EXPERIMENTAL DESIGN

Lydia Yang
Kansas State University

LibreTexts™

Experimental Design (Yang)

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.

LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 04/19/2026

TABLE OF CONTENTS

Licensing

1: Introduction to Research Designs

- 1.1: Research Designs
- 1.2: Internal and External Validity
- 1.3: Threats to Internal Validity and Different Control Techniques
- 1.4: Common Experimental Designs
- 1.5: Common Quasi-Experimental Designs
- 1.6: Non-Experimental Research

2: Between-Subjects Single Factor Design

- 2.1: Between-Subjects Design
- 2.2: One-Way ANOVA
- 2.3: Computing the F statistics
- 2.4 Post Hoc Tests

3: Between-Subjects Factorial Design

- 3.1: Factorial Designs
- 3.2: Factorial ANOVA - Main Effects
- 3.3: Factorial ANOVA - Interaction Effects
- 3.4: Factorial ANOVA - Simple Effects
- 3.5: Partioning the Sum of Squares

4: Between-Subjects Design with a Control Variable

- 4.1: Overview of the Control Variable
- 4.2: Radomized Block Design
- 4.4: Analysis of Covariance (ANCOVA)

5: Within-Subjects Design

- 5.1: Within-Subjects Design
- 5.2: Partioning the Sums of Squares
- 5.3: Repeated Measures ANOVA
- 5.4: More Notes on Repeated Measures ANOVA

6: Multiple Linear Regression

- 6.1: Simple Linear Equation Refresher
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

Index

Glossary

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

1: Introduction to Research Designs

1.1: Research Designs
1.2: Internal and External Validity
1.3: Threats to Internal Validity and Different Control Techniques
1.4: Common Experimental Designs
1.5: Common Quasi-Experimental Designs
1.6: Non-Experimental Research

This page titled 1: Introduction to Research Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

1.1: Research Designs

Research Designs

In the early 1970's, a man named Uri Geller tricked the world: he convinced hundreds of thousands of people that he could bend spoons and slow watches using only the power of his mind. In fact, if you were in the audience, you would have likely believed he had psychic powers. Everything looked authentic—this man had to have paranormal abilities! So, why have you probably never heard of him before? Because when Uri was asked to perform his miracles in line with scientific experimentation, he was no longer able to do them. That is, even though it seemed like he was doing the impossible, when he was tested by science, he proved to be nothing more than a clever magician.

When we look at dinosaur bones to make educated guesses about extinct life, or systematically chart the heavens to learn about the relationships between stars and planets, or study magicians to figure out how they perform their tricks, we are forming observations—the foundation of science. Although we are all familiar with the saying "seeing is believing," conducting science is more than just what your eyes perceive. Science is the result of systematic and intentional study of the natural world. And soical science is no different. In the movie *Jerry Maguire*, Cuba Gooding, Jr. became famous for using the phrase, "Show me the money!" In education, as in all sciences, we might say, "Show me the data!"

One of the important steps in scientific inquiry is to test our research questions, otherwise known as hypotheses. However, there are many ways to test hypotheses in educational research. Which method you choose will depend on the type of questions you are asking, as well as what resources are available to you. All methods have limitations, which is why the best research uses a variety of methods.

Experimental Research

If somebody gave you $20 that absolutely had to be spent today, how would you choose to spend it? Would you spend it on an item you've been eyeing for weeks, or would you donate the money to charity? Which option do you think would bring you the most happiness? If you're like most people, you'd choose to spend the money on yourself (duh, right?). Our intuition is that we'd be happier if we spent the money on ourselves.

Knowing that our intuition can sometimes be wrong, Professor Elizabeth Dunn (2008) at the University of British Columbia set out to conduct an experiment on spending and happiness. She gave each of the participants in her experiment $20 and then told them they had to spend the money by the end of the day. Some of the participants were told they must spend the money on themselves, and some were told they must spend the money on others (either charity or a gift for someone). At the end of the day she measured participants' levels of happiness using a self-report questionnaire.

In an experiment, researchers manipulate, or cause changes, in the **independent variable**, and observe or measure any impact of those changes in the **dependent variable**. The independent variable is the one under the researcher's control, or the variable that is intentionally altered between groups. In the case of Dunn's experiment, the independent variable was whether participants spent the money on themselves or on others. The dependent variable is the variable that is not manipulated at all, or the one where the effect happens. One way to help remember this is that the dependent variable "depends" on what happens to the independent variable. In our example, the participants' happiness (the dependent variable in this experiment) depends on how the participants spend their money (the independent variable). Thus, any observed changes or group differences in happiness can be attributed to whom the money was spent on. What Dunn and her colleagues found was that, after all the spending had been done, the people who had spent the money on others were happier than those who had spent the money on themselves. In other words, spending on others causes us to be happier than spending on ourselves. Do you find this surprising?

But wait! Doesn't happiness depend on a lot of different factors—for instance, a person's upbringing or life circumstances? What if some people had happy childhoods and that's why they're happier? Or what if some people dropped their toast that morning and it fell jam-side down and ruined their whole day? It is correct to recognize that these factors and many more can easily affect a person's level of happiness. So how can we accurately conclude that spending money on others causes happiness, as in the case of Dunn's experiment?

The most important thing about experiments is **random assignment**. Participants don't get to pick which condition they are in (e.g., participants didn't choose whether they were supposed to spend the money on themselves versus others). The experimenter

assigns them to a particular condition based on the flip of a coin or the roll of a die or any other random method. Why do researchers do this? With Dunn's study, there is the obvious reason: you can imagine which condition most people would choose to be in, if given the choice. But another equally important reason is that random assignment makes it so the groups, on average, are similar on all characteristics except what the experimenter manipulates.

By randomly assigning people to conditions (self-spending versus other-spending), some people with happy childhoods should end up in each condition. Likewise, some people who had dropped their toast that morning (or experienced some other disappointment) should end up in each condition. As a result, the distribution of all these factors will generally be consistent across the two groups, and this means that on average the two groups will be relatively equivalent on all these factors. Random assignment is critical to experimentation because if the only difference between the two groups is the independent variable, we can infer that the independent variable is the cause of any observable difference (e.g., in the amount of happiness they feel at the end of the day).

Here's another example of the importance of random assignment: Let's say your class is going to form two basketball teams, and you get to be the captain of one team. The class is to be divided evenly between the two teams. If you get to pick the players for your team first, whom will you pick? You'll probably pick the tallest members of the class or the most athletic. You probably won't pick the short, uncoordinated people, unless there are no other options. As a result, your team will be taller and more athletic than the other team. But what if we want the teams to be fair? How can we do this when we have people of varying height and ability? All we have to do is randomly assign players to the two teams. Most likely, some tall and some short people will end up on your team, and some tall and some short people will end up on the other team. The average height of the teams will be approximately the same. That is the power of random assignment!

Other considerations

In addition to using random assignment, you should avoid introducing confounding variables into your experiments. **Confounding variables** are things that could undermine your ability to draw causal inferences. For example, if you wanted to test if a new happy pill will make people happier, you could randomly assign participants to take the happy pill or not (the independent variable) and compare these two groups on their self-reported happiness (the dependent variable). However, if some participants know they are getting the happy pill, they might develop expectations that influence their self-reported happiness. This is sometimes known as a **placebo effect**. Sometimes a person just knowing that he or she is receiving special treatment or something new is enough to actually cause changes in behavior or perception: In other words, even if the participants in the happy pill condition were to report being happier, we wouldn't know if the pill was actually making them happier or if it was the placebo effect—an example of a confound. Even **experimenter expectations** can influence the outcome of a study. For example, if the experimenter knows who took the happy pill and who did not, and the dependent variable is the experimenter's observations of people's happiness, then the experimenter might perceive improvements in the happy pill group that are not really there.

One way to prevent these confounds from affecting the results of a study is to use a double-blind procedure. In a double-blind procedure, neither the participant nor the experimenter knows which condition the participant is in. For example, when participants are given the happy pill or the fake pill, they don't know which one they are receiving. This way the participants shouldn't experience the placebo effect, and will be unable to behave as the researcher expects (participant demand). Likewise, the researcher doesn't know which pill each participant is taking (at least in the beginning—later, the researcher will get the results for data-analysis purposes), which means the researcher's expectations can't influence his or her observations. Therefore, because both parties are "blind" to the condition, neither will be able to behave in a way that introduces a confound. At the end of the day, the only difference between groups will be which pills the participants received, allowing the researcher to determine if the happy pill actually caused people to be happier.

Quasi-Experimental Designs

What if you want to study the effects of marriage on a variable? For example, does marriage make people happier? Can you randomly assign some people to get married and others to remain single? Of course not. So how can you study these important variables? You can use a **quasi-experimental design**. A quasi-experimental design is similar to experimental research, except that random assignment to conditions is not used. Instead, we rely on existing group memberships (e.g., married vs. single). We treat these as the independent variables, even though we don't assign people to the conditions and don't manipulate the variables. As a result, with quasi-experimental designs causal inference is more difficult. For example, married people might differ on a variety of characteristics from unmarried people. If we find that married participants are happier than single participants, it will be hard to say

that marriage causes happiness, because the people who got married might have already been happier than the people who have remained single.

