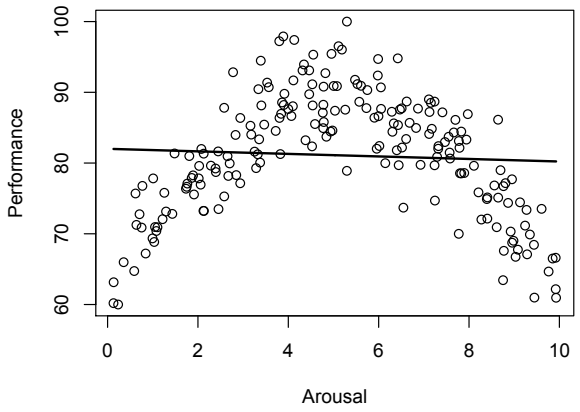
## Overview

The previous chapter introduced the correlation coefficient, and this chapter deals with ways that it can mislead the unwary. The size, and even the sign, of the correlation coefficient can be affected by characteristics of the data that you might not notice unless you check carefully.

## Nonlinear Relationship

The correlation coefficient quantifies the extent to which the data points in a scatterplot tend to cluster around a line. If there's a strong relationship between two variables, but it's nonlinear, the correlation will underestimate the strength of the relationship.

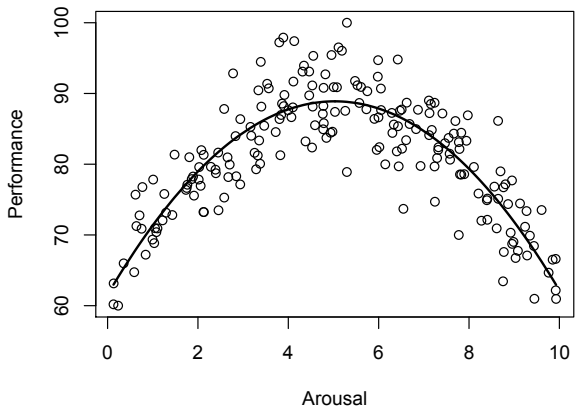
For example, the well-known Yerkes-Dodson law predicts a nonlinear relationship between physiological or mental arousal and performance on a task. Specifically, the relationship is expected to resemble an inverted U, with the optimal level of arousal being moderate. Performance suffers with too little arousal, due to lack of attention or interest, or too much arousal, due to overstimulation or anxiety. Illustrative data are plotted below, along with the line of best fit:[41]




As you can see, the line is a poor fit to these data. The correlation of $r$ = -.06, which is very close to 0, underestimates the strength of the relationship. This demonstrates the importance of examining a scatterplot whenever you calculate a correlation. You could be fooled by the low correlation if you didn't check the plot.

Whenever you inspect a scatterplot and find that a curve fits better than a line, the best response is to fit a curve to the data. Determining what kind of curve to use is an art in itself, and one we will not explore here. In this particular case, it turns out that a parabola (an equation with coefficients for the $X$ variable and $X^2$) fits the scatterplot much better than a line (an equation based only on $X$ values). Here's the same data, this time modeled with the best-fitting parabola rather than the best-fitting line:

[41] We'll see how "best fit" is defined, plus how to find the equation of the best-fitting line, in the next chapter.

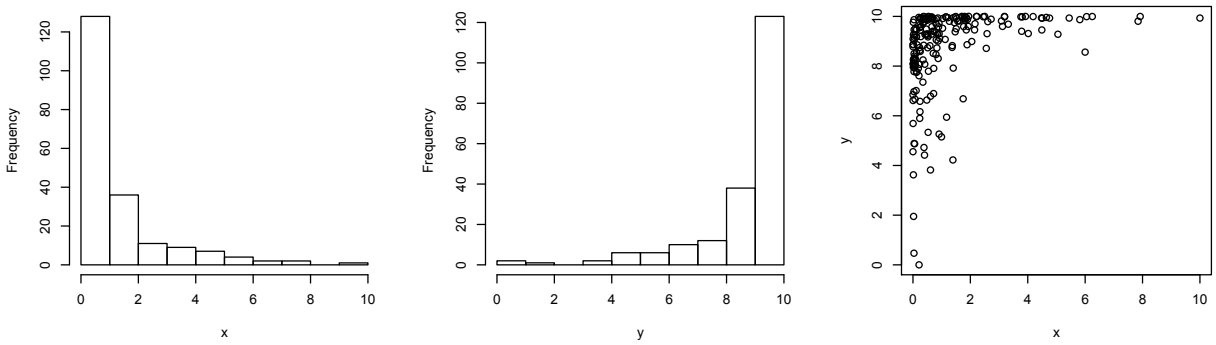



When the data are modeled using a parabola rather than a line, the correlation increases from a meager $r$ = -.06 to a very large $R$ = .86. This reflects what we see in the scatterplot, a very strong relationship between arousal and performance. We'll see why the correlation coefficient for the parabola is expressed as a capital $R$ in the next chapter.

## Different Distributions

When the distributions of the $X$ and $Y$ variables differ, the points in a scatterplot cannot fall along a straight line. This, by itself, reduces the correlation relative to what it would have been had the variables had more similar distributions.

For example, suppose $X$ is positively skewed and $Y$ is negatively skewed. This would force most of the points in the scatterplot into the upper-left corner, where values are low for $X$ but high for $Y$. Below are histograms for such an $X$-$Y$ pair, followed by the scatterplot:

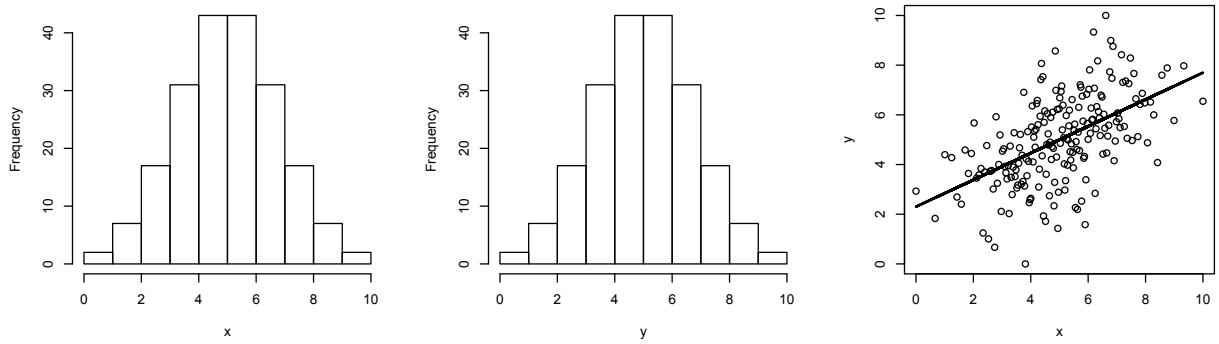



There's no way for a line to fit this plot well, and the correlation is only $r$ = .33. To get a better estimate of the strength of the relationship between $X$ and $Y$, you'd need to transform the data. There are at least two ways to do this.

First, you could calculate Spearman's rank-order correlation rather than the usual Pearson correlation. Recall that Spearman's correlation is for ranked data. By converting both $X$ and $Y$ to ranks, you force their distributions to be equivalent. Specifically, they'd be uniform distributions: The highest score is ranked 1, the second-highest is ranked 2, and so forth, all the way through the lowest score, which is ranked $N$. When this is done for the data shown above, the correlation increases from $r$ = .33 to $r_S$ = .57. Using Cohen's rules of thumb, that's the difference between a medium and a large correlation. Another way to

express the difference is with the coefficient of determination. This increases from $r^2 = .11$ to $r_S^2 = .32$, which means nearly three times as much variance is explained once data are converted to ranks.

Second, you could normalize the data by using nonlinear transformations for each variable. This would also force their distributions to be equivalent, this time normal rather than uniform. When a percentile transformation is used to normalize these data,[42] the correlation increases from $r = .33$ to $r = .54$. Once again, that's the difference between a medium and a large correlation, or between $r^2 = .11$ and $r^2 = .29$. Here's what the histograms and the scatterplot look like after normalizing the variables:




Whether ranked or normalized, the distributions of *X* and *Y* became more similar to one another and the size of the correlation increased substantially. This better reflects the true strength of the relationship between *X* and *Y* for these data.

## Outliers

The presence of one or more outliers can exert a strong influence on the correlation. One way to identify outliers is to examine histograms for *X* and *Y*, but the scatterplot should also be checked for the presence of **multivariate outliers**. A multivariate outlier is a case whose score isn't extreme on either *X* or *Y*, but it is extreme when *X* and *Y* are considered together. For example, an adult who is 6 feet tall isn't an outlier, neither is an adult who weighs 120 pounds. However, an adult who's 6 feet tall and weighs 120 pounds is a multivariate outlier. That's a highly unusual combination of height and weight.

Below is a series of four scatterplots showing the relationship between SAT scores and college GPA. Each graph contains $N = 19$ cases plotted as open circles, plus the line of best fit for these cases only. The first plot (upper left) contains nothing more and serves as a point of reference, with $r = .71$.

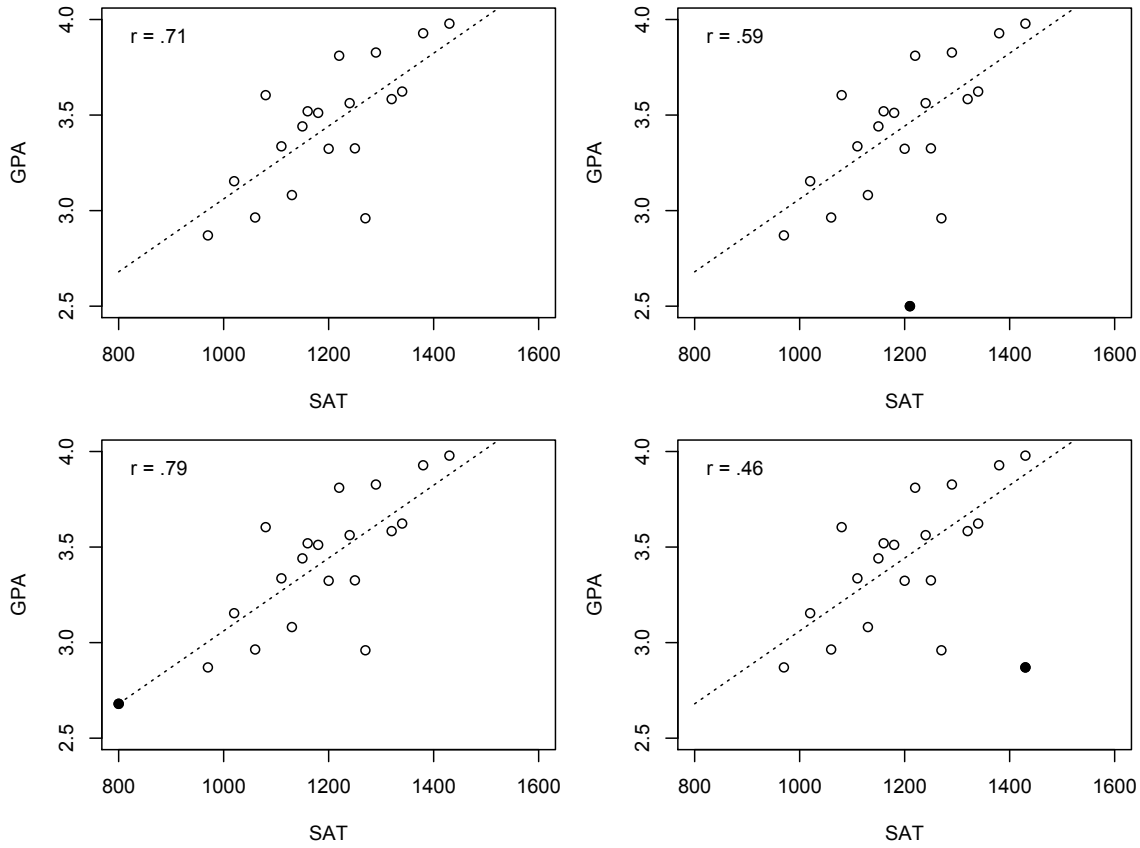
The second plot (upper right) adds to the original 19 cases a single outlier that's near the mean SAT but at a very low GPA. This point doesn't fit the trend in the original data, and as a result the correlation is reduced from $r = .71$ to $r = 59$. Expressing this using the coefficient of determination shows that $r^2$ drops from .51 to .34, a one-third reduction in the proportion of variance explained. Checking the histogram or the scatterplot would have

[42] To perform the percentile transformation, you first convert each score to a percentile and then calculate the *z* score that corresponds to each of these percentiles.

revealed this outlier, which could have been removed to recalculate the correlation without the influence of the extreme score.

The third plot (lower left) adds a single outlier to the original 19 cases that happens to fall along the line that best fit those data. This outlier, also identifiable in a histogram, increases the correlation from $r = .71$ to $r = .79$. This is a fortunate coincidence in the sense that the extreme score happened to be perfectly consistent with the trend in the data.

The fourth plot (lower right) adds a single multivariate outlier that matches both the highest SAT and the lowest GPA in the sample. This point is very far from the line that best fit the original data, and as a result the correlation is reduced from $r = .71$ to $r = .46$. The coefficient of determination drops from $r^2 = .51$ to $r^2 = .21$, a reduction of more than half in the proportion of variance explained. Histograms would not have revealed this outlier because it's not extreme on *X* or *Y*. It only stands out in the scatterplot.

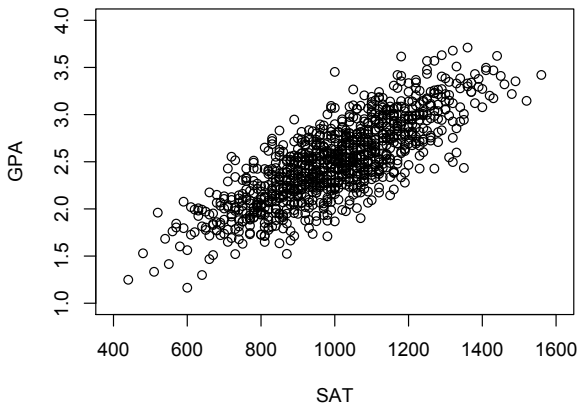



Whenever you identify one or more outliers, it's a good idea to calculate and report the correlation both with and without the outliers. That way, a reader can see how strong an influence the outlier(s) exert on the results.

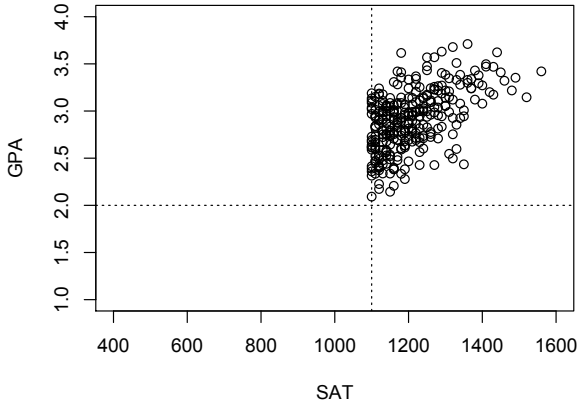
## Restriction of Range

The correlation between two variables usually will be reduced if there is a **restriction of range** on *X* or *Y*, meaning that scores do not vary across the full range of possible values.

For example, suppose that 1,000 students apply to a college, all are accepted, and their SAT (Math + Verbal) scores are plotted along with their first-year college GPAs. Here's what that might look like; the correlation in this case is $r = .80$:




What would happen if this had been a college that only admitted students at or above an SAT of 1100, and that also dismissed students if their GPA fell below 2.00? Thresholds like these are common at many selective colleges. Of the 1,000 applicants shown above, only $N = 299$ would be admitted (based on a high enough SAT) and remain enrolled (based on a high enough GPA). Here's what the scatterplot would look like for these individuals, with dotted lines showing the SAT and GPA thresholds:




The correlation among these 299 students is only $r = .53$. Expressed using the coefficient of determination, that's a drop from $r^2 = .64$ for the 1,000 applicants to $r^2 = .29$ for the 299 enrolled students, or a reduction of more than half the variance explained.

What this demonstrates is that correlations calculated in samples with restricted ranges can seriously underestimate the true correlation between *X* and *Y*. The preferred way to get a better estimate is to design research so that samples will vary along the full ranges on all relevant variables. Unfortunately, that's not always feasible. For example, it's hard to find a college that admits all applicants and that lets all students continue regardless of how low their GPAs fall. In other words, in most real-world settings, the ranges of SAT scores and college GPAs are restricted.

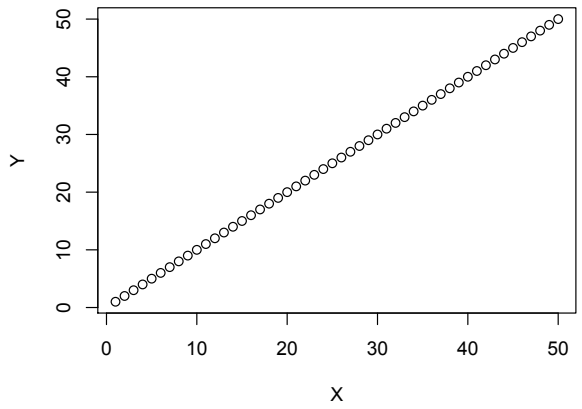
A second-best strategy for getting a better estimate of the correlation between *X* and *Y* is to use an appropriate formula to estimate what the true correlation would have been if

their ranges had not been restricted.[43] For instance, given the variability of the observed SAT scores among the 299 enrolled students and an estimate of the variability of SAT scores among all students who took the SAT, the observed correlation of $r = .53$ can be adjusted to obtain an estimate of $r' = .78$. Making a further correction for range restriction in GPAs would require an estimate of how much their variability was reduced. If one was willing to provide such an estimate, the adjusted correlation between SAT and GPA would increase again.

## Measurement Error

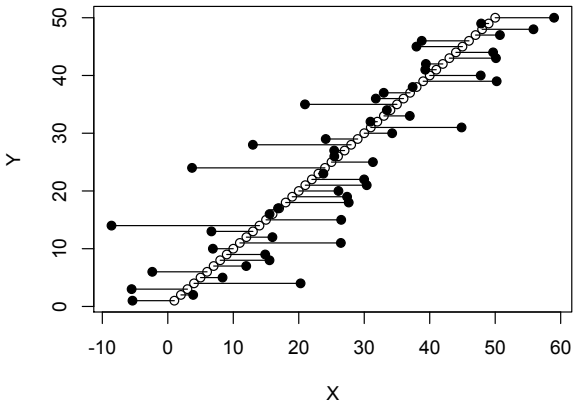
The final factor considered in this chapter influences every correlation to some extent. Nothing can be measured with perfect reliability, hence there will always be some random error in measurements. For example, when we administer an IQ test, the observed IQ scores are imperfect estimates of true IQ scores due to various sources of measurement error. Recall that whereas bias is systematic, error is randomly distributed. Measurement error, as noted in the discussion of regression toward the mean as a threat to internal validity, consists of the random differences between true scores and observed scores.

This is important because the correlation between any two variables will be reduced as the amount of measurement error in each one increases. To understand why, consider a hypothetical case in which *X* and *Y*, if measured free of error, would be perfectly correlated. All the data points, representing true scores, would fall on a line:

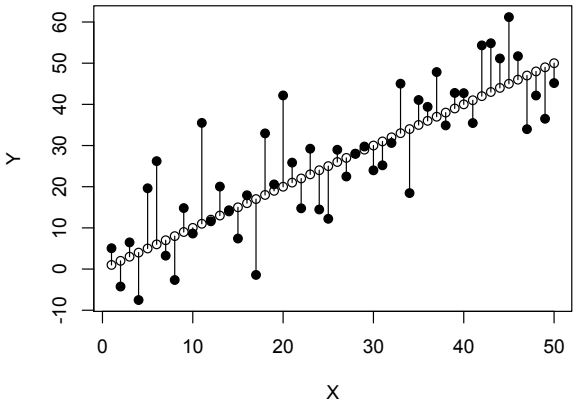



The correlation between true scores is 1.00, but we can never observe true scores. What happens when measurement error is introduced? The error in measuring *X* will nudge each data point a bit to the left or the right, at random, in the scatterplot. In the plot shown below, each true score (open circle) is connected to its corresponding observed score (filled circle) by a line segment that represents random measurement error:

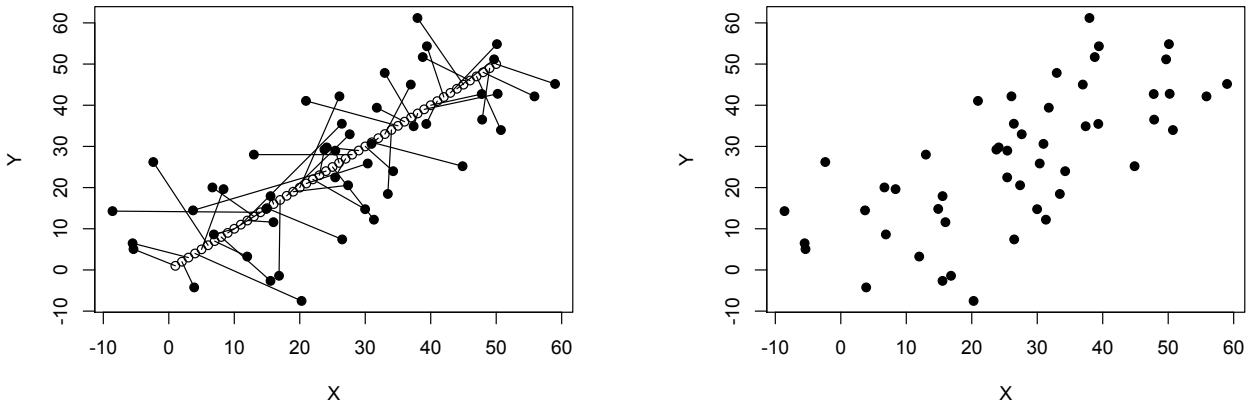
[43] There are many formulas for different kinds of range restriction (e.g., direct vs. indirect range restriction; restriction on *X*, *Y*, or both). An excellent source on this subject is Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112-118.




Of course, measurement error is not limited to *X*. The *Y* variable will also be affected, with scores nudged up or down. Here's a look at the effect of measurement error on *Y*:




Finally, here are scatterplots that show the combined influence of measurement error on *X* and *Y*. The plot on the left includes the true scores and the lines that connect them to the corresponding observed scores. The plot on the right shows only the observed scores:




Notice in the plot on the left that very few of the observed scores (the filled circles) fall on the original line running through the true scores (the open circles). Measurement error disperses observed scores at random to create the cloud of points shown in the plot on the right. In this case, measurement error resulted in a correlation between observed scores of $r = .69$, a serious underestimate of the correlation between true scores of $r = 1.00$.

Measurement error affects all correlations, whether the correlation between true scores is large or small. Fortunately, there's an easy way to estimate the correlation between true scores. This is done by adjusting the correlation between observed scores to take into account the reliability with which each variable was measured.

Reliability is expressed as a correlation, with $r_{xx}$ representing the reliability of $X$ and $r_{yy}$ representing the reliability of $Y$. If there were no measurement error for $X$, $r_{xx} = 1.00$. At the other extreme, if $X$ is measured with no reliability, meaning that it consists of nothing but measurement error, $r_{xx} = .00$. In the example shown above, $r_{xx} = .74$ and $r_{yy} = .66$. These values are typical of the reliability of variables measured in social and behavioral science.

There are many ways to estimate the reliability with which a variable is measured.[44] For present purposes, we'll presume that this information has been provided. Given the observed correlation and estimates of $r_{xx}$ and $r_{yy}$ , a simple formula can be used to estimate the correlation between true scores:

$$r' = r \,/\, \text{sqrt}(r_{xx} \times r_{yy})$$

This formula adjusts the correlation upward using the average reliability of the two measures.[45] Only if one or both reliabilities are only slightly below 1.00 there will not be much of an adjustment, but it will always be the case that $r' \geq r$. When one or both reliabilities are well below 1.00 there will be a more substantial adjustment.

For example, consider the artificial data used in the scatterplots shown above. The correlation between true scores was 1.00. The formula estimates this correlation to be $r' = .69 \,/\, \text{sqrt}(.74 \times .66) = .99$. For actual data, the true correlation would seldom approach 1.00; this calculation is shown only to illustrate how to use the formula.

As a more realistic example, suppose that a sample of employees' IQ test scores correlate with ratings of their job performance at $r = .40$. This is a typical value observed in research on this subject. If the reliability of the IQ test is estimated to be $r_{xx} = .90$ and the reliability of the job performance ratings is estimated to be $r_{yy} = .70$, the estimated correlation between true scores would be $r' = .40 \,/\, \text{sqrt}(.90 \times .70) = .50$. This is very close to what meta-analysis estimates as the correlation between IQ test scores and job performance ratings.

## Problems

1. The faculty at a competitive Ph.D. program find that the correlation between current students' GRE scores and graduate GPAs is only $r = .30$. They conclude that the GRE is not a very useful predictor of success in graduate school and should not be used for graduate admissions. Why is $r = .30$ probably an underestimate of the strength with which GRE scores predict graduate GPAs? What can be done to get a better estimate?
2. Increasing the dosage of a medication has strong and positive effects up to a certain point, and then the response reaches a plateau. There appears to be a strong dose-

[44] An excellent source on this subject is Schmidt, F. L. & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199-223.

[45] The denominator is a type of average known as the geometric mean, which is calculated as the $N^{th}$ root of the product of $N$ scores.

response relationship, but the correlation is only $r = .20$. Why is this probably an artificially low estimate of the strength of the dose-response relationship? What can be done to get a better estimate?

3. There is an old saying that people "drive as they live," meaning that one's personality is reflected in their driving habits. An investigator collects data on impulsivity (measured using a brief self-report questionnaire) and unsafe driving (indexed as the number of tickets for traffic violations recorded by an insurance company), expecting to find a large correlation. It turns out that $r = .10$. Why is this probably an underestimate of the strength of the relationship between impulsivity and unsafe driving? What can be done to get a better estimate?
4. An investigator wonders whether the length of articles appearing in a journal (measured as the number of words) is related to the scholarly impact of the articles (measured as the number of citations it receives in other articles published within the next 5 years). The distribution of article lengths is close to normal, and the distribution of citations is extremely positively skewed. The correlation between articles' length and impact is $r = .10$, which the investigators find to be surprisingly small. Why is this probably an underestimate of the strength of the relationship between article length and article scholarly impact? What can be done to get a better estimate?
5. In a sample of psychotherapists, the relationship between clinical experience (indexed by years of practice) and accuracy of judgment (measured using a series of diagnostic problems) is studied. This sample includes a couple of individuals who have been practicing much longer than everyone else, but their training is obsolete and their knowledge of current diagnostic guidelines is quite poor. The correlation between experience and judgment accuracy is only $r = .10$. Why is this probably an underestimate of the strength of the relationship between clinical experience and accuracy of judgment? What can be done to get a better estimate?

* * *

6. A clinical psychologist is interested in the comorbidity, or co-occurrence, of Post-Traumatic Stress Disorder (PTSD) and Major Depressive Disorder (MDD). Structured interviews are administered to assess the symptoms of each disorder, and a sample of patients who meet diagnostic criteria for both disorders is formed. Within this sample, the correlation between PTSD and MDD symptoms is $r = .30$. Why is this probably an underestimate of the strength of the relationship between PTSD and MDD symptoms? What can be done to get a better estimate?
7. An investigator hypothesizes that people who can remember more of their dreams are more creative individuals. To test this, a sample of undergraduate students is asked to recall how many dreams they experienced during the past week and to write a Haiku (a short poem that typically has 5, 7, and 5 syllables per line). English professors rate the creativity of these Haikus, and this correlates $r = .20$ with the number of dreams. Why is this probably an underestimate of the strength of the relationship between dream frequency and creativity? What can be done to get a better estimate?
8. All twenty students taking their first algebra class in middle school are asked to estimate how many hours they spend on homework and study outside of class. Most

students spend at least 10 hours per week mastering the material, but one student finds it so easy to grasp that she never brings the book home. She's able to skim it during class while paying just enough attention to her teacher to attain a perfect score on every quiz and test. The correlation between study time and grades in the course is $r = .30$. Why is this probably an underestimate of the strength of the relationship between study time and grades? What can be done to get a better estimate?

9. One of the three key features of prospect theory, a cornerstone of behavioral economics, is that there are diminishing returns for gains or losses. For example, gaining $200 feels less than twice as good as gaining $100. A neuroscientist measures brain activity in response to various levels of financial gain or loss and finds that this correlates $r = .50$ with subjects' ratings of how good or bad this makes them feel. Why is this probably an underestimate of the strength of the relationship between actual and perceived gains or losses? What can be done to get a better estimate?

10. A history instructor wonders whether students who complete a 10-item quiz the fastest score the highest. On a typical quiz, most students turn in their quizzes after about 5 minutes, a handful finish within the next 5 minutes, and a few take even longer. Many students get 9 or 10 items correct and most get at least 7, but a few score as low as 3 or 4 correct. The correlation between time to complete the quiz and the number of items correct is $r = .20$. Why is this probably an underestimate of the strength of the relationship between time and score? What can be done to get a better estimate?

* * *

11. A developmental psychologist tests the expressive vocabulary of a sample of children varying in age between 12 and 36 months. Because the number of words children use is expected to double every few months or so, she is surprised that the correlation between age and vocabulary is only $r = .50$. Why is this probably an underestimate of the strength of the relationship between these variables? What can be done to get a better estimate?

12. An industrial/organizational psychologist wonders how much income increases with education. She gathers data for a sample of practicing attorneys, and finds that years of schooling correlates only $r = .10$ with yearly earnings. Why is this probably an underestimate of the strength of the relationship between these variables? What can be done to get a better estimate?

13. A biopsychologist is interested in the relationship between alcohol consumption and academic performance. A sample of college seniors provides their GPAs, which are mostly very high but range down to barely above the minimum for graduation, and the number of drinks they consume in a typical week when classes are in session, which is usually zero or very few but ranges up into the dozens for some students. The correlation between drinking and grades is $r = .20$. Why is this probably an underestimate of the strength of the relationship between these variables? What can be done to get a better estimate?

14. A personality psychologist wonders whether people who are more extraverted, who tend to have more friends and acquaintances, also have larger social networks on social media. A sample of college students with Facebook accounts completes a standard

personality inventory that includes an extraversion scale. The average number of friends they have on Facebook was about 650, though a few students had dramatically larger networks with more than 2,000 friends. The correlation between extraversion scores and number of Facebook friends was $r = .30$. Why is this probably an underestimate of the strength of the relationship between these variables? What can be done to get a better estimate?

15. A health psychologist administered a lengthy questionnaire to a sample of college students. One item asked students whether they consider themselves to be healthy or unhealthy, and another whether they experience a high or low level of stress on a daily basis. The correlation between stress and health was $r = .20$. Why is this probably an underestimate of the strength of the relationship between these variables? What can be done to get a better estimate?

**Problems 1 – 5 are due at the beginning of class.**

# 17. Regression

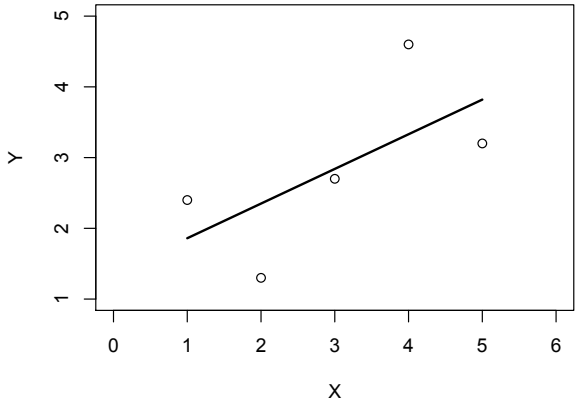
## Overview

The correlation coefficient quantifies the extent to which a line fits the data in a scatterplot. A **regression equation** identifies the best-fitting line. If all that you want to know is how strongly two variables are related, perhaps to test whether this relationship is statistically significant, then correlation is sufficient. If you want to make predictions of *Y* from scores on *X*, you need to know the equation of the **regression line**. To use more than one *X* variable to make predictions, you can extend this to **multiple regression**.

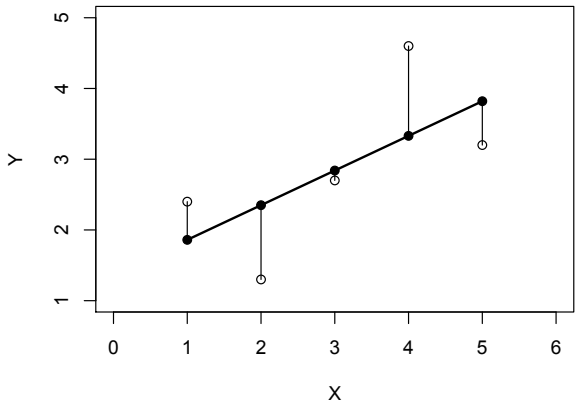
## Prediction Equation

In geometry, we learn that any two points can be connected by a line with the equation $Y = mX + b$, where $m$ is the slope and $b$ is the intercept. In statistics, we know that for any two variables *X* and *Y* that are imperfectly correlated (meaning that $|r| < 1.00$), the points in a scatterplot will not fall on a line. A few points might be on, or close to, the line, but most will be scattered around it. Regression analysis identifies the line that best fits the scatterplot, where best fit is defined as minimizing the sum of the squared errors in prediction, or **residuals**. For each data point, the residual is the difference between the actual *Y* value and the *Y* value predicted by the regression equation.

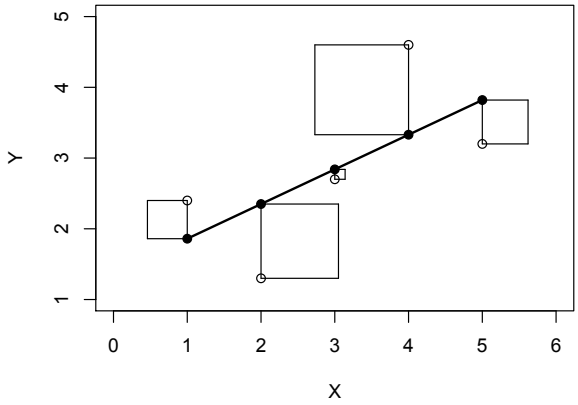
To illustrate, consider a very simple scatterplot with just 5 values. Here's the plot, including the regression line:



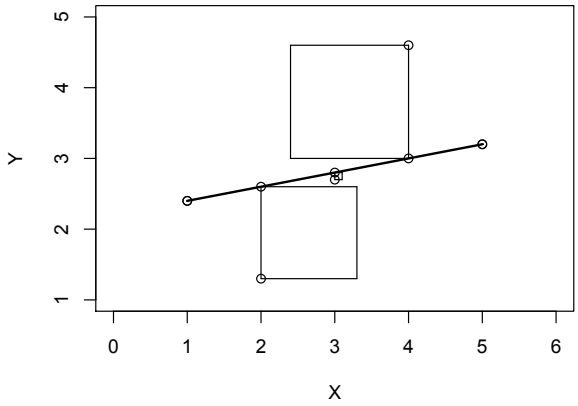

Here's the same plot, this time adding the *Y* values predicted by the regression equation (plotted as filled circles, each of which falls on the regression line) and the residuals (vertical lines connecting each observed *Y* value to its corresponding predicted *Y* value):




Finally, here's a plot that squares those residuals:




This regression line minimizes the sum of the areas inside these squares. Here's another line fit to the same data, with the new squared residuals plotted:




This line fits the 1st and 5th data points perfectly, so those residuals are 0, and the residual is also very small for the 3rd data point. However, the 2nd and 4th data points are fit badly by this line. As a result, the sum of the squared residuals is greater than the sum for the best-fitting regression line. We'll let a computer calculate the slope and the intercept of a regression line, but it's important to understand what it's doing. It's minimizing the sum of the squared residuals.

Though it might not be obvious, a regression line is very similar to the mean. The mean minimizes the sum of the squared deviation scores, the distance from each data point to the

mean. The regression line minimizes the sum of the squared residuals, the distance from each data point to the regression line. Thus, you can think of a regression line as a "running mean." It's like an average value of *Y* for scores at each level of *X*.