Because experimental and quasi-experimental designs can seem pretty similar, let's take another example to distinguish them. Imagine you want to know who is a better professor: Dr. Smith or Dr. Khan. To judge their ability, you're going to look at their students' final grades. Here, the independent variable is the professor (Dr. Smith vs. Dr. Khan) and the dependent variable is the students' grades. In an experimental design, you would randomly assign students to one of the two professors and then compare the students' final grades. However, in real life, researchers can't randomly force students to take one professor over the other; instead, the researchers would just have to use the preexisting classes and study them as-is (quasi-experimental design). Again, the key difference is random assignment to the conditions of the independent variable. Although the quasi-experimental design (where the students choose which professor they want) may seem random, it's most likely not. For example, maybe students heard Dr. Smith sets low expectations, so slackers prefer this class, whereas Dr. Khan sets higher expectations, so smarter students prefer that one. This now introduces a confounding variable (student intelligence) that will almost certainly have an effect on students' final grades, regardless of how skilled the professor is. So, even though a quasi-experimental design is similar to an experimental design (i.e., it has a manipulated independent variable), because there's no random assignment, you can't reasonably draw the same conclusions that you would with an experimental design.

Non-Experimental Studies

When scientists passively observe and measure phenomena it is called non-experimental research. Here, we do not intervene and change behavior, as we do in experiments. In non-experimental research, we identify patterns of relationships, but we usually cannot infer what causes what. Importantly, with non-experimental research, you can examine only two variables at a time, no more and no less.

So, what if you wanted to test whether spending on others is related to happiness, but you don't have $20 to give to each participant? You could use a non-experimental research — which is exactly what Professor Dunn did, too. She asked people how much of their income they spent on others or donated to charity, and later she asked them how happy they were. Do you think these two variables were related? Yes, they were! The more money people reported spending on others, the happier they were. This indicates a positive correlation!

If generosity and happiness are positively correlated, should we conclude that being generous causes happiness? Similarly, if height and pathogen prevalence are negatively correlated, should we conclude that disease causes shortness? From a correlation alone, we can't be certain. For example, in the first case it may be that happiness causes generosity, or that generosity causes happiness. Or, a third variable might cause both happiness *and* generosity, creating the illusion of a direct link between the two. For example, wealth could be the third variable that causes both greater happiness and greater generosity. This is why **correlation does not mean causation**—an often repeated phrase among psychologists.

One particular type of non-experimental research is the **longitudinal study**. Longitudinal studies are typically observational in nature. They track the same people over time. Some longitudinal studies last a few weeks, some a few months, some a year or more. Some studies that have contributed a lot to a given topic by following the same people over decades. For example, one study followed more than 20,000 Germans for two decades. From these longitudinal data, psychologist Rich Lucas (2003) was able to determine that people who end up getting married indeed start off a bit happier than their peers who never marry. Longitudinal studies like this provide valuable evidence for testing many theories in social sciences, but they can be quite costly to conduct, especially if they follow many people for many years.

Tradeoffs in Research

Even though there are serious limitations to non-experimental and quasi-experimental research, they are not poor cousins to experiments designs. In addition to selecting a method that is appropriate to the question, many practical concerns may influence the decision to use one method over another. One of these factors is simply resource availability—how much time and money do you have to invest in the research? Often, we survey people even though it would be more precise—but much more difficult—to track them longitudinally. Especially in the case of exploratory research, it may make sense to opt for a cheaper and faster method first. Then, if results from the initial study are promising, the researcher can follow up with a more intensive method.

Beyond these practical concerns, another consideration in selecting a research design is the ethics of the study. For example, in cases of brain injury or other neurological abnormalities, it would be unethical for researchers to inflict these impairments on healthy participants. Nonetheless, studying people with these injuries can provide great insight into human mind (e.g., if we learn that damage to a particular region of the brain interferes with emotions, we may be able to develop treatments for emotional irregularities). In addition to brain injuries, there are numerous other areas of research that could be useful in understanding the human mind but which pose challenges to a true experimental design — such as the experiences of war, long-term isolation, abusive parenting, or prolonged drug use. However, none of these are conditions we could ethically experimentally manipulate and randomly assign people to. Therefore, ethical considerations are another crucial factor in determining an appropriate research design.

Research Methods: Why You Need Them

Just look at any major news outlet and you'll find research routinely being reported. Sometimes the journalist understands the research methodology, sometimes not (e.g., correlational evidence is often incorrectly represented as causal evidence). Often, the media are quick to draw a conclusion for you. After reading this module, you should recognize that the strength of a scientific finding lies in the strength of its methodology. Therefore, in order to be a savvy producer and/or consumer of research, you need to understand the pros and cons of different methods and the distinctions among them.

References

Dunn, E. W., Aknin, L. B., & Norton, M. I. (2008). Spending money on others promotes happiness. Science, 319(5870), 1687–1688. doi:10.1126/science.1150952

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2003). Re-examining adaptation and the setpoint model of happiness: Reactions to changes in marital status. Journal of Personality and Social Psychology, 84, 527–539.

This page titled 1.1: Research Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **2.4: Research Designs** by Christie Napa Scollon is licensed CC BY-NC-SA 4.0.

1.2: Internal and External Validity

Internal and External Validity

Developing a research design should be more than just a matter of convenience (although practicality is an important element, which we touched on in the last section). Not all designs are created equally and there are trade-offs we make when opting for one type of design over another. The two major components of an assessment of a research design are its internal validity and its external validity. **Internal validity** basically means we can make a causal statement within the context of our study. We have internal validity if, for our study, we can say our independent variable caused our dependent variable. Often times, the major challenge is the issue of spuriousness. We have to ask if our design allows us to say our independent variable makes our dependent variable vary systematically as it changes and that those changes in the dependent variable are not due to some third or extraneous variable/factor.

The second basis for evaluating your research design is to assess its **external validity**. External validity means that we can generalize the results of our study. It asks whether our findings are applicable in other settings. Here we consider what population we are interested in generalizing to. We might be interested in adult Americans, but if we have studied a sample of first-year college students then we might not be able to generalize to our target population. As you can see here, the sampling method is the key. The quality of the sampling method you choose is directly tied to your ability to generalize the findings of one particular study to the entire population. Typically a representative sampling method gives us the best chance to generalize the findings to our target population, thus gives the study high external validity. By contrast, when a non-representative sampling methods is used, it reduces generalizability, i.e., the external validity of the study.

This page titled 1.2: Internal and External Validity is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **2.2: Internal and External Validity** by Jenkins-Smith et al. is licensed CC BY 4.0. Original source: https://bookdown.org/josiesmith/qrmbook.

1.3: Threats to Internal Validity and Different Control Techniques

Internal validity is often the focus from a research design perspective. To understand the pros and cons of various designs and to be able to better judge specific designs, we identify specific **threats to internal validity**. Before we do so, it is important to note that the primary challenge to establishing internal validity in social sciences is the fact that most of the phenomena we care about have multiple causes and are often a result of some complex set of interactions. For example, X may be only a partial cause of Y or X may cause Y, but only when Z is present. Multiple causation and interactive effects make it very difficult to demonstrate causality. Turning now to more specific threats, Figure 1.3.1 below identifies common threats to internal validity.

Figure **1.3.1**: Common Threats to Internal Validity

Threat	
History	Any event that occurs while the experiment is in progress might be an alternation; using a control group mitigates this concern.
Maturation	Normal changes over time (e.g., fatigue or aging) might affect the dependent variable; using a control group mitigates this concern
Selection Bias	If randomization is not used to assign participants, the groups may not be equivalent
Experimental Mortality	If groups lost participants (e.g., due to dropping out of the experiment) they may not be equivalent.
Testing	A pre-test may confound the influence of the experimental treatment; using a control group mitigates this concern
Instrumentation	Changes or difference in the process of measurements might alternatively account for differences
Statistical Regression	The natural tendency for extreme scores to regress or move towards the mean

Different Control Techniques

All of the common threats mentioned above can introduce extraneous variables into your research design, which will potentially confound your research findings. In other words, we won't be able to tell whether it is the independent variable (i.e., the treatment we give participants), or the extraneous variable, that causes the changes in the dependent variable. Controlling for extraneous variables reduces its threats on the research design and gives us a better chance to claim the independent variable causes the changes in the dependent variable, i.e., internal validity. There are different techniques we can use to control for extraneous variables.

Random assignment

Random assignment is the single most powerful control technique we can use to minimize the potential threats of the confounding variables in research design. As we have seen in Dunn and her colleagues' study earlier, participants are not allowed to self select into either conditions (spend $20 on self or spend on others). Instead, they are randomly assigned into either group by the researcher(s). By doing so, the two groups are likely to be similar on all other factors except the independent variable itself. One confounding variable mentioned earlier is whether individuals had a happy childhood to begin with. Using random assignment, those who had a happy childhood will likely end up in each condition group. Similarly, those who didn't have a happy childhood will likely end up in each condition group too. As a consequence, we can expect the two condition groups to be very similar on this confounding variable. Applying the same logic, we can use random assignment to minimize all potential confounding variables (assuming your sample size is large enough!). With that, the only difference between the two groups is the condition participants are assigned to, which is the independent variable, then we are confident to infer that the independent variable actually causes the differences in the dependent variables.

It is critical to emphasize that random assignment is the only control technique to control for both known and unknown confounding variables. With all other control techniques mentioned below, we must first know what the confounding variable is before controlling it. Random assignment does not. With the simple act of randomly assigning participants into different conditions, we take care both the confounding variables we know of and the ones we don't even know that could threat the internal validity of our studies. As the saying goes, "what you don't know will hurt you." Random assignment take cares of it.

Matching

Matching is another technique we can use to control for extraneous variables. We must first identify the extraneous variable that can potentially confound the research design. Then we want to rank order the participants on this extraneous variable or list the participants in a ascending or descending order. Participants who are similar on the extraneous variable will be placed into different treatment groups. In other words, they are "matched" on the extraneous variable. Then we can carry out the intervention/treatment as usual. If different treatment groups do show differences on the dependent variable, we would know it is not the extraneous variables because participants are "matched" or equivalent on the extraneous variable. Rather it is more likely to the independent variable (i.e., the treatments) that causes the changes in the dependent variable. Use the example above (self-spending vs. others-spending on happiness) with the same extraneous variable of whether individuals had a happy childhood to begin with. Once we identify this extraneous variable, we do need to first collect some kind of data from the participants to measure how happy their childhood was. Or sometimes, data on the extraneous variables we plan to use may be already available (for example, you want to examine the effect of different types of tutoring on students' performance in Calculus I course and you plan to match them on this extraneous variable: college entrance test scores, which is already collected by the Admissions Office). In either case, getting the data on the identified extraneous variable is a typical step we need to do before matching. So going back to whether individuals had a happy childhood to begin with. Once we have data, we'd sort it in a certain order, for example, from the highest score (meaning participants reporting the happiest childhood) to the lowest score (meaning participants reporting the least happy childhood). We will then identify/match participants with the highest levels of childhood happiness and place them into different treatment groups. Then we go down the scale and match participants with relative high levels of childhood happiness and place them into different treatment groups. We repeat on the descending order until we match participants with the lowest levels of childhood happiness and place them into different treatment groups. By now, each treatment group will have participants with a full range of levels on childhood happiness (which is a strength...thinking about the variation, the representativeness of the sample). The two treatment groups will be similar or equivalent on this extraneous variable. If the treatments, self-spending vs. other-spending, eventually shows the differences on individual happiness, then we know it's not due to how happy their childhood was. We will be more confident it is due to the independent variable.

You may be thinking, but wait we have only taken care of one extraneous variable. What about other extraneous variables? Good thinking.That's exactly correct. We mentioned a few extraneous variables but have only matched them on one. This is the main limitation of matching. You can match participants on more than one extraneous variables, but it's cumbersome, if not impossible, to match them on 10 or 20 extraneous variables. More importantly, the more variables we try to match participants on, the less likely we will have a similar match. In other words, it may be easy to find/match participants on one particular extraneous variable (similar level of childhood happiness), but it's much harder to find/match participants to be similar on 10 different extraneous variables at once.

Holding Extraneous Variable Constant

Holding extraneous variable constant control technique is self-explanatory. We will use participants at one level of extraneous variable only, in other words, holding the extraneous variable constant. Using the same example above, for example we only want to study participants with the low level of childhood happiness. We do need to go through the same steps as in Matching: identifying the extraneous variable that can potentially confound the research design and getting the data on the identified extraneous variable. Once we have the data on childhood happiness scores, we will only include participants on the lower end of childhood happiness scores, then place them into different treatment groups and carry out the study as before. If the condition groups, self-spending vs. other-spending, eventually shows the differences on individual happiness, then we know it's not due to how happy their childhood was (since we already picked those on the lower end of childhood happiness only). We will be more confident it is due to the independent variable.

Similarly to Matching, we have to do this one extraneous variable at a time. As we increase the number of extraneous variables to be held constant, the more difficult it gets. The other limitation is by holding extraneous variable constant, we are excluding a big chunk of participants, in this case, anyone who are NOT low on childhood happiness. This is a major weakness, as we reduce the variability on the spectrum of childhood happiness levels, we decreases the representativeness of the sample and generalizabiliy suffers.

Building Extraneous Variables into Design

The last control technique building extraneous variables into research design is widely used. Like the name suggests, we would identify the extraneous variable that can potentially confound the research design, and include it into the research design by treating it as an independent variable. This control technique takes care of the limitation the previous control technique, holding extraneous variable constant, has. We don't need to excluding participants based on where they stand on the extraneous variable(s). Instead we can include participants with a wide range of levels on the extraneous variable(s). You can include multiple extraneous variables into the design at once. However, the more variables you include in the design, the large the sample size it requires for statistical analyses, which may be difficult to obtain due to limitations of time, staff, cost, access, etc.

This page titled 1.3: Threats to Internal Validity and Different Control Techniques is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **2.4: Threats to Validity** by Jenkins-Smith et al. is licensed CC BY 4.0. Original source: https://bookdown.org/josiesmith/qrmbook.

1.4: Common Experimental Designs

Post-Test Only Control Group Design

In this section we look at some common research designs, the notation used to symbolize them, and then consider the internal validity of the designs. We start with the most basic experimental design, the Post-test Control Group Design (Figure 1.4.3). In this design, subjects are randomly assigned to one of two groups with one group receiving the experimental treatment. This type of design is called between-subjects design because different participants/subjects are assigned into different groups. In the figure below, the symbol R means there is a random assignment to the group. X symbolizes exposure to experimental treatment. O is an observation or measurement. There are advantages to this design in that it is relatively inexpensive and eliminates the threats associated with pre-testing. If randomization worked the (unobserved) pre-test measures would be the same so any differences in the observations would be due to the experimental treatment. The problem is that randomization could fail us, especially if the sample size is small.

$$\begin{array}{ccc} \mathrm{R} & \mathrm{X} & O_1 \\ \mathrm{R} & & O_2 \end{array}$$

Figure 1.4.3: Post-test Only (with a Control Group) Experimental Design

Pretest-Posttest Control Group Design

Many experimental groups are small and many researchers are not comfortable relying on randomization without empirical verification that the groups are the same, so another common between-subjects design is the Pre-test, Post-test Control Group Design (Figure 1.4.4). By conducting a pre-test, we can be sure that the groups are identical when the experiment begins. The disadvantages are that adding groups drives the cost up (and/or decreases the size of the groups).

$$\begin{array}{cccc} \mathrm{R} & O_1 & \mathrm{X} & O_2 \\ \mathrm{R} & O_3 & & O_4 \end{array}$$

Figure 1.4.4: Pre-test, Post-Test (with a Control Group) Experimental Design

This brief discussion illustrates common research designs and the challenges to maximize internal validity. The designs mentioned above involves random assignment. Therefore they are considered strong experimental designs. With these experimental designs, we worry about external validity, but since we have said we seek the ability to make causal statements, the preference is given to research via experimental designs.

This page titled 1.4: Common Experimental Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **2.5: Some Common Designs** by Jenkins-Smith et al. is licensed CC BY 4.0. Original source: https://bookdown.org/josiesmith/qrmbook.

1.5: Common Quasi-Experimental Designs

Recall that when participants in a between-subjects designs are randomly assigned to treatment conditions, the resulting groups are likely to be quite similar. In fact, researchers consider them to be equivalent. When participants are not randomly assigned to conditions, however, the resulting groups are likely to be dissimilar in some ways. For this reason, researchers consider them to be non-equivalent. A **non-equivalent comparison group design** , then, is a between-subjects design in which participants have not been randomly assigned to conditions. There are several types of nonequivalent groups designs we will consider.

Posttest Only Non-equivalent Comparison Group Design

The first non-equivalent groups design we will consider is the **posttest only non-equivalent comparison group design.** In this design, participants in one group are exposed to a treatment, a nonequivalent group is not exposed to the treatment, and then the two groups are compared. Imagine, for example, a researcher who wants to evaluate a new method of teaching fractions to third graders. One way would be to conduct a study with a treatment group consisting of one class of third-grade students and a comparison group consisting of another class of third-grade students. This design would be a nonequivalent groups design because the students are not randomly assigned to classes by the researcher, which means there could be important differences between them. For example, the parents of higher achieving or more motivated students might have been more likely to request that their children be assigned to Ms. Williams's class. Or the principal might have assigned the "troublemakers" to Mr. Jones's class because he is a stronger disciplinarian. Of course, the teachers' styles, and even the classroom environments might be very different and might cause different levels of achievement or motivation among the students. If at the end of the study there was a difference in the two classes' knowledge of fractions, it might have been caused by the difference between the teaching methods—but it might have been caused by any of these confounding variables.

Of course, researchers using a posttest only nonequivalent groups design can take steps to ensure that their groups are as similar as possible. In the present example, the researcher could try to select two classes at the same school, where the students in the two classes have similar scores on a standardized math test and the teachers are the same sex, are close in age, and have similar teaching styles. Taking such steps would increase the internal validity of the study because it would eliminate some of the most important confounding variables. But without true random assignment of the students to conditions, there remains the possibility of other important confounding variables that the researcher was not able to control.