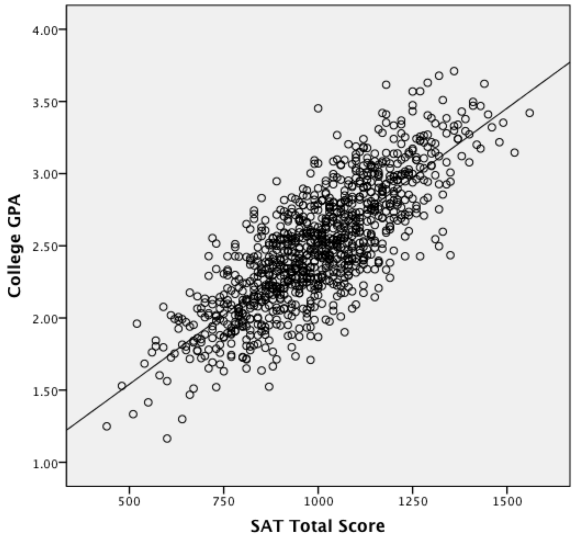
The equation for a regression line is expressed as follows:

$$Y' = bX + a$$

*Y*' is the predicted *Y* value, *X* is the observed *X* value, *b* is the slope, and *a* is the intercept. For the regression line shown above, the equation is:

$$Y' = 0.49X + 1.37$$

This first example was simple, containing just 5 cases, but highly artificial. Let's look at the 1,000 students whose SAT scores and college GPAs were used to illustrate the influence of restriction of range on correlation in the previous chapter. Here's the scatterplot:




Here's the regression equation:

$$GPA' = .0019 \times SAT + .5849$$

An admissions officer could use this equation to predict the college GPA of a new applicant. For example, an applicant with an SAT score of 1200 would be predicted to attain a GPA of $.0019 \times 1200 + .5849 = 2.86$. An applicant with an SAT score of 800 would be predicted to attain a GPA of $.0019 \times 800 + .5849 = 2.10$.

## Accuracy

There are three ways to measure the accuracy of a regression equation. Only one of these is new.

The first measure of accuracy is the correlation itself. For the SAT and GPA data, $r = .80$, a very large correlation.

The second measure of accuracy is the coefficient of determination. For the SAT and GPA data, $r^2 = .64$, which means that SAT scores account for 64% of the variance in GPAs. That's a very large effect.

The third measure of accuracy is the **standard error of the estimate** ($SE_{est}$). This is the typical distance from a predicted *Y* value to the observed *Y* value. It's calculated just like a *SD*, but rather than using deviation scores (distances from the mean) we use residuals (distances from a regression line). For the SAT and GPA data, $SE_{est} = 0.26$. That means that a typical predicted GPA is about 0.26 points away from the observed GPA.

Whereas *r* and $r^2$ are on a standardized scale, ranging from .00 to 1.00, $SE_{est}$ provides a measure of accuracy that's scaled in the units of the *Y* variable. Suppose that someone making admissions decisions wants to know how accurately, in terms of actual GPAs, predictions based on SAT scores would be. Neither *r* nor $r^2$ is helpful because they have nothing to do with the GPA scale. $SE_{est}$, on the other hand, is scaled in GPA units. Predictions would be accurate with a margin of error of $\pm 0.26$.

In sum, we've seen that a regression line is similar in important ways to the *M* (both are located in the middle of a distribution) and the $SE_{est}$ is similar in important ways to the *SD* (both represent a typical distance from a data point to the middle).

## Multiple Regression

When an equation with a single predictor (*X* variable) is used to predict an outcome (*Y* variable), this is called **simple linear regression**. When more than one predictor is included in the equation, this is called **multiple regression**. Multiple regression is a fairly straightforward extension of simple linear regression that can be used for many purposes.

Recall the equation of the best-fitting line in simple linear regression:

$$Y' = bX + a$$

$Y'$ is the predicted value of the outcome (*Y*), *X* is the predictor, *b* is the slope, and *a* is the intercept. Another way of expressing this is that *b* is the regression coefficient, or weight, for the predictor. This indicates how heavily it counts when making predictions. If $b = 0$, then the predictor counts for nothing and all predictions equal the intercept, a constant.

The general equation for multiple regression with *k* predictors is this:

$$Y' = b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + b_0$$

In this formula, each predictor ($X_i$) gets its own regression coefficient ($b_i$), and the constant is labeled $b_0$ rather than *a*. Once again, the regression coefficients indicate how much weight is given to each predictor. The measures of accuracy for simple linear regression extend to multiple regression. The only difference is the notation. Rather than using *r* and $r^2$, we use *R* and $R^2$. The capital *R* indicates that more than one predictor variable was used in the regression equation.

Multiple regression is a very popular data-analytic tool, for many reasons. One reason is that using multiple predictors can increase the accuracy of prediction. If the goal is to explain as much variance as possible in the outcome variable, including more than one predictor will help as long as each is valid and not redundant with others already included in the equation.

A second reason to use multiple regression is that it avoids the problems with splitting cases into groups. Researchers sometimes make such splits in order to compare group means using ANOVA. A better approach is to use multiple regression, which allows you to

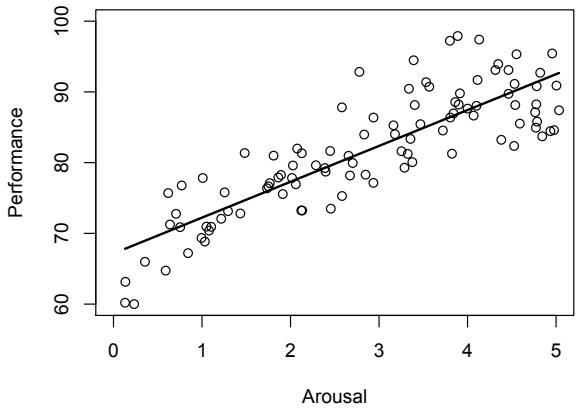
leave quantitative variables in their original, continuous form. This provides better estimates of true effect sizes and retains as much statistical power as possible.

A third reason to use multiple regression is to model a curve. For example, recall the data illustrating the Yerkes-Dodson law (see below for some Yerkes-Dodson data, too). Earlier it was shown that a curve fit the data better than a straight line. Multiple regression allows you model a parabola by including both $X$ and $X^2$ as separate predictors. This is just one example of a curve that can be fit using multiple regression.

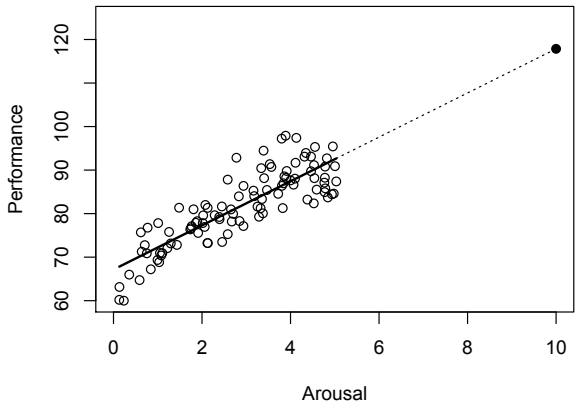
## Extrapolation

Each of the factors influencing correlation (nonlinear relationships, different distributions, outliers, restriction of range, and measurement error) can also affect regression. A new concern is **extrapolation**, or making predictions beyond the range of observed values. The problem with extrapolation is that it's based on an assumption that a trend line will continue indefinitely. By definition, though, there's no evidence available to test this assumption. When you move beyond the range of predictor values in the data, the trend may change. The further beyond the range of observed values you go, the greater the danger that this extrapolation will be inaccurate.

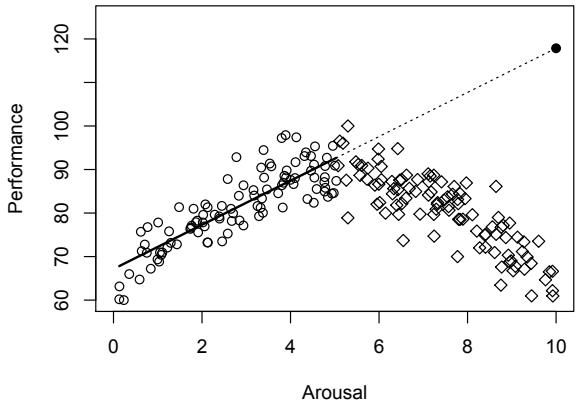
For example, consider what would happen if an investigator studied performance on a challenging crossword puzzle across under experimental conditions that induced levels of mental arousal from very low to moderate. According to the Yerkes-Dodson law, performance should increase along with mental arousal in this range. A scatterplot between arousal and performance might reveal an upward linear trend. To illustrate this possibility, here is a selection from the arousal and performance data shown in the last chapter, specifically all scores below the median level of arousal:




The regression line fits the data quite well, with $r = .85$ and $r^2 = .72$. So far, so good. But what if someone inferred from these findings that subjects would perform even better at much higher levels of mental arousal? Consider an extrapolation up to the maximum arousal level on this scale, a value of 10:




The dotted line extends the regression line well beyond the range of observed scores. Such an extrapolation is unlikely to be borne out. In fact, the Yerkes-Dodson law also predicts that performance will decline at very high levels of mental arousal. Here's the scatterplot for the full range of values, with data points above the median arousal level plotted as diamonds to distinguish them from the data points from which the regression was calculated (plotted as circles).




The regression line is a very poor fit for the higher levels of arousal, failing to support the extrapolation. Be wary of assuming that trends continue into unstudied regions.

## Using SPSS

To perform regression in SPSS, you set up the data file just as you would for a correlational analysis. Enter your data into two variables (columns), one for the predictor (*X*) and one for the outcome (*Y*) to be predicted. To serve as an illustration, the SAT and GPA data from earlier in this chapter are used. The variables are "SAT" and "GPA". The full data set didn't fit onto the screen, but here's the beginning:

| SAT | GPA |
|---|---|
| 440 | 1.25 |
| 480 | 1.53 |
| 510 | 1.33 |
| 520 | 1.96 |
| 540 | 1.68 |
| 550 | 1.42 |
| 560 | 1.76 |
| 570 | 1.84 |
| 570 | 1.80 |
| 580 | 1.60 |
| 590 | 2.08 |
| 590 | 1.80 |
| 600 | 1.56 |
| 600 | 1.16 |
| 610 | 2.02 |
| 610 | 1.73 |
| 620 | 1.75 |
| 620 | 1.99 |
| 630 | 1.98 |
| 630 | 1.93 |
| 630 | 2.01 |
| 640 | 1.99 |

Next, you use the following commands:

```
graph
/scatterplot(bivar) = sat with gpa

reg vars = sat gpa
/dep = gpa
/enter sat
```

The "graph" command was described in the chapter on correlation, and the scatterplot was shown earlier in this chapter. SPSS will not automatically include a regression line, but you can add one.[46]

To run the "reg" (short for regression) command, list the predictor ($X$) variable (here, "SAT") on the first and third lines and the outcome ($Y$) variable (here, "GPA") on the first and second lines. SPSS will produce several tables, only two of which you'll need.

The first table you'll need is labeled "Model Summary". This table provides all measures of accuracy described in this chapter: $r$, $r^2$ (labeled "R Square"), and $SE_{est}$ (labeled "Std. Error of the Estimate"). Note that SPSS always uses capital letters in the output. If you're doing simple linear regression, the output provides $r$ and $r^2$ even though each is listed with the capital $R$ rather than the lowercase $r$.

[46] Double-click on a scatterplot to open the chart editor. From the "Elements" menu, choose "Fit Line at Total", select "Linear" in the "Fit Method" area of the dialogue box, uncheck the "Attach label to line" option near the bottom, click "Apply" and then "Close", and close the Chart Editor to return to the output window.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .798[a] | .637 | .636 | .258746036 |

The second table you'll need is labeled "Coefficients". This contains the slope (*b*) and intercept (*a*) of the regression equation. The values appear in the column labeled "B" in the "Unstandardized Coefficients" section of the table. The row labeled "(Constant)" contains the intercept, and the row beneath it (labeled with your predictor variable, here "SAT Total Score") contains the slope. If you want to know whether the slope is statistically significantly different from 0, the *p* value appears in the column labeled "Sig." and the row labeled with your predictor variable (here, "SAT Total Score").

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | .585 | .046 | | 12.592 | .000 |
| | SAT Total Score | .002 | .000 | .798 | 41.822 | .000 |

a. Dependent Variable: College GPA

This example shows results for simple linear regression. You can perform multiple regression by entering more than one predictor (*X*) variable into the data file and command syntax.

## APA Style

Scatterplots are seldom presented in research reports, but they're invaluable tools for you to check for potentially problematic influences on regression. Regression results can be reported with or without commenting on statistical significance. Here's how the regression results shown above could be reported in APA style:

SAT total scores predicted college GPAs according to the following regression equation:

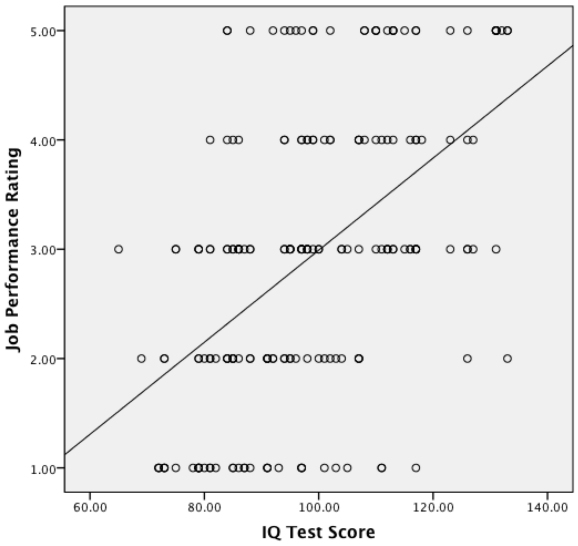
GPA' = .002 × SAT + .585. Not only was the slope statistically significant, $p < .001$, but this

equation was very accurate: $r = .80$, $r^2 = .64$, $SE_{est} = 0.26$.

Notice that an extra decimal place was used for the slope and intercept in the regression equation. Without doing this, the slope would have rounded down to .00, suggesting that SAT was given no weight in predicting GPA. In fact, it was a very strong predictor. The slope is small only because of the difference between SAT units (on the scale of 400 to 1600) and GPA units (on the scale of 0 to 4). You have to multiply SAT scores by a very small number to predict GPAs. SPSS only provides 3 decimal places for the slope and intercept. Earlier, the regression equation was expressed using 4 decimal places, which is even better. You can use more than the usual 2 decimal places in APA style when it's important to do so.

## Problems

Below is a scatterplot and a regression analysis for 200 employees' scores on an IQ test (for which $\mu = 100$, $\sigma = 15$) and ratings of their job performance on a 5-point scale. These data appeared in the chapter on correlation, and a regression line has been added to the scatterplot.




**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .510[a] | .260 | .256 | 1.13500 |

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -1.220 | .503 | | -2.427 | .016 |
| | IQ Test Score | .042 | .005 | .510 | 8.342 | .000 |

a. Dependent Variable: Job Performance Rating

1. Write the equation of the regression line, including the slope and intercept.
2. Is the slope statistically significantly different from 0? How can you tell?
3. What are the values for each measure of accuracy described in this chapter? Briefly explain what each value means, in plain English.
4. Write the results for this regression analysis in APA style.
5. Use the regression equation to predict job performance ratings for individuals with IQ scores of 100 (considered "average" in the general population), 130 (at the border between "superior" and "gifted"), and 160 (at the upper end of the "very gifted" range).
6. In what way is one of the predicted values in #5 more problematic than the others? Why did this happen?

7. Suppose these 200 employees had also been given a test of conscientiousness. Given this additional data, why might it be worthwhile to use multiple regression, rather than simple linear regression?

* * *

8. In his classic work *An Essay on the Principle of Population*, first published in 1798, Thomas Malthus observed that the size of the human population was increasing much more rapidly (an exponential trend) than the size of our food supply (a linear trend). Over a sufficiently long time frame, exponential growth overwhelms linear growth. As a consequence, Malthus believed that mass starvation was ultimately inevitable.

   It's been more than 200 years. The human population has grown dramatically, and famines are relatively rare and usually caused by politics, not an actual (let alone global) food shortage. Malthus' prediction shows no sign of being correct. What might have led Malthus to make this mistake?

* * *

The next series of problems uses the parole data introduced earlier. A simple linear regression analysis was performed to predict years of education from total scores on the Lifestyle Criminality Screening Form (LCSF). The SPSS output is shown below:

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .480[a] | .230 | .224 | 1.566 |

**Coefficients[a]**

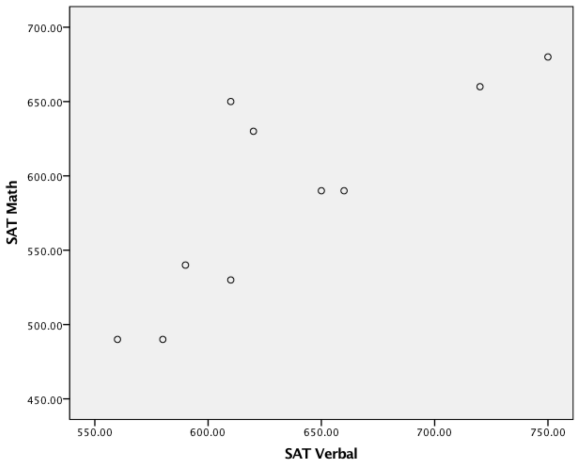
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 12.491 | .354 | | 35.253 | .000 |
| | Lifestyle Criminality Screening Form | -.270 | .047 | -.480 | -5.792 | .000 |

a. Dependent Variable: Years of education

9. Write the equation of the regression line, including the slope and intercept.
10. Is the slope statistically significantly different from 0? How can you tell?
11. What are the values for the measures of accuracy described in this chapter? Briefly explain what each value means, in plain English.
12. Write the results for this regression analysis in APA style.

* * *

13. Below are the scatterplot and regression analysis for the very small sample of SAT data from the chapter on the related samples *t* test. SAT Verbal scores are used to predict SAT Math scores. Report the results in APA style.




**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .806[a] | .650 | .606 | 43.89356 |

a. Predictors: (Constant), SAT Verbal

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -2.541 | 153.028 | | -.017 | .987 |
| | SAT Verbal | .925 | .240 | .806 | 3.855 | .005 |

a. Dependent Variable: SAT Math

14. Enter the SAT data (shown below) into SPSS. Follow the instructions in the text for how to organize the data file and enter the commands to generate a scatterplot and calculate a regression equation. Check that your output matches what's shown above.

| Math | Verbal |
|---|---|
| 540 | 590 |
| 630 | 620 |
| 590 | 650 |
| 530 | 610 |
| 490 | 580 |
| 660 | 720 |
| 590 | 660 |
| 490 | 560 |
| 650 | 610 |
| 680 | 750 |

**Problems 1 – 8 are due at the beginning of class.**

# 18. $\chi^2$ Goodness of Fit Test

## Overview

The final statistic we'll examine, $\chi^2$, is used to analyze nominal data. When cases belong to categories (e.g., marital status, race/ethnicity), $\chi^2$ can be used to test certain kinds of hypotheses. This chapter reviews what's called the **goodness of fit test**. The goal is to determine whether the counts, or frequencies, observed for each of a series of categories differs from what's expected according to a model being tested. You can also compute $\chi^2$ for two or more models to determine which one fits the data better.

## Nonparametric Statistic

Most of the statistics presented in this text are **parametric statistics**, meaning that they make assumptions about population distributions. For example, to test the statistical significance of *z*, *t*, *F*, or *r*, we assume a normal distribution of scores in the population. The $\chi^2$ statistic makes no parametric assumptions. It is, instead, a **nonparametric statistic**.