Pretest-Posttest Non-equivalent Comparison Group Design

Another way to improve upon the posttest only nonequivalent groups design is to add a pretest. In the **pretest-posttest non-equivalent comparison group design**, there is a treatment group that is given a pretest, receives a treatment, and then is given a posttest. But at the same time there is a non-equivalent comparison group that is given a pretest, does not receive the treatment, and then is given a posttest. The question, then, is not simply whether participants who receive the treatment improve, but whether they change *more* than participants who do not receive the treatment.

Imagine, for example, that students in one school are given a pretest on their attitudes toward drugs, then are exposed to an anti-drug program, and finally, are given a posttest. Students in a similar school are given the pretest, not exposed to an anti-drug program, and finally, are given a posttest. Again, if students in the treatment condition become more negative toward drugs, this change in attitude could be an effect of the treatment, but it could also be a matter of history or maturation. If it really is an effect of the treatment, then students in the treatment condition should become more negative than students in the comparison condition. But if it is a matter of history (e.g., news of a celebrity drug overdose) or maturation (e.g., improved reasoning), then students in the two conditions would be likely to show similar amounts of change. This type of design does not completely eliminate the possibility of confounding variables, however. Something could occur at one of the schools but not the other (e.g., a student drug overdose), so students at the first school would be affected by it while students at the other school would not.

Returning to the example of evaluating a new measure of teaching third graders, this study could be improved by adding a pretest of students' knowledge of fractions. The changes in scores from pretest to posttest would then be evaluated and compared across conditions to determine whether one group demonstrated a bigger improvement in knowledge of fractions than another. Of course, the teachers' styles, and even the classroom environments might still be very different and might cause different levels of achievement or motivation among the students that are independent of the teaching intervention. Once again, differential history

also represents a potential threat to internal validity. If extremely high level of radon is found in one of the schools causing it to be shut down for a month, then this interruption in teaching could produce differences across groups on posttest scores.

Pretest-Posttest Design With Switching Replication

Some of these non-equivalent comparison group designs can be further improved by adding a switching replication. Using a **pretest-posttest design with switching replication design,** non-equivalent comparison groups are administered a pretest of the dependent variable, then one group receives a treatment while a nonequivalent comparison group does not receive a treatment, the dependent variable is assessed again, and then the treatment is added to the control group, and finally the dependent variable is assessed one last time.

As a concrete example, let's say we wanted to introduce an exercise intervention for the treatment of depression. We recruit one group of patients experiencing depression and a nonequivalent control group of students experiencing depression. We first measure depression levels in both groups, and then we introduce the exercise intervention to the patients experiencing depression, but we hold off on introducing the treatment to the students. We then measure depression levels in both groups. If the treatment is effective we should see a reduction in the depression levels of the patients (who received the treatment) but not in the students (who have not yet received the treatment). Finally, while the group of patients continues to engage in the treatment, we would introduce the treatment to the students with depression. Now and only now should we see the students' levels of depression decrease.

One of the strengths of this design is that it includes a built in replication. In the example given, we would get evidence for the efficacy of the treatment in two different samples (patients and students). Another strength of this design is that it provides more control over history effects. It becomes rather unlikely that some outside event would perfectly coincide with the introduction of the treatment in the first group and with the delayed introduction of the treatment in the second group. For instance, if a change in the weather occurred when we first introduced the treatment to the patients, and this explained their reductions in depression the second time that depression was measured, then we would see depression levels decrease in both the groups. Similarly, the switching replication helps to control for maturation and instrumentation. Both groups would be expected to show the same rates of spontaneous remission of depression and if the instrument for assessing depression happened to change at some point in the study the change would be consistent across both of the groups.

This page titled 1.5: Common Quasi-Experimental Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **8.3: Non-Equivalent Groups Designs** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

1.6: Non-Experimental Research

What Is Non-Experimental Research?

Non-experimental research is research that lacks the manipulation of an independent variable. Rather than manipulating an independent variable, researchers conducting non-experimental research simply measure variables as they naturally occur (in the lab or real world).

Most researchers in social sciences consider the distinction between experimental and non-experimental research to be an extremely important one. This is because although experimental research can provide strong evidence that changes in an independent variable cause differences in a dependent variable, non-experimental research generally cannot. As we will see, however, this inability to make causal conclusions does not mean that non-experimental research is less important than experimental research. It is simply used in cases where experimental research is not able to be carried out.

When to Use Non-Experimental Research

As we saw earlier, experimental research is appropriate when the researcher has a specific research question or hypothesis about a causal relationship between two variables—and it is possible, feasible, and ethical to manipulate the independent variable. It stands to reason, therefore, that non-experimental research is appropriate—even necessary—when these conditions are not met. There are many times in which non-experimental research is preferred, including when:

- the research question or hypothesis relates to a single variable rather than a statistical relationship between two variables (e.g., how accurate are people's first impressions?).
- the research question pertains to a non-causal statistical relationship between variables (e.g., is there a correlation between verbal intelligence and mathematical intelligence?).
- the research question is about a causal relationship, but the independent variable cannot be manipulated or participants cannot be randomly assigned to conditions or orders of conditions for practical or ethical reasons (e.g., does damage to a person's hippocampus impair the formation of long-term memory traces?).

Again, the choice between the experimental and non-experimental approaches is generally dictated by the nature of the research question(s).

Types of Non-Experimental Research

The most common type of non-experimental research conducted in social sciences is correlational research. Correlational research is considered non-experimental because it focuses on the statistical relationship between two variables but does not include the manipulation of an independent variable. More specifically, in **correlational research**, the researcher measures two variables with little or no attempt to control extraneous variables and then assesses the relationship between them. As an example, a researcher interested in the relationship between self-esteem and school achievement could collect data on students' self-esteem and their GPAs to see if the two variables are statistically related. Another example is a researcher interested in the relationship between education levels and annual income can collect data on individuals highest education levels and their annual income to see if the two variables are statistically related. In the first example, both variables are interval (continuous). In the second example, one variable is categorical (educational levels) and one is continuous. In either case, we are studying the variables as they naturally occur or have occurred.

Cross-Sectional, Longitudinal, and Cross-Sequential Studies

When social scientists wish to study change over time (for example, when developmental psychologists wish to study aging) they usually take one of three non-experimental approaches: cross-sectional, longitudinal, or cross-sequential. **Cross-sectional studies** involve comparing two or more pre-existing groups of people (e.g., children at different stages of development). What makes this approach non-experimental is that there is no manipulation of an independent variable and no random assignment of participants to groups. Using this design, developmental psychologists compare groups of people of different ages (e.g., young adults spanning from 18-25 years of age versus older adults spanning 60-75 years of age) on various dependent variables (e.g., memory, depression, life satisfaction). Of course, the primary limitation of using this design to study the effects of aging is that differences between the groups other than age may account for differences in the dependent variable. For instance, differences between the groups may

reflect the generation that people come from (a **cohort effect**) rather than a direct effect of age. For this reason, **longitudinal studies**, in which one group of people is followed over time as they age, offer a superior means of studying the effects of aging. However, longitudinal studies are by definition more time consuming and so require a much greater investment on the part of the researcher and the participants. A third approach, known as **cross-sequential studies**, combines elements of both cross-sectional and longitudinal studies. Rather than measuring differences between people in different age groups or following the same people over a long period of time, researchers adopting this approach choose a smaller period of time during which they follow people in different age groups. For example, they might measure changes over a ten year period among participants who at the start of the study fall into the following age groups: 20 years old, 30 years old, 40 years old, 50 years old, and 60 years old. This design is advantageous because the researcher reaps the immediate benefits of being able to compare the age groups after the first assessment. Further, by following the different age groups over time they can subsequently determine whether the original differences they found across the age groups are due to true age effects or cohort effects.

Internal Validity Revisited

Recall that internal validity is the extent to which the design of a study supports the conclusion that changes in the independent variable caused any observed differences in the dependent variable. Figure 1.6.1 shows how experimental, quasi-experimental, and non-experimental (correlational) research vary in terms of internal validity. Experimental research tends to be highest in internal validity because the use of manipulation (of the independent variable) and control (of extraneous variables) help to rule out alternative explanations for the observed relationships. If the average score on the dependent variable in an experiment differs across conditions, it is quite likely that the independent variable is responsible for that difference. Non-experimental (correlational) research is lowest in internal validity because these designs fail to use manipulation or control. Quasi-experimental research falls in the middle because it contains some, but not all, of the features of a true experiment. For instance, it may fail to use random assignment to assign participants to groups. Imagine, for example, that a researcher finds two similar schools, starts an anti-bullying program in one, and then finds fewer bullying incidents in that “treatment school” than in the “control school.” While a comparison is being made with a control condition, the inability to randomly assign children to schools could still mean that students in the treatment school differed from students in the control school in some other way that could explain the difference in bullying (e.g., there may be a selection effect).