Many nonparametric statistics avoid the usual parametric assumptions by working with ranked data rather than quantitative data. Spearman's rank-order correlation coefficient is one example. There are nonparametric analogues of *t* and *F* tests, too, that deal with ranked rather than quantitative data. One reason that parametric statistics are more popular is that when their assumptions are satisfied, they provide greater statistical power than the nonparametric alternatives. When one or more parametric assumptions is violated to a problematic extent (e.g., when distributions are extremely skewed), a nonparametric statistic might be a better choice.[47]

Most parametric statistics are used to analyze quantitative data, and most of the nonparametric alternatives are used to analyze ranked data. In contrast, $\chi^2$ is a very popular statistic because it's used to analyze nominal data. For example, consider how you might test the success of systematic desensitization as a treatment for specific phobias. Using this therapeutic technique, the patient learns to relax, and remain relaxed, while progressing through a hierarchy of anxiety-inducing steps related to the feared object or situation. A patient with a fear of snakes would try to work through steps that include imagining snakes, looking at pictures of snakes, and holding live snakes. The key is to remain relaxed at each step before moving to the next. Suppose 30 subjects diagnosed with specific phobias are given treatment using systematic desensitization, and when treatment is complete 24 patients are cured (i.e., they're free of their phobias) and 6 are not.

[47] If you're curious, check out the Wilcoxon rank-sum test (to compare two independent groups), the Kruskal-Wallis one-way ANOVA, or the Friedman two-way ANOVA.

## Performing the $\chi^2$ Test

The $\chi^2$ goodness of fit test can be used to compare the two counts, or **observed frequencies**, of 24 cured and 6 not cured to the **expected frequencies** that correspond to a null hypothesis, or model, being tested. To perform the test, you need to provide the expected frequencies. How many patients would you expect to be cured vs. not cured? Perhaps you want to test a null hypothesis of equal chances of being cured vs. not cured (which means that 1/2 of all patients would be cured and 1/2 would not). That implies expected frequencies of 15 and 15 for the two cells in the design. These are calculated by multiplying the total $N$ of 30 by 1/2 for the cured cell ($30 \times 1/2 = 15$) and 1/2 for the not cured cell ($30 \times 1/2 = 15$). We'll call this model 1. It's helpful to organize the observed frequencies ($f_O$) and expected frequencies ($f_E$) in a table:

| | $f_O$ | $f_E$ |
|---|---|---|
| Cured | 24 | 15 |
| Not Cured | 6 | 15 |

Perhaps you want to test a null hypothesis based on the success rate for another treatment, say a 2/3 success rate. That implies expected frequencies of 20 and 10 for the two cells in the design. These are calculated by multiplying the total $N$ of 30 by 2/3 for the cured cell ($30 \times 2/3 = 20$) and 1/3 for the not cured cell ($30 \times 1/3 = 10$). We'll call this model 2, and here's the table:

| | $f_O$ | $f_E$ |
|---|---|---|
| Cured | 24 | 20 |
| Not Cured | 6 | 10 |

To test each model, all that you need to do is compare the observed and expected frequencies using the formula for $\chi^2$:

$$\chi^2 = \Sigma((f_O - f_E)^2 / f_E)$$

For each cell, you take the difference between $f_O$ and $f_E$, square it, and divide by $f_E$. Once you've done this for all cells, you sum the results to get $\chi^2$. Here's what this looks like when testing model 1, equal chances of being cured vs. not cured:

$$\chi^2 = ((24 - 15)^2 / 15) + ((6 - 15)^2 / 15) = 81/15 + 81/15 = 10.80$$

To determine whether this is statistically significant, you look up a critical value for $\chi^2$ in a table (e.g., the one in Appendix A). The critical value is based on the $\alpha$ level (usually .05) and the *df*, which is the number of cells minus 1. In this case, using $\alpha = .05$ and $df = 2 - 1 = 1$, we find that the critical value is $\chi^2 = 3.84$. This is always a nondirectional test, so you reject $H_0$ whenever the $\chi^2$ calculated for your data exceeds the critical value. Here, $10.80 > 3.84$, so we'd reject the $H_0$ of model 1. It fits the data poorly.

To describe and interpret the findings, you examine the observed and expected frequencies to see where and how they differ. In this case, there were more people cured than expected (24 vs. 15), so the treatment is effective relative to the baseline of model 1.

We can follow the same procedure to test model 2. Here's the calculation of $\chi^2$:

$$\chi^2 = ((24 - 20)^2 / 20) + ((6 - 10)^2 / 10) = 16/20 + 16/10 = 2.40$$

The critical value remains 3.84 because the $\alpha$ level and *df* are the same as for the test of model 1. In this case, 2.40 < 3.84, so we'd retain $H_0$. Model 2 fits the data well.

In addition to testing the statistical significance of one or more models, you can compare the relative fit of competing models by comparing their $\chi^2$ values. The lower the $\chi^2$, the better the fit. In this case, 2.40 < 10.80, so model 2 fits better than model 1.

One final note on calculating $\chi^2$ is important. In the examples shown above, all of the expected frequencies happened to be whole numbers. That will not always be true. When expected frequencies include fractions, you should not round them to whole numbers. If you must round at all, such as when doing hand calculations, you should retain at least a couple more decimal places for the expected frequencies than you'll use when you round the final value for $\chi^2$. Because we usually round to two decimal places for APA style, retaining four or five when doing calculations is a good idea.

## Using SPSS

To perform the $\chi^2$ goodness of fit test in SPSS, you need to enter the raw data, not the frequencies themselves. Enter your data into a single variable (column) using numerical codes to represent the categories; it makes no difference what numbers you use to represent the categories. To serve as an illustration, the data analyzed above are used. The variable is "Outcome", coded as 1 = cured and 2 = not cured. The full data set wouldn't fit onto the screen, and it's not shown even in part because all you'd see is a single column of numbers, with 24 rows containing a 1 and 6 rows containing a 2.

Next, you use the following command:

```
npar test
/chisquare = outcome
/expected = 15 15
```

To run the command, list the variable (here, "Outcome") on the second line and specify the expected frequencies on the third line. It's important to list the expected frequencies for the cells in the order that corresponds to your coding. The command shown above would test model 1 (equal expected frequencies). To test model 2, the command would be modified as follows:

```
npar test
/chisquare = outcome
/expected = 20 10
```

Notice that the expected frequencies are listed as "20 10" because the categories were coded as 1 = cured and 2 = not cured. The value of 20 corresponds to how many patients were expected to be cured, and 10 to how many patients were expected not to be cured.

SPSS will produce two tables that provide what you'll need. The first table shows the observed and expected frequencies, labeled as "Observed N" and "Expected N". Check to make sure that you listed the expected frequencies in the correct order. If there are statistically siginficant results, you'd compare the observed and expected frequencies to describe and interpret the findings. SPSS calculates the difference ($f_O - f_E$) for each cell

(labeled as "Residual") to help you interpret the results. The second table contains the $\chi^2$ value, the *df*, and the *p* value (labeled "Asymp. Sig."). Here are the results for model 1:

**Tretament Outcome**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| Cured | 24 | 15.0 | 9.0 |
| Not Cured | 6 | 15.0 | -9.0 |
| Total | 30 | | |

**Test Statistics**

| | Tretament Outcome |
|---|---|
| Chi-Square | 10.800[a] |
| df | 1 |
| Asymp. Sig. | .001 |

Here are the results for model 2:

**Tretament Outcome**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| Cured | 24 | 20.0 | 4.0 |
| Not Cured | 6 | 10.0 | -4.0 |
| Total | 30 | | |

**Test Statistics**

| | Tretament Outcome |
|---|---|
| Chi-Square | 2.400[a] |
| df | 1 |
| Asymp. Sig. | .121 |

## APA Style

To report the results of a $\chi^2$ goodness of fit test, the statistical information is presented in much the same way as for *z*, *t*, or *F*. You provide the name of the statistic, with *df* in parentheses, followed by the statistic's value and the *p* value. There is no widely accepted measure of effect size for $\chi^2$ goodness of fit tests. Instead, *N* is reported along with the *df* so that readers can consider the sample size when thinking about the size of the effect.

If the test is not statistically significant, you can state this simply. If the test is statistically significant, you need to explain the pattern of results. There are many ways to do this, and the goal is to help the reader understand how the observed and expected frequencies differed. In much the same way that you report the *M* and *SD* when comparing groups, you can report the frequencies (or percentages) for cells to help describe $\chi^2$ results. Here's how you might report the $\chi^2$ test of model 1:

Among 30 subjects diagnosed with specific phobias, 80% were cured by treatment with systematic desensitization, which differs statistically significantly from an expected 50% success rate, $\chi^2(1, N = 30) = 10.80$, $p = .001$.

Here's how you might report the $\chi^2$ test of model 2:

Data are consistent with the hypothesis that two-thirds of subjects diagnosed with specific phobias would be cured by treatment with systematic desensitization, $\chi^2(1, N = 30) = 2.40$, $p = .121$.

If you wanted to present the results for both models in a way that indicates you were interested in comparing their relative fit, it might look like this:

Among subjects diagnosed with specific phobias and treated using systematic desensitization, data are more consistent with the hypothesis that two-thirds would be cured, $\chi^2(1, N = 30) = 2.40$, than the hypothesis that one-half would be cured, $\chi^2(1, N = 30) = 10.80$.

## Problems

A total of 100 students are enrolled in an introductory psychology course, and the final grades of all students are categorized as As, Bs, Cs, Ds, or Fs as follows:

A = 32
B = 28
C = 22
D = 14
F = 4

There are at least two models that might fit these data. Model 1 is a flat grade distribution, meaning that the number of students in each grade category is equal. If 100 students split evenly into 5 grades, you'd expect 20 students to receive each grade.

1. If you perform a $\chi^2$ goodness of fit test for model 1, what null hypothesis will be tested?
2. Construct a table showing the observed and expected frequencies for model 1.
3. What is the *df* for this test?
4. What is the critical region for this test?
5. Calculate $\chi^2$.
6. Are the results statistically significant? How can you tell?
7. Examine the differences between the observed and expected frequencies in your table to help interpret the results. How would you describe what you see?
8. Write the results for this $\chi^2$ goodness of fit test in APA style.

* * *

This series of problems continues to use the grades for the same introductory psychology course as the previous series. Model 2 is the grade distribution for all introductory courses. The number of students who receive each grade equals the college average for all 100-level courses. Specifically, the expected frequencies are:

A = 30
B = 25
C = 20
D = 15
F = 10

9. If you perform a $\chi^2$ goodness of fit test for model 2, what null hypothesis will be tested?
10. Construct a table showing the observed and expected frequencies for model 2.
11. What is the *df* for this test?
12. What is the critical region for this test?
13. Calculate $\chi^2$.
14. Are the results statistically significant? How can you tell?
15. Examine the differences between the observed and expected frequencies in your table to help interpret the results. How would you describe what you see?
16. Write the results for this $\chi^2$ goodness of fit test in APA style.
17. Compare the findings for models 1 and 2. Write the results in APA style.

* * *

18. Using SPSS, enter the treatment data that served as the illustration in this chapter. Follow the instructions in the text for how to organize the data file and enter the commands to perform $\chi^2$ goodness of fit tests for models 1 and 2. Check that your output matches what's shown in the text.
19. Using SPSS, enter the grade distribution data used in the previous series of problems. Follow the instructions in the text for how to organize the data file and enter the commands to perform $\chi^2$ goodness of fit tests for models 1 and 2. Check that your results match what you found when you ran the test by hand.

**Problems 1 – 8 are due at the beginning of class.**

# 19. $\chi^2$ Test of Independence

## Overview

In this chapter, we'll explore how to determine whether two nominal variables are related to one another. The same $\chi^2$ statistic that's used to perform the goodness of fit test can also be used to perform a **test of independence**. This is very much like a correlation, but it operates on nominal rather than quantitative data.

## Performing the $\chi^2$ Test

To illustrate the $\chi^2$ test of independence, let's expand the treatment study from the previous chapter. Rather than having only 30 subjects all receive systematic desensitization for the treatment of specific phobias, suppose 75 subjects were randomly assigned to treatment conditions: systematic desensitization, psychodynamic therapy, or no treatment. At completion, each subject is scored as either cured of the phobia or not. We can use the $\chi^2$ test of independence to determine whether treatment condition is related to outcome.

The null hypothesis is no relationship between the two variables. That's why this is called a test of independence. Unlike the $\chi^2$ goodness of fit test, this test provides a unique set of expected frequencies to which the observed frequencies are compared. In other words, when you perform the $\chi^2$ test of independence, you'd don't have to figure out what expected frequencies to use.

Let's see how this is done. The first step is to arrange the observed frequencies into a table crossing the two variables. The categories for one variable form the columns, and the categories for the other variable form the rows:

| | Systematic Desensitization | Psychodynamic Therapy | No Treatment | Total |
|---|---|---|---|---|
| Cured | 24 | 12 | 4 | 40 |
| Not Cured | 6 | 13 | 16 | 35 |
| Total | 30 | 25 | 20 | 75 |

In addition to the observed frequencies for each cell in the table, totals are calculated for each column, each row, and the entire table, and these totals are placed in the margins. It's these marginal totals that allow us to calculate the expected frequencies.

For the expected frequencies to represent $H_0$, they have to exhibit no association between the two variables. For example, in this case we can see that overall, 40 out of 75 subjects were cured. This proportion, 40/75, would have to remain constant across treatment conditions for there to be no relationship between treatment and outcome. That means that out of the 30 subjects who received systematic desensitization, 40/75 of them would need to be cured. This quantity, $30 \times 40/75$, is the column total ($T_C$) multiplied by the row total ($T_R$) and divided by the overall total ($T$). This formula provides the expected frequency not only for this cell in the table, but for every cell in the table:

$f_E = T_R \times T_C / T$

To obtain the expected frequencies, we use this formula for each cell in the table:

Systematic Desensitization, Cured: $30 \times 40 / 75 = 16$

Systematic Desensitization, Not Cured: $30 \times 35 / 75 = 14$

Psychodynamic Therapy, Cured: $25 \times 40 / 75 = 13.3333$

Psychodynamic Therapy, Not Cured: $25 \times 35 / 75 = 11.6667$

No Treatment, Cured: $20 \times 40 / 75 = 10.6667$

No Treatment, Not Cured: $20 \times 35 / 75 = 9.3333$

Remember that you should not round expected frequencies. Once you've calculated them all, it's a good idea to double-check your work. The sum of the expected frequencies must equal *T*, the overall total, so add them together and check:

$16 + 14 + 13.3333 + 11.6667 + 10.6667 + 9.3333 = 75$

Next, let's include these expected frequencies in the table. You can place them in parentheses to distinguish them from the observed frequencies:

| | Systematic Desensitization | Psychodynamic Therapy | No Treatment | Total |
|---|---|---|---|---|
| Cured | 24 (16) | 12 (13.3333) | 4 (10.6667) | 40 |
| Not Cured | 6 (14) | 13 (11.6667) | 16 (9.3333) | 35 |
| Total | 30 | 25 | 20 | 75 |

To calculate $\chi^2$, we use the same formula as for the goodness of fit test:

$\chi^2 = \Sigma((f_O - f_E)^2 / f_E)$

Once again, you work through the expression in the outer parentheses for each cell and then you sum the results for all cells to get $\chi^2$. Here's what that looks like for these data:

$(24 - 16)^2 / 16 = 64 / 16 = 4.0000$
$(6 - 14)^2 / 14 = 64 / 14 = 4.5714$
$(12 - 13.3333)^2 / 13.3333 = 0.1333$
$(13 - 11.6667)^2 / 11.6667 = 0.1524$
$(4 - 10.6667)^2 / 10.6667 = 4.1667$
$(16 - 9.3333)^2 / 9.3333 = 4.7620$

$\chi^2 = 4.0000 + 4.5714 + 0.1333 + 0.1524 + 4.16667 + 4.7620 = 17.7858 = 17.79$

Notice that four decimal places were retained for all calculations, and only the final value of $\chi^2$ was rounded to two decimals for reporting in APA style. If you round off at earlier steps, the final answer might be incorrect.

To determine whether your result is statistically significant, you look up a critical value for $\chi^2$ in a table (e.g., the one in Appendix A). The critical value is based on the $\alpha$ level (usually .05) and the *df*, which is $(C - 1) \times (R - 1)$, where *C* is the number of columns and *R*

is the number of rows. In this case, using $\alpha = .05$ and $df = (3 - 1) \times (2 - 1) = 2$, we find that the critical value is $\chi^2 = 5.99$. This is always a nondirectional test, so you'd reject $H_0$ whenever the $\chi^2$ calculated for your data exceeds the critical value. Here, $17.79 > 5.99$, so we'd reject $H_0$. This means that there is a statistically significant association between treatment and outcome.

To describe and interpret the findings, you examine the observed and expected frequencies to see where and how they differ. In this case, systematic desensitization yielded more cures than expected, psychodynamic therapy yielded about as many cures as expected, and no treatment yielded fewer cures than expected.

## Effect Size

There is no widely accepted measure of effect size for the $\chi^2$ test of independence, except in one special case. If both variables are dichotomous, the frequencies can be organized into a $2 \times 2$ table and you can calculate $\phi$ (the phi coefficient). That's a type of correlation, and therefore it's also a measure of effect size with the usual rules of thumb (.10 = small, .30 = medium, .50 = large).

For a $2 \times 2$ table of frequencies, the $\chi^2$ test of independence can also be calculated. In fact, the $p$ value for $\phi$ and the $\chi^2$ test would be identical. These are equivalent statistics. If you've performed the $\chi^2$ test, you can convert the result into $\phi$ to report this as a measure of effect size:

$$\phi = \text{sqrt}(\chi^2 / N)$$

Keep in mind that this only applies to $2 \times 2$ tables. If there are more than two categories for either of the variables, $\phi$ cannot be calculated.

## Using SPSS

To perform the $\chi^2$ test of independence in SPSS, you need to enter the raw data, not the frequencies themselves. Enter your data into two variables (columns) using numerical codes to represent the categories for each; it makes no difference what numbers you use to represent the categories. To serve as an illustration, the treatment data analyzed above are used. The first variable is "Treatment", coded as 1 = systematic desensitization, 2 = psychodynamic therapy, and 3 = no treatment. The second variable is "Outcome", coded as 1 = cured and 2 = not cured. The full data set wouldn't fit onto the screen, but here's a portion that shows some variability in scores:

| Treatment | Outcome |
|---|---|
| 2.00 | 2.00 |
| 2.00 | 2.00 |
| 3.00 | 1.00 |
| 3.00 | 1.00 |
| 3.00 | 1.00 |
| 3.00 | 1.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |
| 3.00 | 2.00 |

Next, you use the following command:

```
crosstabs
/tables = outcome by treatment
/stats = chisq
/cells = count exp
```

To run the command, list the two variables (here, "Outcome" and "Treatment") on the second line. The first variable you list will form the rows in the table, and the second variable you list (after the word "by") will form the columns in the table. Reversing their order will not affect the statistical results.

SPSS will produce three tables, but you can ignore the first one (labeled "Case Processing Summary"). The second table shows the observed and expected frequencies for each cell, labeled as "Count" and "Expected Count".[48] The third table contains the test results. Use the first row of this table to find the $\chi^2$ value (labeled "Pearson Chi-Square"), the *df*, and the *p* value (labeled "Asymp. Sig. (2-tailed)"). The bottom row contains *N* (labeled "N of Valid Cases"). Here are the results for the test shown earlier:

[48] SPSS rounds expected frequencies to one decimal place in the output, but it uses many more decimals to perform calculations.

**Outcome * Treatment Condition Crosstabulation**

| | | | Treatment Condition | | | |
|---|---|---|---|---|---|---|
| | | | Systematic Desensitization | Psychodynamic Therapy | No Treatment | Total |
| Outcome | Cured | Count | 24 | 12 | 4 | 40 |
| | | Expected Count | 16.0 | 13.3 | 10.7 | 40.0 |
| | Not Cured | Count | 6 | 13 | 16 | 35 |
| | | Expected Count | 14.0 | 11.7 | 9.3 | 35.0 |
| Total | | Count | 30 | 25 | 20 | 75 |
| | | Expected Count | 30.0 | 25.0 | 20.0 | 75.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 17.786[a] | 2 | .000 |
| Likelihood Ratio | 18.981 | 2 | .000 |
| Linear-by-Linear Association | 17.523 | 1 | .000 |
| N of Valid Cases | 75 | | |

## APA Style

To report the results of a $\chi^2$ test of independence, the statistical information is provided just as it is for the $\chi^2$ goodness of fit test. If the test is not statistically significant, you can state this simply. If the test is statistically significant, you need to explain the pattern of results. There are many ways to do this, and the goal is to help the reader understand the nature of the association between the two variables. Here's how you might report the $\chi^2$ test shown above:

When subjects diagnosed with specific phobias were randomly assigned to treatment conditions, cure rates were statistically significantly different for systematic desensitization (80%), psychodynamic therapy (48%), and no-treatment control (20%), $\chi^2(2, N = 75) = 17.79, p < .001$.

Notice that percentages revealed the pattern of results. Percentages were calculated from the observed frequencies (e.g., for desensitization, 24 cured out of 30 = 80%).

## Problems

A total of 100 students, including 60 psychology majors and 40 students in other majors, are enrolled in an introductory psychology course. The final course grades of all students are categorized as As, Bs, Cs, Ds, or Fs. The observed frequencies are as follows:

| | | | | | |
|---|---|---|---|---|---|
| Psychology Majors: | A = 23 | B = 20 | C = 11 | D = 5 | F = 1 |
| Other Majors: | A = 9 | B = 8 | C = 11 | D = 9 | F = 3 |

1. If you perform a $\chi^2$ test of independence, what null hypothesis will be tested?
2. Construct a table showing the observed frequencies. Leave room in each cell to show the expected frequencies.
3. Calculate the expected frequency for each cell in the table.
4. What is the *df* for this test?
5. What is the critical region for this test?
6. Calculate $\chi^2$.
7. Are the results statistically significant? How can you tell?
8. Is there an appropriate measure of effect size for this test? If so, calculate it.
9. Write the results for this $\chi^2$ test of independence in APA style.

* * *

A total of 114 prisoners released on parole were classified by race (white vs. non-white) and high school diploma (no vs. yes). Among white individuals, 14 had their HS diploma and 14 did not. Among non-white individuals, 27 had their HS diploma and 59 did not.

10. If you perform a $\chi^2$ test of independence, what null hypothesis will be tested?
11. Construct a table showing the observed frequencies. Leave room in each cell to show the expected frequencies.
12. Calculate the expected frequency for each cell in the table.
13. What is the *df* for this test?
14. What is the critical region for this test?
15. Calculate $\chi^2$.
16. Are the results statistically significant? How can you tell?
17. Is there an appropriate measure of effect size for this test? If so, calculate it.
18. Write the results for this $\chi^2$ test of independence in APA style.

* * *

19. Using SPSS, enter the treatment data that served as the illustration in this chapter. Follow the instructions in the text for how to organize the data file and enter the commands to perform a $\chi^2$ test of independence. Check that your output matches what's shown in the text.
20. Using SPSS, enter the grade distribution data used in the first series of problems. Follow the instructions in the text for how to organize the data file and enter the commands to perform a $\chi^2$ test of independence. Check that your results match what you found when you ran the test by hand.
21. Using SPSS, enter the race and education data used in the second series of problems. Follow the instructions in the text for how to organize the data file and enter the

commands to perform a $\chi^2$ test of independence. Check that your results match what you found when you ran the test by hand.

**Problems 1 – 9 are due at the beginning of class.**

# 20. Selecting a Statistical Test

## Overview

A total of 15 statistical tests were introduced in this text. They can be grouped into a few broad types based on the kinds of research designs and data for which they're used. In this final chapter, we'll explore how to select the most appropriate statistical test to address a particular research question. A summary of the key decision points is provided in Appendix C.

## The Menu of Choices

The 15 statistical tests introduced in this text are listed below. To select the most appropriate test to address a particular research question, the key is to determine what kind of research design and data are involved. This chapter deal with ways to test for differences between means, test relationships between two or more variables, or test the goodness of fit between observed and expected frequencies.

1. One sample *z* test
2. One sample *t* test
3. Independent groups *t* test
4. Related samples *t* test
5. Independent groups ANOVA
6. Related samples ANOVA
7. Factorial ANOVA
8. Correlation (Pearson product-moment correlation)
9. Spearman rank-order correlation
10. Point-biserial correlation
11. Phi coefficient
12. Regression (simple linear regression)
13. Multiple regression
14. $\chi^2$ goodness of fit test
15. $\chi^2$ test of independence

## Testing Differences Between Means

The *z*, *t*, and *F* tests can all be used to compare means. One important clue as to whether the most appropriate test to address a particular research question is of this type is that these tests all require a quantitative (interval or ratio scale) dependent variable, or outcome measure, whose mean can be calculated for comparison. If you have nominal (qualitative) or ordinal (ranked) data only, none of these tests should be used.

### One Variable Whose *M* Will Be Compared to Population μ

When you want to compare the average scores on a single variable to a population average, either a **one sample *z* test** or a **one sample *t* test** should be used. Each requires that you specify a value for μ, the population mean, that serves as the null hypothesis. The test result will tell you whether *M*, the sample mean, is sufficiently far from μ that the difference is statistically significant.

To choose between *z* and *t*, all that you need to consider is whether or not you know σ, the population standard deviation. If you have this information, you use the *z* test. If not, you use the *t* test, which uses the sample standard deviation (*SD*) as an estimate of σ.

### Categorical Independent Variable(s)

When you want to compare means across a series of conditions in a study, either a *t* test or an ANOVA should be used. Each of these begins with a null hypothesis of no difference across conditions, and the test result will tell you whether sample means differ enough to be statistically significantly different.

If the research design involves a single between-subjects independent variable, meaning that subjects are divided into groups along one factor, you use an **independent groups *t* test** or an **independent groups ANOVA**. To choose between *t* and ANOVA, all that you need to consider is how many groups are being compared. If there are only two groups, you use the *t* test. If there are three or more groups, you use the ANOVA.

If the research design involves a single within-subjects independent variable, meaning that the same subjects are measured in all conditions or specific subjects are matched to one another and then assigned to conditions, you use a **related samples *t* test** or a **related samples ANOVA**. To choose between *t* and ANOVA, all that you need to consider is how many conditions are being compared. If there are only two conditions, you use the *t* test. If there are three or more conditions, you use the ANOVA.

If the research design involves two or more independent variables, you use a **factorial ANOVA**. This test can be used with any combination of between-subjects or within-subjects factors. For simplicity, this text only detailed the procedure for how to perform a factorial ANOVA with two between-subjects factors.

## Testing Relationships Between Two or More Variables

Correlation, regression, and the $\chi^2$ test of independence can all be used to test the relationship between two or more variables. Because means are not compared across conditions, the data need not be quantitative. Some of these tests can accommodate nominal or ordinal data.

### Association Between Two Variables

When you want to determine whether two variables are associated with one another, some type of correlation or the $\chi^2$ test of independence should be used. Each of these begins with a null hypothesis of no association between variables, and the test result will tell you whether the association observed in the sample is strong enough to be statistically significant.

To choose between the various correlational analyses, all that you need to consider is what types of data the two variables are. There are many possible combinations, and this text reviewed the five that are encountered most frequently.

If you have two quantitative variables, you use an ordinary **correlation** (aka Pearson product-moment correlation coefficient, symbolized $r$).

If you have two ranked variables, you use **Spearman's rank-order correlation** (symbolized $r_S$).

If you have one quantitative and one dichotomous variable, meaning that the latter identifies members of two groups, you use a **point-biserial correlation** (symbolized $r_{pb}$). This is equivalent to using an independent groups $t$ test, meaning that their $p$ values would be identical and you'd reach the same conclusion regarding your null hypothesis.

If you have two dichotomous variables, meaning that each variable identifies members of two groups, you use a **phi coefficient** (symbolized $\phi$).

If you have two nominal variables, meaning that each variable identifies members of two or more categories, you use the **$\chi^2$ test of independence**. The phi coefficient is a special case of this test, and they're equivalent when both variables are dichotomous.

### Making Predictions

When you want to use one or more variables to determine how, and how accurately, they predict scores on a single outcome variable, either **regression** or **multiple regression** should be used. Each of these begins with a null hypothesis of no predictive validity, or no variance in the outcome variable explained by the predictor(s), and the test result will tell you whether the predictive validity observed in the sample is strong enough to be statistically significant.

To choose between regression and multiple regression, all that you need to consider is how many predictor variables will be entered into the regression equation. If you have one predictor, you use **regression** (aka simple linear regression). If you have more than one predictor, you use **multiple regression**.

## Testing Goodness of Fit Between Observed and Expected Frequencies

One remaining statistical test is unique. The **$\chi^2$ goodness of fit test** doesn't involve a comparison of means across conditions or the association between two or more variables. Instead, this is used with a single categorical variable when you want to determine whether the observed frequencies, or counts, in the cells of a table differ sufficiently from the expected frequencies to be statistically significant. Identifying research questions for which this test is most appropriate should be very easy because it's the only test we've considered that involves a single categorical variable.

## Problems

For each of the following studies, indicate which of the 15 statistical tests introduced in this text should be used and explain why this is the most appropriate choice. Make sure to name the test fully (e.g., "$t$ test" is not sufficiently specific because there are three kinds, and if you mean "point-biserial correlation" you need to specify that rather than writing

only "correlation"). Also, make sure you justify your choice. An example of a well-justified choice would be "Related samples *t* test; the design is within-subjects and there were only two conditions."

1. A demographer working for the U.S. Census Bureau wants to compare salaries for urban vs. rural areas. She gets a sample of psychologists, some who live in urban areas and some who live in rural areas. Do earnings differ across these areas?
2. A cognitive psychologist wonders whether talking on a cell phone impairs the ability to concentrate while driving. An experiment is performed using a driving simulator, and subjects asked to drive a standard, challenging course under one of three randomly assigned conditions: (1) driving while holding and talking on a cell phone, (2) driving while talking on a hands-free phone, and (3) driving while talking with a passenger seated in the simulator. After a practice period to become accustomed to the task, the test begins and the dependent variable is whether a passing score is earned. Is there a difference in passing rates across experimental conditions?
3. First-year college students were surveyed about how much they liked their roommates at three points in time: within five minutes of meeting them, after the first week of classes, and at the end of the semester. Ratings were made on a 7-point Likert scale. Does degree of liking change over the course of the semester?
4. People are measured to determine how fair-skinned they are; this is assessed using a quantitative scale. A dermatologist then counts the number of suspicious moles on each person's skin. Is there a relationship between skin fairness and the number of suspicious moles?
5. A clinical psychologist wondered whether adults with attention deficit hyperactivity disorder (ADHD) had reflexes that differed in speed from those of the general population. She located a test of reaction time that was normed on adults in the U.S. ($\mu$ = 200 msec). From treatment centers in her home state, a random sample of 141 adults diagnosed with ADHD were tested for reaction time ($M$ = 220, $SD$ = 27). Do adults with ADHD differ in reaction time from the general population?
6. A large sample of adults living together in self-reported "committed relationships" (which includes, but is not limited to, marriage) in an urban area is studied to determine whether there is an association between the employment status of partners. Both members of each couple are classified independently as working full time or not working full time. Is there an association between the employment status of men and women in committed relationships?
7. Each child in the 4$^{th}$ grade at a large elementary school is classified by the teacher as predominantly right-handed, left-handed, or ambidextrous. The children's art teachers rate their artistic ability on a 10-point scale. Does artistic ability differ by handedness?
8. A clinical psychologist wants to use scores on a childhood behavior checklist to predict the severity of depression among young adults. A sample of children who were assessed for behavioral problems is followed over time. Among those who later seek counseling services for any mood disturbance, one of the measures that is administered assesses

their level of depression. How can scores on the childhood behavior checklist best be used to predict severity of depression?

9. A sociologist wanted to see if there was a relationship between a family's educational status and the eliteness of the college that their oldest child attended. She measured educational status by counting how many years of education the parents had received and she counted colleges that accepted fewer than one-third of their applicants as elite. Is there an association between family educational status and college eliteness?
10. A nutritionist wanted to find out if coffee and tea, as served in restaurants, differed in caffeine content. She went to 30 restaurants, ordered coffee and tea in each one, and had the caffeine content of each beverage tested. Do these servings of coffee and tea differ in caffeine levels?
11. Teenagers in a small community believe that the local police single them out for traffic stops more often than adult drivers. To investigate this, a researcher randomly selected six traffic tickets from each month in one year, for a total of 72 tickets. Because the age of the driver was recorded on the ticket, the investigator was able to determine that 11 tickets went to teen drivers and the other 61 tickets went to adults. According to the Department of Motor Vehicles, 8% of licensed drivers in the town are teenagers. Does the percentage of tickets given to teen drivers differ from the percentage of teen drivers?
12. A developmental psychologist is interested in the study of aggression. She observes aggressive behavior on school playgrounds to test for gender differences in both physical and verbal aggression. Does the level of aggression differ by gender, by type of aggression, or both?
13. A kinesiologist and a psychologist collaborated on a study to investigate the relationship between exercise and mental health in a random sample of adult men. Exercise was measured as the number of minutes of aerobic activity per week and mental health was measured using a self-report scale. The investigators noticed that whereas the distribution of exercise was positively skewed, the distribution of mental health better approximated normality. Because of this, they converted both quantitative variables to ranks. Is there an association between exercise and mental health?
14. Researchers assisted a large, metropolitan psychiatric hospital in predicting the length of stay of newly admitted patients. Among a sample of 800 patients that spent an average of 16.3 days in the hospital, a wealth of information was available. How well does a model that includes five variables (primary diagnosis of schizophrenia, primary diagnosis of mood disorder, secondary diagnosis of alcohol or drug problem, number of previous admissions, and age) predict the length of stay?
15. A scientific supply company has developed a new breed of lab rat, which it claims weighs the same as the classic white rat ($\mu = 485$ grams, $\sigma = 50$ grams). A researcher obtained a sample of 76 of the new breed of rats, weighed them, and found $M = 515$ grams. Is the company's claim true?

* * *

16. In 1997, Nabisco came out with a clever advertising campaign, the Chips Ahoy Challenge. Nabisco guaranteed that there were more than 1,000 chocolate chips in every bag, and they challenged consumers to count. Suppose 25 people go to the trouble of counting the chips in one bag apiece. Do their findings statistically significantly refute Nabisco's claim?
17. A developmental psychologist seeks to determine whether children's exposure to pets in the home affects the likelihood of keeping similar kinds of pets later in life. A sample of middle-aged adults is asked what kind of pet (if any) predominated in their home when growing up, using five categories: furry (dogs, cats, hamsters, etc.), finned (fish), feathered (birds), scaly (reptiles, amphibians), or none. They also indicate whether they have kept any pets of that category as adults. Is there an association between the kinds of pets kept as kids and as adults?
18. A clinical psychologist wanted to compare three treatments for Generalized Anxiety Disorder (GAD). She put an ad in the local paper to find people with GAD. Based on severity of symptoms, she matched the volunteers for her study into triads and randomly assigned each of the matched cases to one of the three treatments. Outcomes were assessed individually by a clinician blind to treatment assignments. Are the treatments equally effective?
19. An economist wants to predict how much an increase in the minimum wage will increase unemployment among low-skilled workers. He collects data on the minimum wage in different places and at different times, along with the corresponding unemployment rates among low-skilled workers. How much does an increase in the minimum wage increase the unemployment rate?
20. A dentist wanted to determine whether childhood fluoride supplements reduced the number of cavities. She took a sample of adults who were raised in regions without fluoride in the water supply, asked whether each had regularly taken fluoride supplements, and tallied the number of cavities in the dental record. Is there a relationship between fluoride supplements and cavities?
21. An investigator wonders whether the reduced mental alertness due to sleep deprivation can be counteracted by consuming caffeine. Three groups of volunteers are subjected to varying amounts of sleep deprivation (0 hours, 1 hour, or 2 hours). One-half of all volunteers is given a standardized dose of caffeine, the other half is not. Does mental alertness differ by sleep deprivation, caffeine intake, or both?
22. A behavioral therapist had patients with spider phobias rate the level of their fear on a 10-point scale. He then asked each patient, in turn, to enter a room with a spider in a cage and come as close to the spider as they felt comfortable. Do people with more self-rated fear stay a greater distance away from the spider?
23. A developmental psychologist wondered if birth order had an impact on academic performance. She found families with two children and obtained the high school GPA of each child. Is there a difference in GPA between first-born and second-born children?
24. A real estate agent wonders how accurately the selling price of homes can be predicted. She constructs a data set that includes the square footage of the home, the acreage of

the lot, the number of bedrooms, the number of bathrooms, and the average price of homes sold within the same development over the past two years. How accurately can the selling price of homes be predicted from these variables?

25. A psychologist wanted to investigate the relationship between the technical skill and creativity of children's drawings. A sample of kids provided drawings, and art teachers ranked these from highest to lowest in terms of technical proficiency at drawing and then from highest to lowest in terms of the creativity of artistic expression. Is technical skill related to creativity in drawing?