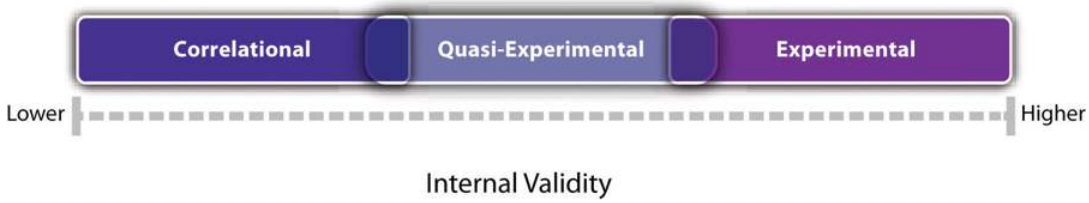


Figure 1.6.1: Internal Validity of Correlation, Quasi-Experimental, and Experimental Studies. Experiments are generally high in internal validity, quasi-experiments lower, and correlation (non-experimental) studies lower still.

Notice also in Figure 1.6.1 that there is some overlap in the internal validity of experiments, quasi-experiments, and correlational (non-experimental) studies. For example, a poorly designed experiment that includes many confounding variables can be lower in internal validity than a well-designed quasi-experiment with no obvious confounding variables.

This page titled 1.6: Non-Experimental Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **6.2: Overview of Non-Experimental Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

2: Between-Subjects Single Factor Design

2.1: Between-Subjects Design
2.2: One-Way ANOVA
2.3: Computing the F statistics
2.4 Post Hoc Tests

This page titled 2: Between-Subjects Single Factor Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

2.1: Between-Subjects Design

Between-Subjects Design

One way to differentiate different research designs is based on how many treatment(s) or condition(s) a participant receives. In a **between-subjects design**, each participant is tested in only one condition. When each participant is tested in more than one treatment or condition, it is considered a different type of research design, within-subjects design, which we will look at later on. Going back to between-subjects design, as an example, a researcher with a sample of 100 university students might assign half of them to write about a traumatic event and the other half write about a neutral event. Or a researcher with a sample of 60 people with severe agoraphobia (fear of open spaces) might assign 20 of them to receive each of three different treatments for that disorder. In both examples, there is only one independent variable, or a single factor. These are called **between-subjects single factor design**.

Random Assignment

Like as we have seen earlier, the primary way that researchers control for extraneous variables across conditions is called **random assignment** , which means using a random process to decide which participants are tested in which conditions. Do not confuse random assignment with random sampling. Random sampling is a method for selecting a sample from a population. Random assignment is a method for assigning participants in a sample to the different conditions, and it is an important element of all experimental research.

In its strictest sense, random assignment should meet two criteria. One is that each participant has an equal chance of being assigned to each condition (e.g., a 50% chance of being assigned to each of two conditions). The second is that each participant is assigned to a condition independently of other participants. Thus one way to assign participants to two conditions would be to flip a coin for each one. If the coin lands heads, the participant is assigned to Condition A, and if it lands tails, the participant is assigned to Condition B. For three conditions, one could use a computer to generate a random integer from 1 to 3 for each participant. If the integer is 1, the participant is assigned to Condition A; if it is 2, the participant is assigned to Condition B; and if it is 3, the participant is assigned to Condition C. In practice, a full sequence of conditions—one for each participant expected to be in the experiment—is usually created ahead of time, and each new participant is assigned to the next condition in the sequence as they are tested. When the procedure is computerized, the computer program often handles the random assignment.

Random assignment is not guaranteed to control all extraneous variables across conditions. The process is random, so it is always possible that just by chance, the participants in one condition might turn out to be substantially older, less tired, more motivated, or less depressed on average than the participants in another condition. However, there are some reasons that this possibility is not a major concern. One is that random assignment works much better than one might expect, especially for large samples. Another reason is that even if random assignment does result in a confounding variable and therefore produces misleading results, this confound is likely to be detected when the experiment is replicated. The upshot is that random assignment to conditions—although not infallible in terms of controlling extraneous variables—is always considered a strength of a research design.

This page titled 2.1: Between-Subjects Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **5.3: Experimental Design** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.2: One-Way ANOVA

The one-way ANOVA is sometimes also called a between-subjects ANOVA, or an independent factor ANOVA. The critical ingredient for a one-way between-subjects ANOVA, is that you have one independent variable, with at least two-levels. When you have one IV with two levels, you can run a t-test. You can also run an ANOVA. Interestingly, they give you almost the exact same results. You will get a p-value from both tests that is identical (they are really doing the same thing under the hood). The t-test gives a t-value as the important sample statistic. The ANOVA gives you the F-value (for Fisher, the inventor of the test, who is one of the most important statisticians in the history of the field) as the important sample statistic. It turns out that t^2 equals F, when there are only two groups in the design.

F is computed directly from the data. In fact, the idea behind F is the same basic idea that goes into making t. Here is the general idea behind the formula, it is again a ratio of the effect we are measuring (in the numerator), and the variation associated with the effect (in the denominator).

$$\text{name of statistic} = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

The difference with F, is that we use variances to describe both the measure of the effect and the measure of error. So, F is a ratio of two variances.

When the variance associated with the effect is the same size as the variance associated with sampling error, we will get two of the same numbers, this will result in an F-value of 1. When the variance due to the effect is larger than the variance associated with sampling error, then F will be greater than 1. When the variance associated with the effect is smaller than the variance associated with sampling error, F will be less than one.

Let's rewrite in plainer English. We are talking about two concepts that we would like to measure from our data. 1) A measure of what we can explain, and 2) a measure of error, or stuff about our data we can't explain. So, the F formula looks like this:

$$\text{F} = \frac{\text{Can Explain}}{\text{Can't Explain}}$$

When we can explain as much as we can't explain, F = 1. This isn't that great of a situation for us to be in. It means we have a lot of uncertainty. When we can explain much more than we can't we are doing a good job, F will be greater than 1. When we can explain less than what we can't, we really can't explain very much, F will be less than 1. That's the concept behind making F.

If you saw an F in the wild, and it was .6. Then you would automatically know the researchers couldn't explain much of their data. If you saw an F of 5, then you would know the researchers could explain 5 times more than the couldn't, that's pretty good. And the point of this is to give you an intuition about the meaning of an F-value, even before you know how to compute it.

This page titled 2.2: One-Way ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **7.2: One-factor ANOVA** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

2.3: Computing the F statistics

Computing the F-value

Fisher's ANOVA is very elegant in my opinion. It starts us off with a big problem we always have with data. We have a lot of numbers, and there is a lot of variation in the numbers, what to do? Wouldn't it be nice to split up the variation into to kinds, or sources. If we could know what parts of the variation were being caused by our experimental manipulation (i.e., the independent variable we choose as researchers), and what parts were being caused by sampling error, we would be making really good progress. We would be able to know if our experimental manipulation was causing more change in the data than sampling error, or chance alone. If we could measure those two parts of the total variation, we could make a ratio, and then we would have an F value. This is what the ANOVA does. It splits the total variation in the data into two parts. The formula is:

Total Variation = Variation due to Manipulation + Variation due to sampling error

This is a nice idea, but it is also vague. We haven't specified our measure of variation. What should we use?

Remember the sums of squares that we used to make the variance and the standard deviation? That's what we'll use. Let's take another look at the formula, using sums of squares for the measure of variation:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

SS Total

The total sums of squares, or $SS\text{Total}$ is a way of thinking about all of the variation in a set of data. It's pretty straightforward to measure. No tricky business. All we do is find the difference between each score and the grand mean, then we square the differences and add them all up.

Let's imagine we had some data in three groups, A, B, and C. For example, we might have 3 scores in each group. The data could look like this:

groups	scores	diff	diff_squared
A	20	13	169
A	11	4	16
A	2	-5	25
B	6	-1	1
B	2	-5	25
B	7	0	0
C	2	-5	25
C	11	4	16
C	2	-5	25
Sums	63	0	302
Means	7	0	33.5555555555556

The data is organized in long format, so that each row is a single score. There are three scores for the A, B, and C groups. The mean of all of the scores is called the **Grand Mean**. It's calculated in the table, the Grand Mean = 7.

We also calculated all of the difference scores **from the Grand Mean**. The difference scores are in the column titled `diff` . Next, we squared the difference scores, and those are in the next column called `diff_squared` .

Remember, the difference scores are a way of measuring variation. They represent how far each number is from the Grand Mean. If the Grand Mean represents our best guess at summarizing the data, the difference scores represent the error between the guess and each actual data point. The only problem with the difference scores is that they sum to zero (because the mean is the balancing point in the data). So, it is convenient to square the difference scores, which gets rid of the negative signs (or values) and turns all of them into positive numbers. The size of the squared difference scores still represents error between the mean and each score. And, the squaring operation exacerbates the differences as the error grows larger (squaring a big number makes a really big number, squaring a small number still makes a smallish number).

OK fine! We have the squared deviations from the grand mean, we know that they represent the error between the grand mean and each score. What next? SUM THEM UP!

When you add up all of the individual squared deviations (difference scores) you get the sums of squares. That's why it's called the sums of squares (SS).

Now, we have the first part of our answer:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

$$SS_{\text{total}} = 302$$

and

$$302 = SS_{\text{Effect}} + SS_{\text{Error}}$$

What next? If you think back to what you learned about algebra, and solving for X, you might notice that we don't really need to find the answers to both missing parts of the equation. We only need one, and we can solve for the other. For example, if we found SS_{Effect}, then we could solve for SS_{Error}.