26. A conservation biologist wonders whether placement on the endangered species list improves the chances that a species will avoid extinction. He catalogues reptile species that were considered equally threatened 25 years ago, but among which one-half were subsequently placed on the endangered species list and the other half were not, and classifies the rate of population decline in its natural habitat as either sped up or slowed down. Is status as an endangered species associated with whether the decline sped or slowed?

27. A behavioral economist wonders whether portion sizes influence weight change. Rather than performing a one-shot experiment in the laboratory, she arranges for dining halls on three college campuses to systematically vary their portion sizes for one full semester. One serves small portions, another serves medium-sized portions, and the third serves large portions. Students who regularly eat in these dining halls are asked to weigh themselves at the beginning and the end of the semester, and their change in weight is the dependent variable. Does portion size affect weight change?

28. Across U.S. cities, the average vacancy rate for apartments is $\mu = 10\%$ ($\sigma = 4.6\%$). An urban studies major obtained a sample of 15 rust-belt cities and found that the average vacancy rate was $M = 13.3\%$. Does the vacancy rate for these cities differ from the U.S. average?

29. In Los Angeles County, members of grand juries serve for a one-year term and are paid at the rate of $25/working day. Individuals are either self-nominated or nominated by judges. Twenty-three percent of all citizens that are eligible to serve on the grand jury are Hispanic. Of 144 nominees in a given year, only 8 were Hispanic. Does this represent evidence of racism?

30. An exercise physiologist classifies people—on the basis of their body mass index, heart rate, and lung capacity—as above or below average in terms of fitness. He then directs the same people to walk on a treadmill, individually, at an increasing speed until they can no longer walk. The speed when a person maxes out is the dependent variable. Is there a difference in maximum walking speed based on fitness level?

* * *

31. Individuals who behave in hard-driving, competitive, and ambitious ways have been described as exhibiting the Type A personality, in contrast to the Type B personality that's characterized as more relaxed and easy-going. When people classified as Type A or Type B are confronted with an experimental task designed to induce frustration, do they experience different levels of frustration?

32. A researcher asks male and female volunteers to describe their most recent dream. Each dream is rated by an expert as low, medium, or high in aggressive content. Do men have more aggressive dreams than women?
33. A manufacturer of computer components is trying to improve its keyboards. A sample of 12 administrative assistants spends one hour typing on each of six newly designed keyboards. The performance of each keyboard is rated on a 7-point scale from "very poor" to "very good". Are there systematic differences in the performance ratings of the keyboards?
34. Members of married couples complete a questionnaire that measures how liberal or conservative their attitudes are. Do the data support the notion that similarities in attitudes are important for interpersonal attraction?
35. Nationwide, 5$^{th}$ graders achieve $\mu = 70$ on a standardized test of reading achievement. A particular teacher notices that for the 25 students in her class, $M = 75$. Does this suggest that her students are reading better than average?
36. A researcher hypothesizes that a particular chemical contained in the urine of male rats affects the behavior of other males—but not females—in the colony, specifically that it makes them more anxious. To test this hypothesis, the investigator measures the activity levels of male and female rats that are each placed, alone, into a cage that's either sterile or painted with the chemical extracted from male rat urine. Is the researcher's hypothesis supported by the data?
37. An observer visits several large lecture halls across all the major departments at a college over a period of one week. Each laptop computer that's in use during class time is coded as being a Mac or a PC. Are students' preferences equally divided between these types of computer?
38. An instructor records students' scores on both a midterm and a final exam. How accurately do midterm exam scores predict final exam scores?
39. The Pepsi Challenge, which began in the 1970s, involves a blind taste test between Coke and Pepsi colas. A researcher wonders whether preferences depend on age. One hundred people who complete the Pepsi challenge report their age, and whether they each preferred Coke or Pepsi is revealed and recorded. Is age related to cola preference?
40. For a science fair project, a child exposes newly sprouted bean plants to one of four types of music for several hours each day: classical, pop, rap, or country. The height of each plant is measured after two weeks. Does type of music affect plant growth?
41. Prior to entering the NFL draft, college football players are assessed on a series of seven physical tests that include the 40-yard dash, bench press, and vertical jump (this does not include position-specific drills and physical measurements). One measure of success in the NFL is being a starting player during the first year on a team. How well do these seven measures predict the number of games started in the first year?
42. College players who hope to be drafted by NFL teams also complete the Wonderlic, a brief intelligence test. An investigator ranks the career success of 24 quarterbacks who

played in the NFL. Is success associated with the rank-ordering of Wonderlic scores for these 24 quarterbacks?

43. Over the past 30 years, an average of $\mu = 12$ batters per week are hit by wild pitches in MLB games ($\sigma = 3$). For a sample of $n = 8$ weeks with unusually hot weather, the weekly average was $M = 15.5$. Does hot weather affect the likelihood of being hit by a pitch?

44. The chair of a committee formed to create a new licensing exam wonders whether examinees with severe test anxiety will have greater difficulty passing a written test. A sample of 60 examinees includes equal numbers of individuals who score at very low or very high levels on a measure of test anxiety. Is test anxiety associated with whether the examinees pass or fail the written test?

45. A drop in the number of white blood cells (lymphocytes) in the blood is associated with an increased susceptibility to disease. Lymphocyte counts are taken for a sample of men before and during a period of emotional distress. Does this form of stress decrease the number of white blood cells?

**Problems 1 – 15 are due at the beginning of class.**

# 21. Reproducibility

## Overview

The publication of scientific research in a peer-reviewed scholarly journal lends a seal of approval to the quality of the work and the credibility of the findings. To some extent this confidence is well earned, as editors and peer reviewers generally do a good job of detecting problems and only publishing studies that meet reasonable quality-control standards. However, this process depends on the integrity of authors to report their research fully and honestly as well as the integrity of editors and reviewers to assess it carefully and free of bias. Even at the most prestigious and selective journals, there is no guarantee that published findings are correct. In recent years, psychological scientists have been paying greater attention to the **reproducibility** of research by examining how often findings can be replicated and what can be done to prevent mistaken conclusions from being published. In this chapter, we'll learn about questionable research practices that can lead to false findings being published as well as ways to reduce the chances of this happening.

## False Findings

In 2005, John Ioannidis published a paper that stirred up some controversy, to put it mildly. It was titled "Why Most Published Research Findings Are False".[49] That's a pretty bold claim, and Ioannidis meant exactly what he said. Though his article appeared in a medical journal, he was not criticizing any particular scientific field. Rather, he was pointing out just how weak statistical evidence can be.

Many investigators believe that simply using $\alpha = .05$ provides good protection against Type I errors, but Ioannidis showed that this is mistaken. The actual probability of reaching a false-positive conclusion—of mistaking results that can be explained by chance for evidence of a systematic effect—is often surprisingly high, exceeding 50% much of the time. Ioannidis made his case using simple math and argued that this conclusion holds true across a wide range of plausible assumptions regarding research contexts. He also discussed factors that would increase the likelihood of false findings, such as using smaller sample sizes; studying phenomena with smaller effect sizes; allowing greater flexibility in designs, definitions, outcomes, and analyses; having financial and other interests in certain results, or holding prejudices; and studying hotter topics, with more scientific teams involved.

Ioannidis sounded the alarm using probabilities, and his calculations required making some assumptions. Those uncomfortable with his assumptions, or uninterested in hypothetical calculations, might have found this an easy argument to dismiss. Before long, the concerns he expressed would be hard to ignore.

[49] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* 696-701.

## Replication Crisis

Everyone learns in their research training that replication is a cornerstone of the scientific method. We know that a single finding may be mistaken and that independent corroboration in follow-up studies builds trust that the finding is real. Unfortunately, it's relatively rare to perform replication studies that put this to a test. There are many possible reasons for this, most of which involve greater interest in and reward for generating original ideas than for attempting to replicate prior research. The "publish or perish" culture of scientific research exerts a strong pressure on investigators to be productive in ways that are valued by journal editors, reviewers, and members of committees that make decisions about hiring, tenure, promotion, grant support, and professional honors. Replication has not been valued nearly as highly as original research.

As a consequence, in psychology as in most competitive scientific fields, replication has historically been given little attention. In part because of some high-profile cases of scientific misconduct[50], however, researchers have recently become increasingly curious about how many of the findings in the published literature are trustworthy. A project by the Open Science Collaboration[51] sheds light on this question. Research teams replicated 100 studies published in three leading journals in psychology. They followed all of the methods used in the original studies as closely as possible. The authors of the report on this massive project found that effect sizes in the replication studies were about one-half as large, on average, as those in the original studies. Whereas 97% of the original findings were statistically significant, only 36% were in the replication studies. Based on a variety of criteria for evaluating the outcomes, the authors concluded that fewer than one-half of the original findings had been replicated successfully. Though some critics have argued that this project underestimates reproducibility to some extent, it still underscores the possibility that Ioannidis was on to something, namely that a lot of published research findings may be false.

## Questionable Research Practices

How can so many bright, hard-working scientists publish results that are in fact mistaken? Actually, there are a lot of ways this can happen. Outright fraud might account for a few instances, but the far more common culprits are likely to be a number of biases in how research is designed, how data are analyzed, and how results are reported. There are so many choices that investigators have available that if they take advantage of this flexibility, this can greatly increase their chances of finding something that seems interesting. For example, to demonstrate just how easy it is to generate apparent support even for false hypotheses, one research team reported findings from a pair of studies that

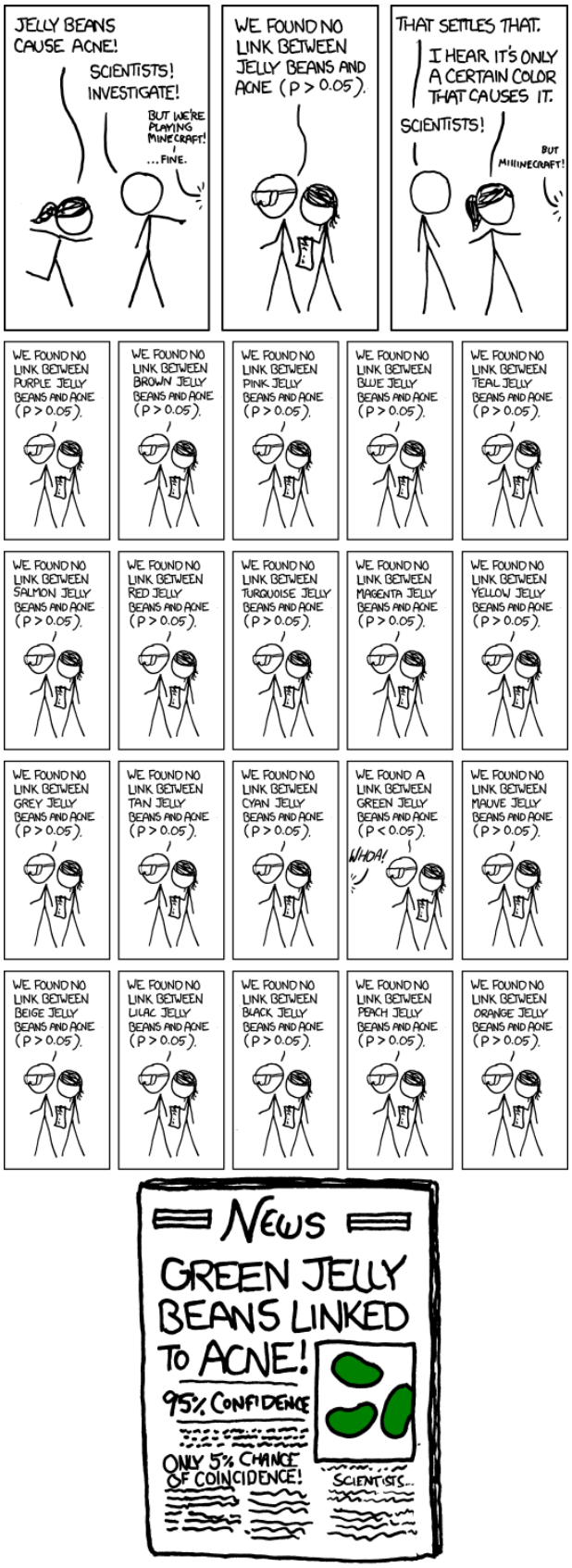
---

[50] For example, Karen Ruggiero (former psychology professor at Harvard University) fabricated data on gender and discrimination and retracted at least two published articles. Brian Wansink (former behavioral economist at Cornell University) committed a variety of types of academic misconduct and has retracted 18 published articles and corrected 15 papers, with concerns remaining about many others. In perhaps the best-known case in the social sciences, Diederik Stapel (former social psychology professor at Tilburg University) fabricated data in dozens of studies and has retracted 58 publications.

[51] Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716.

are hard to believe.[52] First, they showed that listening to a children's song made people feel older. That would be surprising, but it could be true. Second, they showed that listening to a song about older age makes people younger. Not how they feel, but their actual age—music had reversed the aging process!

You're probably sufficiently skeptical that you suspect something is wrong with this research. What the investigators did was to intentionally take advantage of what they called **researcher degrees of freedom** by, for example, collecting and analyzing much more data (e.g., experimental conditions, dependent variables) than they reported. They presented only a few results that appeared to support the hypotheses. The authors demonstrated that if you collect enough data and try enough analyses, eventually you're bound to find something you like, something that will make for an interesting report. Consider this comic by Randall Monroe at xkcd.com, for example.




⇒

Testing each color of jelly bean is like collecting as many dependent variables as possible in a study and performing separate statistical tests for all of them. Running a lot of tests, using many different variables, can increase the chances of obtaining at least one statistically significant result. This makes it easier to publish your research, but it also increases the chances of false-positive findings entering the literature. As noted in the context of ANOVA models, when you make multiple comparisons you increase the experimentwise Type I error rate. It's highly problematic to focus attention only on the statistically significant finding(s) without noting how many tests were performed.

The idea that actual working scientists might test 20 different colors of jelly beans,

but then headlines would emphasize the single statistically significant finding might seem far-fetched. Surely research this poorly done wouldn't survive the peer-review process at a reputable journal? Surely journalists reporting on scientific discoveries would know better than to isolate one finding and ignore the many failed tests? Unfortunately, as noted earlier, even professionals aren't 100% reliable in these regards. For example, consider these actual headlines from the fall of 1999:

"Heart Patients Fared Better after Secret Prayers"
*Toronto Star*, October 26

"Prayer's 'Medicinal' Value Gets an Amen from Study"
San Diego *Union-Tribune*, November 3

"Scientists 'Prove the Power of Prayer'"
London *Daily Telegraph*, November 11

This study received a lot of media attention, and the message was clear and consistent: Scientific data support the effectiveness of prayer as health care. What evidence supported these bold claims?

A team of researchers[53] randomly assigned 990 patients in a coronary care unit to a treatment group that received prayers for their swift recovery or a no-prayer control group. Thirty-five health outcomes were recorded for all patients, including pneumonia, major surgery, cardiac arrest, or death. Because the researchers used $\alpha = .05$ for all tests, sampling error alone would be expected to yield 1 or 2 statistically significant results for their 35 tests. And that's just what they found: The only significant difference between groups was that patients in the prayer condition had better "Swan-Ganz catheter" ratings. Despite the very strong possibility that these results represent nothing more than sampling error, this study was published in a major medical journal. Though the authors themselves are careful to state that "we have not proven that God answers prayer or that God even exists" (Harris et al., 1999, p. 2277), the headlines quoted earlier show that the media proclaimed prayer to be an effective remedy for disease. The evidence was basically the same as that linking green jelly beans to acne.

Performing a lot of tests but focusing attention only on the favorable results—counting the hits and ignoring the misses—is among a number of common research practices that cross the line between rigorously testing a hypothesis and fishing for support in ways that boost the rate of false findings. A study of psychological scientists[54] found that many admitted to engaging in **questionable research practices** such as failing to report all dependent variables (63%) or experimental conditions (28%), deciding whether to collect more data based on whether results were significant (56%), "rounding off" a *p* value that's actually above an $\alpha$ level to make it appear significant (22%), and selectively reporting studies that "worked" (46%). Reported levels of engaging in these questionable research

[52] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22,* 1359-1366.
[53] Harris et al (1999). A randomized, controlled trial of the effects of remote, intercessory prayer on outcomes in patients admitted to the coronary care unit. *Archives of Internal Medicine, 159,* 2273-2278.
[54] John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23,* 524-532.

practices increased when the anonymous respondents were given stronger incentives to be honest and when they were asked about their general prevalence among all scientists.

## Academic Culture

Questionable research practices emerge within research communities. Aspects of the academic culture that shapes these communities, especially those involving the publication of scientific papers, can either encourage or discourage questionable research practices.

### Publish or Perish

The "publish or perish" nature of academic culture is at least partly responsible for the popularity of questionable research practices. Publishing more papers, especially in reputable journals, leads to professional success. Those whose publication records fall short are less likely to be hired, tenured, promoted, or granted honors or awards. The constant pressure to be publishing new papers propels the search for statistically significant results. As we have seen, seeking statistical significance can compromise scientific integrity at many stages of the research process. In this way, the natural inclination to seek professional success can pave the way for questionable research practices that increase the likelihood of false findings.

### Peer Review

When a paper is submitted for publication in a scientific journal, an editor is tasked with deciding whether to accept the paper, reject it, or require revisions such as changing the text, performing new analyses, or perhaps even collecting new data. The editor will ask a few scientists with relevant expertise to read the submitted work, write a review that objectively evaluates its strengths and weaknesses, and offer advice to help reach a decision about publishing the paper. This process generally works well. When questionable research practices are identified as cause for concern, this helps to prevent the publication of false findings. However, reviewers are only human, meaning that they respond to incentives and have their own limitations and biases. Several implications of this observation lead to concerns about peer review as a mechanism of quality control.

First, reviewers are not paid for this work. Instead, this is volunteer activity expected of all those who participate in the field of study. As a consequence, scholars may not be as devoted to conscientious reviewing as they are to other aspects of their jobs. It's easier to read a paper quickly and write some fairly superficial comments than to read it very carefully and write a detailed, technically sound review. Without being paid for their time and effort, individuals may not work as hard as possible, and questionable research practices may be missed.

Second, the default procedure at most journals is for peer reviews to be provided anonymously. The editor who solicits these reviews will know the reviewers' identities, but the authors of the work being evaluated will not. This enables reviewers to be honestly and constructively critical without fear of reprisal from authors, but it also weakens incentives to be conscientious. There is very little professional recognition for work that is done anonymously. Again, this can lead to reviews that overlook questionable research practices.