SS Effect

SS_{Total} gave us a number representing all of the change in our data, how all the scores are different from the grand mean.

What we want to do next is estimate how much of the total change in the data might be due to the experimental manipulation. For example, if we ran an experiment that causes causes change in the measurement, then the means for each group will be different from other. As a result, the manipulation forces change onto the numbers, and this will naturally mean that some part of the total variation in the numbers is caused by the manipulation.

The way to isolate the variation due to the manipulation (also called effect) is to look at the means in each group, and calculate the difference scores between each group mean and the grand mean, and then sum the squared deviations to find SS_{Effect}.

Consider this table, showing the calculations for SS_{Effect}.

groups	scores	means	diff	diff_squared
A	20	11	4	16
A	11	11	4	16
A	2	11	4	16
B	6	5	-2	4
B	2	5	-2	4
B	7	5	-2	4
C	2	5	-2	4
C	11	5	-2	4
C	2	5	-2	4
Sums	63	63	0	72
Means	7	7	0	8

Notice we created a new column called `means`. For example, the mean for group A was 11. You can see there are three 11s, one for each observation in row A. The means for group B and C happen to both be 5. So, the rest of the numbers in the means column are 5s.

What we are doing here is thinking of each score in the data from the viewpoint of the group means. The group means are our best attempt to summarize the data in those groups. From the point of view of the mean, all of the numbers are treated as the same. The mean doesn't know how far off it is from each score, it just knows that all of the scores are centered on the mean.

Now that we have converted each score to it's mean value we can find the differences between each mean score and the grand mean, then square them, then sum them up. We did that, and found that the $SS_{\text{Effect}} = 72$.

SS_{Effect} represents the amount of variation that is caused by differences between the means. I also refer to this as the amount of variation that the researcher can explain (by the means, which represent differences between groups or conditions that were manipulated by the researcher).

Notice also that $SS_{\text{Effect}} = 72$, and that 72 is smaller than $SS_{\text{total}} = 302$. That is very important. SS_{Effect} by definition can never be larger than SS_{total}.

SS Error

Great, we made it to SS Error. We already found SS Total, and SS Effect, so now we can solve for SS Error just like this:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

switching around:

$$SS_{\text{Error}} = SS_{\text{total}} - SS_{\text{Effect}}$$

$$SS_{\text{Error}} = 302 - 72 = 230$$

We could stop here and show you the rest of the ANOVA, we're almost there. But, the next step might not make sense unless we show you how to calculate SS_{Error} directly from the data, rather than just solving for it. We should do this just to double-check our work anyway.

groups	scores	means	diff	diff_squared
A	20	11	-9	81
A	11	11	0	0
A	2	11	9	81
B	6	5	-1	1
B	2	5	3	9
B	7	5	-2	4
C	2	5	3	9
C	11	5	-6	36
C	2	5	3	9
Sums	63	63	0	230
Means	7	7	0	25.5555555555556

Alright, we did almost the same thing as we did to find SS_{Effect}. Can you spot the difference? This time for each score we first found the group mean, then we found the error in the group mean estimate for each score. In other words, the values in the $diff$ column are the differences between each score and it's group mean. The values in the `diff_squared` column are the squared deviations. When we sum up the squared deviations, we get another Sums of Squares, this time it's the SS_{Error}. This is an appropriate name, because these deviations are the ones that the group means can't explain!

Degrees of freedom

Degrees of freedom come into play again with ANOVA. This time, their purpose is a little bit more clear. Dfs can be fairly simple when we are doing a relatively simple ANOVA like this one, but they can become complicated when designs get more complicated.

Let's talk about the degrees of freedom for the SS_{Effect} and SS_{Error}.

The formula for the degrees of freedom for SS_{Effect} is

$df_{\text{Effect}} = \text{Groups} - 1$, where Groups is the number of groups in the design.

In our example, there are 3 groups, so the df is 3-1 = 2. You can think of the df for the effect this way. When we estimate the grand mean (the overall mean), we are taking away a degree of freedom for the group means. Two of the group means can be anything they want (they have complete freedom), but in order for all three to be consistent with the Grand Mean, the last group mean has to be fixed.

The formula for the degrees of freedom for SS_{Error} is

$df_{\text{Error}} = \text{scores} - \text{groups}$, or the number of scores minus the number of groups. We have 9 scores and 3 groups, so our df for the error term is 9-3 = 6. Remember, when we computed the difference score between each score and its group mean, we had to compute three means (one for each group) to do that. So, that reduces the degrees of freedom by 3. 6 of the difference scores could be anything they want, but the last 3 have to be fixed to match the means from the groups.

Mean Squares

OK, so we have the degrees of freedom. What's next? There are two steps left. First we divide the SSes by their respective degrees of freedom to create something new called Mean Squared deviations or Mean Squares. Let's talk about why we do this.

First of all, remember we are trying to accomplish this goal:

$$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

We want to build a ratio that divides a measure of an effect by a measure of error. Perhaps you noticed that we already have a measure of an effect and error! How about the SS_{Effect} and SS_{Error}. They both represent the variation due to the effect, and the leftover variation that is unexplained. Why don't we just do this?

$$\frac{SS_{\text{Effect}}}{SS_{\text{Error}}}$$

Well, of course you could do that. What would happen is you can get some really big and small numbers for your inferential statistic. And, the kind of number you would get wouldn't be readily interpretable like a t value or a z score.

The solution is to **normalize** the SS terms. Don't worry, normalize is just a fancy word for taking the average, or finding the mean. Remember, the SS terms are all sums. And, each sum represents a different number of underlying properties.

For example, the *SS* effect represents the sum of variation for three means in our study. We might ask the question, well, what is the average amount of variation for each mean…You might think to divide *SS* effect by 3, because there are three means, but because we are estimating this property, we divide by the degrees of freedom instead (# groups - 1 = 3-1 = 2). Now we have created something new, it's called the MS_{Effect}.

$$MS_{\text{Effect}} = \frac{SS_{\text{Effect}}}{df_{\text{Effect}}}$$

$$MS_{\text{Effect}} = \frac{72}{2} = 36$$

This might look alien and seem a bit complicated. But, it's just another mean. It's the mean of the sums of squares for the effect. It shows the change in the data due to changes in the means (which are tied to the experimental conditions).

The SS_{Error} represents the sum of variation for nine scores in our study. That's a lot more scores, so the SS_{Error} is often way bigger than than SS_{Effect}. If we left our *SS* error this way and divided them, we would almost always get numbers less than one, because the SS_{Error} is so big. What we need to do is bring it down to the average size. So, we might want to divide our SS_{Error} by 9, after all there were nine scores. However, because we are estimating this property, we divide by the degrees of freedom instead (scores-groups) = 9-3 = 6). Now we have created something new, it's called the MS_{Error}.

$$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}}$$

$$MS_{\text{Error}} = \frac{230}{6} = 38.33$$

Calculate F

Now that we have done all of the hard work, calculating F is easy:

$$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$\text{F} = \frac{MS_{\text{Effect}}}{MS_{\text{Error}}}$$

$$\text{F} = \frac{36}{38.33} = .939$$

Once we have the F statistics, we can find the corresponding significance or *p* value (the statistics program you use for calculation will likely present this value in the output automatically), and compare it to the pre-determined *p* critical value to make a decision about the null hypothesis.

This page titled 2.3: Computing the F statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **7.2: One-factor ANOVA** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

2.4 Post Hoc Tests

If we are able to reject the null hypothesis in the ANOVA F test, it tells us not all group means are equal. In other words, there are some differences among group means. But we don't know where the difference is. This is where post hoc tests come in.

A post hoc test is used only after we find a statistically significant result and need to determine where our differences truly came from. The term “post hoc” comes from the Latin for “after the event”. There are many different post hoc tests that have been developed, and most of them will give us similar answers. We will only focus here on the most commonly used ones. We will also only discuss the concepts behind each and will not worry about calculations.

Bonferroni Test

A Bonferroni test is perhaps the simplest post hoc analysis. A Bonferroni test is a series of *t*-tests performed on each pair of groups. As we discussed earlier, the number of groups quickly grows the number of comparisons, which inflates Type I error rates. To avoid this, a Bonferroni test divides our significance level α by the number of comparisons we are making so that when they are all run, they sum back up to our original Type I error rate. Once we have our new significance level, we simply run independent samples *t*-tests to look for difference between our pairs of groups. This adjustment is sometimes called a Bonferroni Correction, and it is easy to do with statistical program as we compare obtained *p*-values to our the α level.

Tukey’s Honest Significant Difference Test

Tukey’s Honest Significant Difference (HSD) Test is a very popular post hoc analysis. This analysis, like Bonferroni’s, makes adjustments based on the number of comparisons, but it makes adjustments to the test statistic when running the comparisons of two groups. These comparisons give us an estimate of the difference between the groups and a confidence interval for the estimate. We use this confidence interval in the same way that we use a confidence interval for a regular independent samples t-test: if it contains 0.00, the groups are not different, but if it does not contain 0.00 then the groups are different.