Third, even dedicated reviewers can and do make honest mistakes, failing to notice important problems or falsely identifying acceptable practices as problematic. Particularly

when questionable research practices involve things that are hidden (e.g., a large number of statistical tests are run but only those that yield statistically significant results are reported), it will be easy for them to be missed even by conscientious reviewers.

Fourth, reviewers tend to be less critical of articles submitted for publication when they like the authors' conclusions. We all have soft spots, or blind spots. If biases are idiosyncratic (e.g., some people value experimental research especially highly whereas others value correlational research even more highly), they can be identified and discounted or they will tend to cancel out. Either way, a reasonable appraisal of a paper can be reached by obtaining reports from several peer reviewers. However, shared biases can cause substantial problems. To the extent that reviewers tend to share the same biases as authors, important flaws in research can be overlooked. This means that questionable research practices might go unnoticed, with reviewers recommending publication of work despite important flaws. For example, there appears to be a widely-shared, left-leaning political bias among social psychologists.[55] This makes it easier to publish social psychological research when the findings flatter liberals, even when questionable research practices have been followed. Among other problems, this raises the risk of false findings being published despite the approval of peer reviewers.

Fifth, on a related note, one type of bias is shared by all or most members of a research community: The desire to continue publishing their own work. It's unlikely that reviewers will call into question methods they use in their own research. Even if everyone recognizes that a practice is suboptimal and that better alternatives exist, it may be persist because it makes investigators' lives easier. For example, using introductory psychology students as research subjects helps investigators collect data cheaply and fairly quickly. The limitations of such samples are well known (e.g., undergraduate students enrolled in psychology courses may not be representative of the population to which one would like to generalize the findings, the quality of data provided by students required to participate in research may be poor), but reviewers who also like to rely on unpaid convenience samples may not want to point out these limitations very forcefully, if at all. The same goes for data collected online (e.g., through Amazon's Mechanical Turk, or MTurk). Though it costs money, it has greatly streamlined the data collection process for many investigators. That's great, provided the data are of sufficiently high quality for scientific research, but unfortunately the bar for accepting online data may have been set extremely low (e.g., responses were shown to be as reliable as those of introductory psychology students). The point being made here is not that any particular practice is necessarily objectionable. Many research questions can be addressed effectively using introductory psychology students as subjects or by collecting data online. Rather, the issue is that authors may be given a free pass on some questionable research practices if the reviewers also tend to engage in them. The pressure for everyone to continue publishing their work constitutes a shared bias within a community of scholars that can weaken the safeguards against publishing false findings.

[55] Jussim, L., Crawford, J. T., Anglin, S. M., & Stevens, S. T. (2015). Ideological bias in social psychological research. In *Social psychology and politics* (pp. 107-126). Psychology Press.

## Potential Remedies

The picture that emerges is one of researchers frequently engaging in questionable research practices with the goal of obtaining statistically significant results, which has come to be known as ***p* hacking** and leads to an alarmingly high rate of false findings in the published literature. What can be done about this? There are many possible remedies.

### Replication

The standard solution to the problem of false findings is replication. In practice, however, we have seen that this doesn't work as well as one might hope. The publish-or-perish culture of academic research rewards productivity, which is usually assessed through the frequency and impact of scholarly publications. Replication studies are more difficult to publish because they are inherently less interesting, and usually less impactful, than original research. Thus, expecting scientists to perform replication studies because their philosophy of science suggests they should may be unrealistic. Researchers are not the only professionals whose philosophical principles crash on the shores of self-interest.

On the bright side, however, some steps are being taken to encourage replication research. Agencies or foundations supporting research have targeted some of the available funding specifically for replication studies. Journals sometimes solicit papers reporting the results of replication studies. Particularly as more prestigious journals devote space to replication research, this will demonstrate its value and provide a compelling incentive for investigators to undertake such work. At least in some fields of study, journals frequently require that papers contain multiple studies to provide evidence of successful replication within a program of research. Another idea is to require that undergraduate or graduate students perform replication studies as part of their training in research methods.

### Clearly Label Exploratory Research

Research can be designed either to develop ideas or to test them, and it is usually not possible to address both of these goals in a single study. The goal of exploratory research is to develop ideas. This is done by collecting information on a wide range of variables, perhaps with very little experimental control, and then examining the data in many ways to search for interesting patterns. One need not have any hypotheses to perform useful exploratory research, nor is strong evidence required to raise the possibility that observed trends may be worthy of further testing. Findings are tentative, and follow-up research is needed to replicate and better understand them.

The goal of hypothesis-testing research, in contrast, is to subject ideas to rigorous tests. This is done by designing a study such that the evidence will either support or refute a hypothesis. This entails careful experimental control, collecting a large sample of data, performing demanding statistical tests, or other techniques to help rule out alternative explanations for results.

If the last two paragraphs sound familiar, that's because they were copied from the beginning of Chapter 1. They bear repeating here, in the context of reproducibility. There's nothing wrong with doing exploratory research as long as it's clearly labeled. However, it's easy to blur the lines between exploration and hypothesis-testing, and doing so can easily produce false findings. When authors pretend that they had hypothesized the results all along, this is known as **HARKing** (short for Hypothesizing After the Results are Known).

Misrepresenting exploratory research as hypothesis-testing by introducing the study with hypotheses and selectively presenting only those results that support them can lead to unwarranted confidence in the findings.

There are many reasons why HARKing can be tempting. For example, journals often impose strict limits on the length of research reports, so trimming anything not related to statistically significant findings helps to stay within the allowed space. Moreover, a paper that begins with hypotheses and ends with supportive evidence tells a more compelling "story" than a paper that describes a wide range of variables, tested in many ways, that led to a few statistically significant results. Regardless of the motivation, it is ethically irresponsible to disguise exploratory findings as the results of hypothesis-testing research.

The honest way to avoid this problem is to be explicit, throughout the research process, about when ideas are being developed and when they are being tested. The same data should never be used for both purposes. You should keep this in mind when designing a study, analyzing the data, interpreting the results, and reporting the findings.

## Registered Reports

Some journals offer investigators the opportunity to submit a **registered report**. The way this works is that a proposal describing the theory, hypotheses, methods, and planned analyses is subjected to peer review before a study is conducted. If the proposal is accepted and the authors follow the protocol they specified in advance, the paper will be published regardless of the results. Journal editors and reviewers are just as susceptible to confirmation bias as anyone else, and they tend to raise more objections to research that reports findings inconsistent with their preferred beliefs. Using registered reports avoids that problem by evaluating only the theory and methods of proposed research, not its outcomes. Registered reports also reduce or eliminate many of the incentives for *p* hacking or HARKing. When statistically significant findings are no longer a prerequisite for publication, there is less temptation to take advantage of researcher degrees of freedom to obtain them. Though registered reports are not yet common, those journals that do allow them encourage investigators to focus on theory and methods to design worthwhile research, letting the chips fall where they may when it comes to the results.

## Open Science

Whether or not registered reports are used, following guidelines for **open science** can reduce the likelihood of false findings. This entails making science as transparent as possible to enable outsiders to check findings, and there are many ways to do this. For example, one might post a detailed research plan before beginning a study to enhance the credibility of subsequent findings. Knowing what was hypothesized and how it would be tested reduces concerns about HARKing or *p* hacking. Another way to make science more open is to archive research reports, raw data, code used to perform analyses, or other materials in a public repository such as the Open Science Framework offered by the Center for Open Science. The more information is provided, the easier it becomes for others to examine the methods and results of the study for themselves and identify any questionable research practices that might call into question the findings.

### Make and Follow Plans

Even when investigators are not submitting a registered report or adopting an open science approach to documenting their research, they will reduce the likelihood of false findings if they devise a plan for the research and then follow the plan. The more carefully a study is crafted, the more detailed the hypotheses and methods (including the planned analyses), and the more scrupulously these plans are followed, the less opportunity there will be for HARKing or *p* hacking and the more trustworthy the end results.

### Report Comprehensively

Finally, it is important to be comprehensive when reporting research. This means listing all experimental conditions, all variables collected, all analyses performed, and all results obtained. Fully describing the method includes the criteria for removing any data as well as the rationale for the data analysis plan. The more variables are collected, the more ways there are to perform analyses, especially multivariate analyses that statistically control for the influence of one or more variables before examining the relationships between key independent and dependent variables. To interpret results correctly, it is critical to know how many analyses were performed, why each one was done, and what the results were. Space constraints may make it difficult or impossible to include all of this information in the text of an article, but footnotes and online supplementary material can be used to direct interested readers to the comprehensive reporting of all essential details.

## Problems

1. One foundation of the scientific method is that findings can be trusted when they are successfully replicated, ideally independently by researchers unaffiliated with the original investigators.
    a. Why is it important to replicate findings?
    b. Why do investigators seldom attempt to replicate their own or others' findings?
2. In a scientific research report, the method section should read like a recipe that someone else could follow to repeat the study as closely as possible. Even if these instructions are thorough and clear, when a new investigator follows them there is no guarantee that the original findings will themselves be reproduced. What are two distinct reasons why not?
3. In terms of research design and data analysis, what are the differences between exploratory research and hypothesis-testing research? Why is it important to clearly identify exploratory research to reduce the risk of false findings?
4. What are HARKing and *p* hacking, and in what ways are they similar and different from one another?
5. What are registered reports and open science, and in what ways are they similar and different from one another??
6. For several decades, there has been a debate about whether clinical psychologists should seek prescription privileges, the legal authority to prescribe medications. Many

psychologists would like to expand graduate training in clinical psychology to add this mode of treatment, whereas many others believe that those who wanted to prescribe medication should attend medical school, as psychiatrists do. The arguments on both sides are numerous and complex. In an article published in the journal *Professional Psychology: Research and Practice*, Antonuccio, Danton, and DeNelsky (1995) contributed to this debate by arguing that psychotherapy, not medication, should be the treatment of choice for depression.

a. How does the fact that this article was published lend greater credibility to the authors' argument?

b. Despite publication in a scholarly journal, why might it be appropriate to approach the authors' argument with some skepticism?

7. Simmons, Nelson, and Simonsohn (2011, p. 1360) reported two studies that "were conducted with real participants, employed legitimate statistical analyses, and are reported truthfully. Nevertheless, they seem to support hypotheses that are unlikely (Study 1) or necessarily false (Study 2)." The authors performed these studies to demonstrate a variety of factors that can lead to false findings. Read their report (quoted below), making sure you consult the footnote explaining what ANCOVA is.

***Study 1: Musical Contrast and Subjective Age***

In Study 1, we investigated whether listening to a children's song induces an age contrast, making people feel older. In exchange for payment, 30 University of Pennsylvania undergraduates sat at computer terminals, donned headphones, and were randomly assigned to listen to either a control song ("Kalimba," an instrumental song by Mr. Scruff that comes free with the Windows 7 operating system) or a children's song ("Hot Potato," performed by The Wiggles).

After listening to part of the song, participants completed an ostensibly unrelated survey: They answered the question "How old do you feel right now?" by choosing among five options (*very young*, *young*, *neither young nor old*, *old*, and *very old*). They also reported their father's age, allowing us to control for variation in baseline age across participants.

An analysis of covariance (ANCOVA) revealed the predicted effect: People felt older after listening to "Hot Potato" (adjusted $M = 2.54$ years) than after listening to the control song (adjusted $M = 2.06$ years), $F(1, 27) = 5.06$, $p = .033$.

In Study 2, we sought to conceptually replicate and extend Study 1. Having demonstrated that listening to a children's song makes people feel older, Study 2 investigated whether listening to a song about older age makes people *actually* younger.

***Study 2: Musical Contrast and Chronological Rejuvenation***

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba." Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/ yyyy) and their father's age. We used father's age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted $M$ = 20.1 years) rather than to "Kalimba" (adjusted $M$ = 21.5 years), $F(1, 17) = 4.92$, $p = .040$.

Which of the questionable research practices described in this chapter might have allowed the investigators to take advantage of researcher degrees of freedom and thereby produce false findings?

8. What steps could have been taken to prevent the (likely) false findings in these two studies?

**Problems 1 – 6 are due at the beginning of class.**

# Appendix A: Statistical Tables

## Unit Normal Table

**Area in one tail for *z***

| | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **.00** | .50000 | .49601 | .49202 | .48803 | .48405 | .48006 | .47608 | .47210 | .46812 | .46414 |
| **.10** | .46017 | .45620 | .45224 | .44828 | .44433 | .44038 | .43644 | .43251 | .42858 | .42465 |
| **.20** | .42074 | .41683 | .41294 | .40905 | .40517 | .40129 | .39743 | .39358 | .38974 | .38591 |
| **.30** | .38209 | .37828 | .37448 | .37070 | .36693 | .36317 | .35942 | .35569 | .35197 | .34827 |
| **.40** | .34458 | .34090 | .33724 | .33360 | .32997 | .32636 | .32276 | .31918 | .31561 | .31207 |
| **.50** | .30854 | .30503 | .30153 | .29806 | .29460 | .29116 | .28774 | .28434 | .28096 | .27760 |
| **.60** | .27425 | .27093 | .26763 | .26435 | .26109 | .25785 | .25463 | .25143 | .24825 | .24510 |
| **.70** | .24196 | .23885 | .23576 | .23270 | .22965 | .22663 | .22363 | .22065 | .21770 | .21476 |
| **.80** | .21186 | .20897 | .20611 | .20327 | .20045 | .19766 | .19489 | .19215 | .18943 | .18673 |
| **.90** | .18406 | .18141 | .17879 | .17619 | .17361 | .17106 | .16853 | .16602 | .16354 | .16109 |
| **1.00** | .15866 | .15625 | .15386 | .15151 | .14917 | .14686 | .14457 | .14231 | .14007 | .13786 |
| **1.10** | .13567 | .13350 | .13136 | .12924 | .12714 | .12507 | .12302 | .12100 | .11900 | .11702 |
| **1.20** | .11507 | .11314 | .11123 | .10935 | .10749 | .10565 | .10383 | .10204 | .10027 | .09853 |
| **1.30** | .09680 | .09510 | .09342 | .09176 | .09012 | .08851 | .08691 | .08534 | .08379 | .08226 |
| **1.40** | .08076 | .07927 | .07780 | .07636 | .07493 | .07353 | .07215 | .07078 | .06944 | .06811 |
| **1.50** | .06681 | .06552 | .06426 | .06301 | .06178 | .06057 | .05938 | .05821 | .05705 | .05592 |
| **1.60** | .05480 | .05370 | .05262 | .05155 | .05050 | .04947 | .04846 | .04746 | .04648 | .04551 |
| **1.70** | .04457 | .04363 | .04272 | .04182 | .04093 | .04006 | .03920 | .03836 | .03754 | .03673 |
| **1.80** | .03593 | .03515 | .03438 | .03362 | .03288 | .03216 | .03144 | .03074 | .03005 | .02938 |
| **1.90** | .02872 | .02807 | .02743 | .02680 | .02619 | .02559 | .02500 | .02442 | .02385 | .02330 |
| **2.00** | .02275 | .02222 | .02169 | .02118 | .02068 | .02018 | .01970 | .01923 | .01876 | .01831 |
| **2.10** | .01786 | .01743 | .01700 | .01659 | .01618 | .01578 | .01539 | .01500 | .01463 | .01426 |
| **2.20** | .01390 | .01355 | .01321 | .01287 | .01255 | .01222 | .01191 | .01160 | .01130 | .01101 |
| **2.30** | .01072 | .01044 | .01017 | .00990 | .00964 | .00939 | .00914 | .00889 | .00866 | .00842 |
| **2.40** | .00820 | .00798 | .00776 | .00755 | .00734 | .00714 | .00695 | .00676 | .00657 | .00639 |
| **2.50** | .00621 | .00604 | .00587 | .00570 | .00554 | .00539 | .00523 | .00508 | .00494 | .00480 |
| **2.60** | .00466 | .00453 | .00440 | .00427 | .00415 | .00402 | .00391 | .00379 | .00368 | .00357 |
| **2.70** | .00347 | .00336 | .00326 | .00317 | .00307 | .00298 | .00289 | .00280 | .00272 | .00264 |
| **2.80** | .00256 | .00248 | .00240 | .00233 | .00226 | .00219 | .00212 | .00205 | .00199 | .00193 |
| **2.90** | .00187 | .00181 | .00175 | .00169 | .00164 | .00159 | .00154 | .00149 | .00144 | .00139 |
| **3.00** | .00135 | .00131 | .00126 | .00122 | .00118 | .00114 | .00111 | .00107 | .00104 | .00100 |
| **3.10** | .00097 | .00094 | .00090 | .00087 | .00084 | .00082 | .00079 | .00076 | .00074 | .00071 |
| **3.20** | .00069 | .00066 | .00064 | .00062 | .00060 | .00058 | .00056 | .00054 | .00052 | .00050 |
| **3.30** | .00048 | .00047 | .00045 | .00043 | .00042 | .00040 | .00039 | .00038 | .00036 | .00035 |
| **3.40** | .00034 | .00032 | .00031 | .00030 | .00029 | .00028 | .00027 | .00026 | .00025 | .00024 |
| **3.50** | .00023 | .00022 | .00022 | .00021 | .00020 | .00019 | .00019 | .00018 | .00017 | .00017 |
| **3.60** | .00016 | .00015 | .00015 | .00014 | .00014 | .00013 | .00013 | .00012 | .00012 | .00011 |
| **3.70** | .00011 | .00010 | .00010 | .00010 | .00009 | .00009 | .00008 | .00008 | .00008 | .00008 |
| **3.80** | .00007 | .00007 | .00007 | .00006 | .00006 | .00006 | .00006 | .00005 | .00005 | .00005 |
| **3.90** | .00005 | .00005 | .00004 | .00004 | .00004 | .00004 | .00004 | .00004 | .00003 | .00003 |
| **4.00** | .00003 | | | | | | | | | |