Games-Howell Post Hoc Tests

Bonferroni and Tukey HSD post hoc tests can be used when the assumption of homogeneity of variance can be assumed across comparison groups. However, when the assumption of homogeneity of variance is violated, Games-Howell test is the typical post hoc analysis. Like the analyses mentioned above, it makes appropriate adjustments based on the number of tests when comparing differences of each pair of means. Games-Howell tests also provide confidence interval for group mean differences and show whether each pairwise comparison is statistically significant. If the confidence interval contains 0, then the groups are similar. If it doens't contain 0, then the groups are statistically different.

This page titled 2.4 Post Hoc Tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **11.8: Post Hoc Tests** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

3: Between-Subjects Factorial Design

3.1: Factorial Designs
3.2: Factorial ANOVA - Main Effects
3.3: Factorial ANOVA - Interaction Effects
3.4: Factorial ANOVA - Simple Effects
3.5: Partioning the Sum of Squares

This page titled 3: Between-Subjects Factorial Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

3.1: Factorial Designs

Just as it is common for studies in education (or social sciences in general) to include multiple levels of a single independent variable (new teaching method, old teaching method), it is also common for them to include multiple independent variables. Just as including multiple levels of a single independent variable allows one to answer more sophisticated research questions, so too does including multiple independent variables in the same experiment. But including multiple independent variables also allows the researcher to answer questions about whether the effect of one independent variable depends on the level of another. This is referred to as an interaction between the independent variables. As we will see, interactions are often among the most interesting results in empirical research.

Factorial Designs

Overview

By far the most common approach to including multiple independent variables (which are also called factors or ways) in an experiment is the factorial design. In a **between-subjects factorial design**, each level of one independent variable is combined with each level of the others to produce all possible combinations. Each combination, then, becomes a condition in the experiment. Imagine, for example, an experiment on the effect of cell phone use (yes vs. no) and time of day (day vs. night) on driving ability. This is shown in the **factorial design table** in Figure **3.1.1**. The columns of the table represent cell phone use, and the rows represent time of day. The four cells of the table represent the four possible combinations or conditions: using a cell phone during the day, not using a cell phone during the day, using a cell phone at night, and not using a cell phone at night. This particular design is referred to as a 2 × 2 (read “two-by-two”) factorial design because it combines two variables, each of which has two levels.

If one of the independent variables had a third level (e.g., using a handheld cell phone, using a hands-free cell phone, and not using a cell phone), then it would be a 3 × 2 factorial design, and there would be six distinct conditions. Notice that the number of possible conditions is the product of the numbers of levels. A 2 × 2 factorial design has four conditions, a 3 × 2 factorial design has six conditions, a 4 × 5 factorial design would have 20 conditions, and so on. Also notice that each number in the notation represents one factor, one independent variable. So by looking at how many numbers are in the notation, you can determine how many independent variables there are in the experiment. 2 x 2, 3 x 3, and 2 x 3 designs all have two numbers in the notation and therefore all have two independent variables. Some people refer to these are two-way factorial ANOVA. The numerical value of each of the numbers represents the number of levels of each independent variable. A 2 means that the independent variable has two levels, a 3 means that the independent variable has three levels, a 4 means it has four levels, etc. To illustrate, a 3 x 3 design has two independent variables, each with three levels, while a 2 x 2 x 2 design has three independent variables, each with two levels.

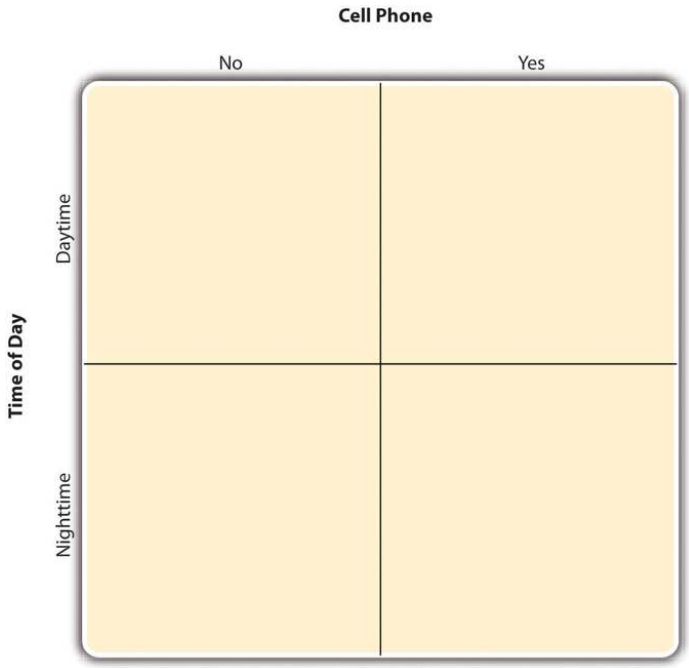


Figure **3.1.1**: Factorial Design Table Representing a 2 (cellphone use: yes or no) × 2 (time of day: daytime or nighttime) Factorial Design

In principle, factorial designs can include any number of independent variables with any number of levels. For example, an experiment could include the type of psychotherapy (cognitive vs. behavioral), the length of the psychotherapy (2 weeks vs. 2 months), and the sex of the psychotherapist (female vs. male). This would be a 2 × 2 × 2 factorial design and would have eight conditions. Figure **3.1.2** shows one way to represent this design. In practice, it is unusual for there to be more than three independent variables with more than two or three levels each. This is for at least two reasons: For one, the number of conditions can quickly become unmanageable. For example, adding a fourth independent variable with three levels (e.g., therapist experience: low vs. medium vs. high) to the current example would make it a 2 × 2 × 2 × 3 factorial design with 24 distinct conditions. Second, the number of participants required to populate all of these conditions (while maintaining a reasonable ability to detect a real underlying effect) can render the design unfeasible. As a result, in the remainder of this section, we will focus on designs with two independent variables. The general principles discussed here extend in a straightforward way to more complex factorial designs.

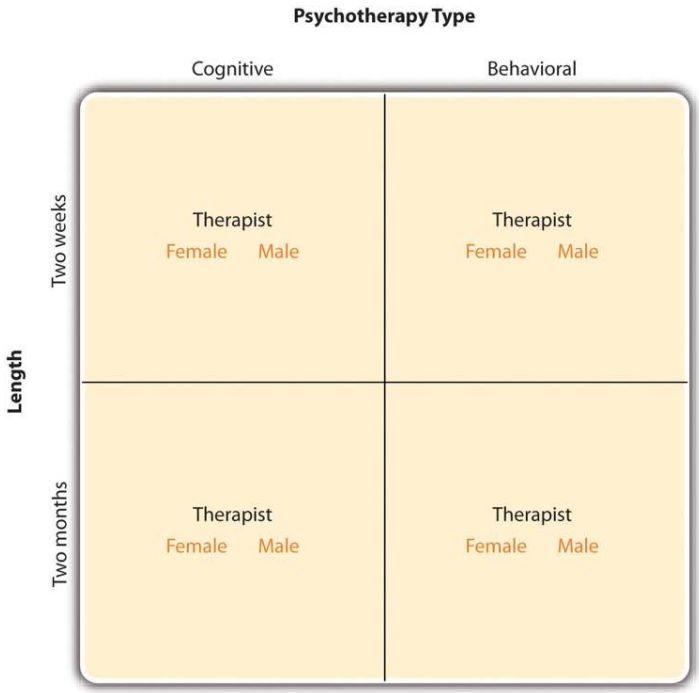


Figure **3.1.2**: Factorial Design Table Representing a 2 (psychotherapy type) × 2 (psychotherapy length) × 2 (sex of psychotherapist) Factorial Design

Assigning Participants to Conditions

Recall that in a between-subjects single factor design, each participant is tested in only one condition. In a **between-subjects factorial design**, all of the independent variables are manipulated between subjects. For example, all participants could be tested either while using a cell phone *or* while not using a cell phone and either during the day *or* during the night. This would mean that each participant would be tested in one and only one condition.

Since factorial designs have more than one independent variable, it is also possible to manipulate one independent variable between subjects and another within subjects. This is called a **mixed factorial design**. For example, a researcher might choose to treat cell phone use as a within-subjects factor by testing the same participants both while using a cell phone and while not using a cell phone. But they might choose to treat time of day as a between-subjects factor by testing each participant either during the day or during the night (perhaps because this only requires them to come in for testing once). Thus each participant in this mixed design would be tested in two of the four conditions. This is a complex design with complex statistical analyses. In the remainder of this section, we will focus on between-subjects factorial designs only. Also, regardless of the design, the actual assignment of participants to conditions is typically done randomly.

Non-Manipulated Independent Variables

In many factorial designs, one of the independent variables is a **non-manipulated independent variable**. The researcher measures it but does not manipulate it. Figure **3.1.2** above is an example in which one independent variable, the sex of the psychotherapist (female vs. male) cannot be manipulated. Another example is a study by Halle Brown and colleagues in which participants were exposed to several words that they were later asked to recall (Brown, Kosslyn, Delamater, Fama, & Barsky, 1999)[1]. The manipulated independent variable was the type of word. Some were negative health-related words (e.g., *tumor, coronary*), and others were not health related (e.g., *election, geometry*). The non-manipulated independent variable was whether participants were high or low in hypochondriasis (excessive concern with ordinary bodily symptoms). The result of this study was that the participants high in hypochondriasis were better than those low in hypochondriasis at recalling the health-related words, but they were no better at recalling the non-health-related words.