# *t* Table

**Critical values for *t***

| | 2-tailed (nondirectional) test | | | 1-tailed (directional) test | | | |
|---|---|---|---|---|---|---|---|
| *df* | α = .05 | α = .01 | α = .001 | *df* | α = .05 | α = .01 | α = .001 |
| 1 | 12.706 | 63.657 | 636.619 | 1 | 6.314 | 31.821 | 318.309 |
| 2 | 4.303 | 9.925 | 31.599 | 2 | 2.920 | 6.965 | 22.327 |
| 3 | 3.182 | 5.841 | 12.924 | 3 | 2.353 | 4.541 | 10.215 |
| 4 | 2.776 | 4.604 | 8.610 | 4 | 2.132 | 3.747 | 7.173 |
| 5 | 2.571 | 4.032 | 6.869 | 5 | 2.015 | 3.365 | 5.893 |
| 6 | 2.447 | 3.707 | 5.959 | 6 | 1.943 | 3.143 | 5.208 |
| 7 | 2.365 | 3.499 | 5.408 | 7 | 1.895 | 2.998 | 4.785 |
| 8 | 2.306 | 3.355 | 5.041 | 8 | 1.860 | 2.896 | 4.501 |
| 9 | 2.262 | 3.250 | 4.781 | 9 | 1.833 | 2.821 | 4.297 |
| 10 | 2.228 | 3.169 | 4.587 | 10 | 1.812 | 2.764 | 4.144 |
| 11 | 2.201 | 3.106 | 4.437 | 11 | 1.796 | 2.718 | 4.025 |
| 12 | 2.179 | 3.055 | 4.318 | 12 | 1.782 | 2.681 | 3.930 |
| 13 | 2.160 | 3.012 | 4.221 | 13 | 1.771 | 2.650 | 3.852 |
| 14 | 2.145 | 2.977 | 4.140 | 14 | 1.761 | 2.624 | 3.787 |
| 15 | 2.131 | 2.947 | 4.073 | 15 | 1.753 | 2.602 | 3.733 |
| 16 | 2.120 | 2.921 | 4.015 | 16 | 1.746 | 2.583 | 3.686 |
| 17 | 2.110 | 2.898 | 3.965 | 17 | 1.740 | 2.567 | 3.646 |
| 18 | 2.101 | 2.878 | 3.922 | 18 | 1.734 | 2.552 | 3.610 |
| 19 | 2.093 | 2.861 | 3.883 | 19 | 1.729 | 2.539 | 3.579 |
| 20 | 2.086 | 2.845 | 3.850 | 20 | 1.725 | 2.528 | 3.552 |
| 21 | 2.080 | 2.831 | 3.819 | 21 | 1.721 | 2.518 | 3.527 |
| 22 | 2.074 | 2.819 | 3.792 | 22 | 1.717 | 2.508 | 3.505 |
| 23 | 2.069 | 2.807 | 3.768 | 23 | 1.714 | 2.500 | 3.485 |
| 24 | 2.064 | 2.797 | 3.745 | 24 | 1.711 | 2.492 | 3.467 |
| 25 | 2.060 | 2.787 | 3.725 | 25 | 1.708 | 2.485 | 3.450 |
| 26 | 2.056 | 2.779 | 3.707 | 26 | 1.706 | 2.479 | 3.435 |
| 27 | 2.052 | 2.771 | 3.690 | 27 | 1.703 | 2.473 | 3.421 |
| 28 | 2.048 | 2.763 | 3.674 | 28 | 1.701 | 2.467 | 3.408 |
| 29 | 2.045 | 2.756 | 3.659 | 29 | 1.699 | 2.462 | 3.396 |
| 30 | 2.042 | 2.750 | 3.646 | 30 | 1.697 | 2.457 | 3.385 |
| 40 | 2.021 | 2.704 | 3.551 | 40 | 1.684 | 2.423 | 3.307 |
| 60 | 2.000 | 2.660 | 3.460 | 60 | 1.671 | 2.390 | 3.232 |
| 120 | 1.980 | 2.617 | 3.373 | 120 | 1.658 | 2.358 | 3.160 |
| ∞ | 1.960 | 2.576 | 3.291 | ∞ | 1.645 | 2.326 | 3.090 |

# *q* Table

**Critical Values of *q* for Tukey's *HSD* Test with α = .05**

| *df* error | # of Means (*k*) 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| **2** | 8.33 | 9.80 | 10.88 | 11.73 | 12.43 | 13.03 |
| **3** | 5.91 | 6.82 | 7.50 | 8.04 | 8.48 | 8.85 |
| **4** | 5.04 | 5.76 | 6.29 | 6.71 | 7.05 | 7.35 |
| **5** | 4.60 | 5.22 | 5.67 | 6.03 | 6.33 | 6.58 |
| **6** | 4.34 | 4.90 | 5.30 | 5.63 | 5.90 | 6.12 |
| **7** | 4.16 | 4.68 | 5.06 | 5.36 | 5.61 | 5.82 |
| **8** | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 |
| **9** | 3.95 | 4.41 | 4.76 | 5.02 | 5.24 | 5.43 |
| **10** | 3.88 | 4.33 | 4.65 | 4.91 | 5.12 | 5.30 |
| **11** | 3.82 | 4.26 | 4.57 | 4.82 | 5.03 | 5.20 |
| **12** | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 |
| **13** | 3.73 | 4.15 | 4.45 | 4.69 | 4.88 | 5.05 |
| **14** | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 |
| **15** | 3.67 | 4.08 | 4.37 | 4.59 | 4.78 | 4.94 |
| **16** | 3.65 | 4.05 | 4.33 | 4.56 | 4.74 | 4.90 |
| **17** | 3.63 | 4.02 | 4.30 | 4.52 | 4.70 | 4.86 |
| **18** | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.82 |
| **19** | 3.59 | 3.98 | 4.25 | 4.47 | 4.65 | 4.79 |
| **20** | 3.58 | 3.96 | 4.23 | 4.45 | 4.62 | 4.77 |
| **21** | 3.56 | 3.94 | 4.21 | 4.42 | 4.60 | 4.74 |
| **22** | 3.55 | 3.93 | 4.20 | 4.41 | 4.58 | 4.72 |
| **23** | 3.54 | 3.91 | 4.18 | 4.39 | 4.56 | 4.70 |
| **24** | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 |
| **25** | 3.52 | 3.89 | 4.15 | 4.36 | 4.53 | 4.67 |
| **26** | 3.51 | 3.88 | 4.14 | 4.35 | 4.51 | 4.65 |
| **27** | 3.51 | 3.87 | 4.13 | 4.33 | 4.50 | 4.64 |
| **28** | 3.50 | 3.86 | 4.12 | 4.32 | 4.49 | 4.62 |
| **29** | 3.49 | 3.85 | 4.11 | 4.31 | 4.47 | 4.61 |
| **30** | 3.49 | 3.85 | 4.10 | 4.30 | 4.46 | 4.60 |
| **32** | 3.48 | 3.83 | 4.09 | 4.28 | 4.45 | 4.58 |
| **34** | 3.47 | 3.82 | 4.07 | 4.27 | 4.43 | 4.56 |
| **36** | 3.46 | 3.81 | 4.06 | 4.25 | 4.41 | 4.55 |
| **38** | 3.45 | 3.80 | 4.05 | 4.24 | 4.40 | 4.53 |
| **40** | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 |
| **42** | 3.44 | 3.78 | 4.03 | 4.22 | 4.38 | 4.51 |
| **44** | 3.43 | 3.78 | 4.02 | 4.21 | 4.37 | 4.50 |
| **46** | 3.42 | 3.77 | 4.01 | 4.20 | 4.36 | 4.49 |
| **48** | 3.42 | 3.76 | 4.01 | 4.20 | 4.35 | 4.48 |
| **50** | 3.42 | 3.76 | 4.00 | 4.19 | 4.34 | 4.47 |
| **55** | 3.41 | 3.75 | 3.99 | 4.18 | 4.33 | 4.46 |
| **60** | 3.40 | 3.74 | 3.98 | 4.16 | 4.31 | 4.44 |
| **65** | 3.39 | 3.73 | 3.97 | 4.15 | 4.30 | 4.43 |
| **70** | 3.39 | 3.72 | 3.96 | 4.14 | 4.29 | 4.42 |
| **80** | 3.38 | 3.71 | 3.95 | 4.13 | 4.28 | 4.40 |
| **100** | 3.36 | 3.70 | 3.93 | 4.11 | 4.26 | 4.38 |
| **125** | 3.35 | 3.68 | 3.91 | 4.09 | 4.24 | 4.36 |
| **150** | 3.35 | 3.67 | 3.90 | 4.08 | 4.23 | 4.35 |
| **200** | 3.34 | 3.66 | 3.89 | 4.07 | 4.21 | 4.33 |
| **400** | 3.33 | 3.65 | 3.88 | 4.05 | 4.19 | 4.31 |
| **1000** | 3.32 | 3.64 | 3.86 | 4.04 | 4.18 | 4.30 |

# $\chi^2$ Table

**Critical Values for $\chi^2$**

| *df* | α = .05 | α = .01 |
|---|---|---|
| **1** | 3.84 | 6.63 |
| **2** | 5.99 | 9.21 |
| **3** | 7.81 | 11.34 |
| **4** | 9.49 | 13.28 |
| **5** | 11.07 | 15.09 |
| **6** | 12.59 | 16.81 |
| **7** | 14.07 | 18.48 |
| **8** | 15.51 | 20.09 |
| **9** | 16.92 | 21.67 |
| **10** | 18.31 | 23.21 |
| **11** | 19.68 | 24.72 |
| **12** | 21.03 | 26.22 |
| **13** | 22.36 | 27.69 |
| **14** | 23.68 | 29.14 |
| **15** | 25.00 | 30.58 |
| **16** | 26.30 | 32.00 |
| **17** | 27.59 | 33.41 |
| **18** | 28.87 | 34.81 |
| **19** | 30.14 | 36.19 |
| **20** | 31.41 | 37.57 |
| **21** | 32.67 | 38.93 |
| **22** | 33.92 | 40.29 |
| **23** | 35.17 | 41.64 |
| **24** | 36.42 | 42.98 |
| **25** | 37.65 | 44.31 |
| **26** | 38.89 | 45.64 |
| **27** | 40.11 | 46.96 |
| **28** | 41.34 | 48.28 |
| **29** | 42.56 | 49.59 |
| **30** | 43.77 | 50.89 |
| **40** | 55.76 | 63.69 |
| **50** | 67.50 | 76.15 |
| **60** | 79.08 | 88.38 |
| **70** | 90.53 | 100.43 |

# Appendix B: Statistical Power

The table presented below provides rough sample size guidelines for various research designs. Three assumptions are made. First, all applicable assumptions of statistical tests (e.g., independence of observations, normality of population distributions, equal population variances) are satisfied.

Second, two-tailed tests with $\alpha = .05$ are used. One-tailed tests or larger values of $\alpha$ (e.g., .10) yield larger values of statistical power, the required sample sizes would be smaller than those listed in the table; the opposite is true for smaller values of $\alpha$ (e.g., .01 or .001), which yield smaller values of statistical power and would require larger sample sizes.

Third, the desired level of statistical power is .80. To achieve higher statistical power, larger sample sizes than those listed in the table are required.

For more information, consult Cohen's (1988) book on statistical power.[56]

| Research Design | Statistical Test | Effect Size Measure | Rules of Thumb | Sample Size Required for Power = .80 |
|---|---|---|---|---|
| Two independent groups[a] | *t* | *d* | 0.20 = small<br>0.50 = medium<br>0.80 = large | *N* = 786 (393 per group)<br>*N* = 128 (64 per group)<br>*N* = 52 (26 per group) |
| Three independent groups[a] | *F* | $\eta^2$ | .01 = small<br>.09 = medium<br>.25 = large | *N* = 966 (322 per group)<br>*N* = 156 (52 per group)<br>*N* = 63 (21 per group) |
| Four independent groups[a] | *F* | $\eta^2$ | .01 = small<br>.09 = medium<br>.25 = large | *N* = 1,096 (274 per group)<br>*N* = 180 (45 per group)<br>*N* = 64 (18 per group) |
| Correlation between two variables | *r* | *r* | .10 = small<br>.30 = medium<br>.50 = large | *N* = 783<br>*N* = 85<br>*N* = 28 |
| $\chi^2$ test of independence (1 *df*) | $\chi^2$ | $\phi$ | .10 = small<br>.30 = medium<br>.50 = large | *N* = 785<br>*N* = 87<br>*N* = 26 |

[a] Tests for related samples usually require smaller samples than those for independent groups; how much smaller depends on the correlations between scores for the related samples.

[56] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

# Appendix C: Selecting a Statistical Test

## Testing Differences Between Means

### One Variable Whose *M* Will Be Compared to Population μ

Population standard deviation (σ) known → **one sample *z* test**

Population standard deviation (σ) unknown → **one sample *t* test**

### Categorical Independent Variable(s)

Subjects form independent groups (nobody participates in more than one condition):

- Two groups → **independent groups *t* test**[57]
- More than two groups → **independent groups ANOVA**

Subjects are tested repeatedly *or* matched (everyone participates in every condition):

- Two conditions → **related samples *t* test**
- More than two conditions → **related samples ANOVA**

More than one categorical independent variable → **factorial ANOVA**

## Testing Relationships Between Two or More Variables

### Association Between Two Variables

Two quantitative variables → **correlation** ($r$)

Two ranked variables → **Spearman's rank-order correlation** ($r_S$)

One quantitative and one dichotomous variable → **point-biserial correlation** ($r_{pb}$)[58]

Two dichotomous variables → **phi coefficient** ($\phi$)[59]

Two categorical variables → **$\chi^2$ test of independence**[60]

### Making Predictions

One predictor variable → **regression** (aka **simple linear regression**)

Two or more predictor variables → **multiple regression**

## Testing Goodness of Fit Between Observed and Expected Frequencies

Sample contains only one categorical variable → **$\chi^2$ goodness of fit test**

---

[57] The independent groups *t* test is equivalent to the point-biserial correlation.

[58] The point-biserial correlation is equivalent to the independent groups *t* test.

[59] The phi coefficient is equivalent to the $\chi^2$ test of independence.

[60] If both variables are dichotomous, the $\chi^2$ test of independence is equivalent to the phi coefficient.

# Appendix D: Symbols and Abbreviations

Unless otherwise noted, the following symbols and abbreviations refer to sample statistics rather than population parameters.

## Descriptive Statistics

| | |
|---|---|
| $N$ | Sample size |
| $n$ | Subsample size |
| $X$ | Score for an individual |
| $M$ | Sample mean |
| $\mu$ | Population mean (Greek letter "mu") |
| $SD$ | Sample standard deviation |
| $\sigma$ | Population standard deviation (Greek letter "sigma") |
| $Mdn$ | Median |
| $IQR$ | Interquartile range |

## *z* Score and *z* Test

| | |
|---|---|
| $z$ | Standard score for individual or sample |
| $H_0$ | Null hypothesis |
| $H_1$ | Alternative hypothesis |
| $\sigma_M$ | Standard error of the mean |
| $\alpha$ | Size of critical region (Greek letter "alpha") |
| $d$ | Cohen's effect size measure for comparing two means |
| $p$ | Probability value |

## *t* Tests

| | |
|---|---|
| $t$ | Value for one sample, related samples, or independent groups test |
| $df$ | Degrees of freedom |
| $SD_M$ | Standard error of the mean |
| $D$ | Difference score |
| $Y_1$ | Score in one condition |
| $Y_2$ | Score in the other condition |
| $M_D$ | Mean of the difference scores |
| $SD_D$ | Standard deviation of the difference scores |
| $\mu_D$ | Population mean difference score |
| $SD_{MD}$ | Standard error of the mean difference score |
| $SD_p$ | Pooled standard deviation |
| $SD_{M1-M2}$ | Standard error of the difference between two groups' means |

## ANOVAs

| | |
|---|---|
| $k$ | Number of conditions being compared |
| $F$ | Ratio of systematic to error variance |
| $\eta^2$ | Effect size measure for comparing two or more means (Greek letter "eta" squared) |

## Correlation

| | |
|---|---|
| $r$ | Correlation coefficient (Pearson's) |
| $\rho$ | Population correlation coefficient (Greek letter "rho") |
| $r_S$ | Spearman's rank-order correlation coefficient |
| $r_{pb}$ | Point-biserial correlation coefficient for one dichotomous and one continuous variable |
| $\phi$ | Correlation coefficient for two dichotomous variables (Greek letter "phi") |
| $r^2$ | Coefficient of determination |
| $r'$ | Correlation coefficient corrected for measurement error |
| $r_{xx}$ | Reliability of variable $X$ |
| $r_{yy}$ | Reliability of variable $Y$ |

## Regression

| | |
|---|---|
| $Y'$ | Predicted value of criterion variable |
| $X$ | Predictor variable (simple linear regression) |
| $b$ | Regression slope (simple linear regression) |
| $a$ | Regression intercept (simple linear regression) |
| $X_1, X_2, X_3, \ldots$ | Predictor variables (multiple regression) |
| $b_1, b_2, b_3, \ldots$ | Regression coefficients (multiple regression) |
| $b_0$ | Regression constant (multiple regression) |
| $SE_{est}$ | Standard error of the estimate |

## $\chi^2$ Tests

| | |
|---|---|
| $\chi^2$ | Value for goodness of fit or independence test (Greek letter "chi" squared) |
| $f_O$ | Observed frequency |
| $f_E$ | Expected frequency |
| $T$ | Total frequency for all cells |
| $T_R$ | Total frequency for one row |
| $T_C$ | Total frequency for one column |
| $\phi$ | Effect size measure for two dichotomous variables (Greek letter "phi") |

# Appendix E: Formulas

See Appendix D for definitions of the symbols and abbreviations that appear in these formulas. When performing calculations, pay close attention to the use of parentheses and the order of operations.

## Descriptive Statistics

$M = \Sigma X / N$

$SD = \text{sqrt}(\Sigma(X - M)^2 / (N - 1))$

## z Score and z Test

$z = (X - \mu) / \sigma$

$z = (M - \mu) / \sigma_M$

$\sigma_M = \sigma / \text{sqrt}(N)$

$d = (M - \mu) / \sigma$

## t Tests

$SD_M = SD / \text{sqrt}(N)$

$t = (M - \mu) / SD_M$

$d = (M - \mu) / SD$

$D = Y_1 - Y_2$

$SD_{MD} = SD_D / \text{sqrt}(N)$

$t = M_D / SD_{MD}$

$SD_p = \text{sqrt}((SD_1^2 + SD_2^2) / 2)$

$d = (M_1 - M_2) / SD_p$

$SD_{M1\text{-}M2} = \text{sqrt}(2) \times SD_p / \text{sqrt}(n)$

$SD_p = \text{sqrt}((SD_1^2 \times df_1 + SD_2^2 \times df_2) / (df_1 + df_2))$

$SD_{M1\text{-}M2} = \text{sqrt}((SD_p^2 / n_1) + (SD_p^2 / n_2))$

$t = (M_1 - M_2) / SD_{M1\text{-}M2}$

## ANOVAs

$F = t^2$

$HSD = q \times \text{sqrt}(MS_{\text{error}} / N)$

## Correlation

$r = \Sigma(z_x \times z_y) / N$

$r_{pb} = \text{sqrt}(t^2 / (t^2 + df))$

$r' = r / \text{sqrt}(r_{xx} \times r_{yy})$

## Regression

$Y' = bX + a$

$Y' = b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k + b_0$

## $\chi^2$ Tests

$\chi^2 = \Sigma((f_O - f_E)^2 / f_E)$

$f_E = T_R \times T_C / T$

$\phi = \text{sqrt}(\chi^2 / N)$