Such studies are extremely common, and there are several points worth making about them. First, non-manipulated independent variables are usually participant background variables (self-esteem, gender, and so on), and as such, they are by definition between-subjects factors. For example, people are either low in self-esteem or high in self-esteem; they cannot be tested in both of these

conditions. Second, such studies are generally considered to be experiments as long as at least one independent variable is manipulated, regardless of how many non-manipulated independent variables are included. Third, it is important to remember that causal conclusions can *only* be drawn about the manipulated independent variable. Thus it is important to be aware of which variables in a study are manipulated and which are not.

Non-Experimental Studies With Factorial Designs

Thus far we have seen that factorial experiments can include manipulated independent variables or a combination of manipulated and non-manipulated independent variables. But factorial designs can also include *only* non-manipulated independent variables, in which case they are no longer experiment designs, but are instead non-experimental in nature. Consider a hypothetical study in which a researcher simply measures both the moods and the self-esteem of several participants—categorizing them as having either a positive or negative mood and as being either high or low in self-esteem—along with their willingness to have unprotected sex. This can be conceptualized as a 2 × 2 factorial design with mood (positive vs. negative) and self-esteem (high vs. low) as non-manipulated between-subjects factors. Willingness to have unprotected sex is the dependent variable.

Again, because neither independent variable in this example was manipulated, it is a non-experimental study rather than an experimental design. This is important because, as always, one must be cautious about inferring causality from non-experimental studies because of the threats of potential confounding variables. For example, an effect of participants' moods on their willingness to have unprotected sex might be caused by any other variable that happens to be correlated with their moods.

References

1. Brown, H. D., Kosslyn, S. M., Delamater, B., Fama, A., & Barsky, A. J. (1999). Perceptual and memory biases for health-related information in hypochondriacal individuals. *Journal of Psychosomatic Research*, *47*, 67–78. ↵

This page titled 3.1: Factorial Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **9.2: Setting Up a Factorial Experiment** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.2: Factorial ANOVA - Main Effects

Graphing the Results of Factorial Designs

The results of between-subjects factorial designs with two independent variables can be graphed by representing one independent variable on the *x*-axis and representing the other by using different colored bars or lines. (The *y*-axis is always reserved for the dependent variable.) Figure **3.2.1** shows results for two hypothetical factorial experiments. The top panel shows the results of a 2 × 2 design. Time of day (day vs. night) is represented by different locations on the *x*-axis, and cell phone use (no vs. yes) is represented by different-colored bars. (It would also be possible to represent cell phone use on the *x*-axis and time of day as different-colored bars. The choice comes down to which way seems to communicate the results most clearly.) The bottom panel of Figure **3.2.1** shows the results of a 4 × 2 design in which one of the variables, psychotherapy length, is represented along the *x*-axis, and the other variable (psychotherapy type) is represented by differently formatted lines. This is a line graph rather than a bar graph because the variable on the x-axis is quantitative with a small number of distinct levels. Line graphs are also appropriate when representing measurements made over a time interval on the *x*-axis.

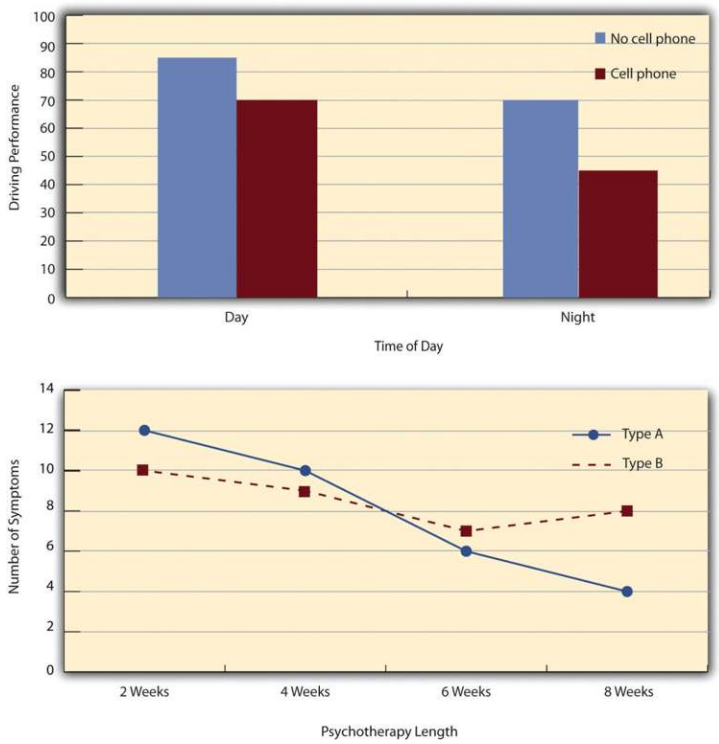


Figure **3.2.1**: Two Ways to Plot the Results of a Factorial Experiment With Two Independent Variables

Main Effects

In factorial designs, there are three kinds of results that are of interest: main effects, interaction effects, and simple effects. A **main effect** is the effect of one independent variable on the dependent variable—averaging across the levels of the other independent variable. Thus there is one main effect to consider for each independent variable in the study. The top panel of Figure **3.2.1** shows a main effect of cell phone use because driving performance was better, on average, when participants were not using cell phones than when they were. The blue bars are, on average, higher than the red bars. It also shows a main effect of time of day because driving performance was better during the day than during the night—both when participants were using cell phones and when they were not. Main effects are independent of each other in the sense that whether or not there is a main effect of one independent variable says nothing about whether or not there is a main effect of the other. The bottom panel of Figure **3.2.1**, for example, shows a clear main effect of psychotherapy length. The longer the psychotherapy, the better it worked.

This page titled 3.2: Factorial ANOVA - Main Effects is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **9.3: Interpreting the Results of a Factorial Experiment** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.3: Factorial ANOVA - Interaction Effects

Interaction Effects

There is an **interaction** effect (or just “interaction”) when the effect of one independent variable depends on the level of another. Although this might seem complicated, you already have an intuitive understanding of interactions. As an everyday example, assume your friend asks you to go to a movie with another friend. Your response to her is, “well it depends on which movie you are going to see and who else is coming.” You really want to see the big blockbuster summer hit but have little interest in seeing the cheesy romantic comedy. In other words, there is a main effect of type of movie on your decision. If your decision to go to see either of these movies further depends on who she is bringing with her then there is an interaction. For instance, if you will go to see the big blockbuster summer hit if she brings her cool friend you get along with, but you will not go to this movie if she brings her boring friend, then there is an interaction.

Let’s now consider some examples of interactions from research. It probably would not surprise you to hear that the effect of receiving psychotherapy is stronger among people who are highly motivated to change than among people who are not motivated to change. This is an interaction because the effect of one independent variable (whether or not one receives psychotherapy) depends on the level of another (motivation to change).

In many studies, the primary research question is about an interaction. The study by Brown and her colleagues was inspired by the idea that people with hypochondriasis are especially attentive to any negative health-related information. This led to the hypothesis that people high in hypochondriasis would recall negative health-related words more accurately than people low in hypochondriasis but recall non-health-related words about the same as people low in hypochondriasis. And of course, this is exactly what happened in this study.

Types of Interactions

The effect of one independent variable can depend on the level of the other in several different ways. Note that these different types of interactions are simply nice to know in case you read about them in the manuscripts. But the ultimate goal is to understand what interaction means (when the effect of one IV on DV depends on the level of another IV) and what it looks like (see the panels and bars below). Here we go.

First, there can be **spreading interactions**. Examples of spreading interactions are shown in the top two panels of Figure 3.3.2. In the top panel, independent variable “B” has an effect at level 1 of independent variable “A” (there is a difference in the height of the blue and red bars on the left side of the graph) but no effect at level 2 of independent variable “A.” (there is no difference in the height of the blue and red bars on the right side of the graph). In the middle panel, independent variable “B” has a stronger effect at level 1 of independent variable “A” than at level 2 (there is a larger difference in the height of the blue and red bars on the left side of the graph and a smaller difference in the height of the blue and red bars on the right side of the graph). So to summarize, for spreading interactions there is an effect of one independent variable at one level of the other independent variable and there is either a weak effect or no effect of that independent variable at the other level of the other independent variable.

The second type of interaction that can be found is a **cross-over interaction.** A cross-over interaction is depicted in the bottom panel of Figure 3.3.2, independent variable “B” again has an effect at both levels of independent variable “A,” but the effects are in opposite directions. An example of a crossover interaction comes from a study by Kathy Gilliland on the effect of caffeine on the verbal test scores of introverts and extraverts (Gilliland, 1980). Introverts perform better than extraverts when they have not ingested any caffeine. But extraverts perform better than introverts when they have ingested 4 mg of caffeine per kilogram of body weight.

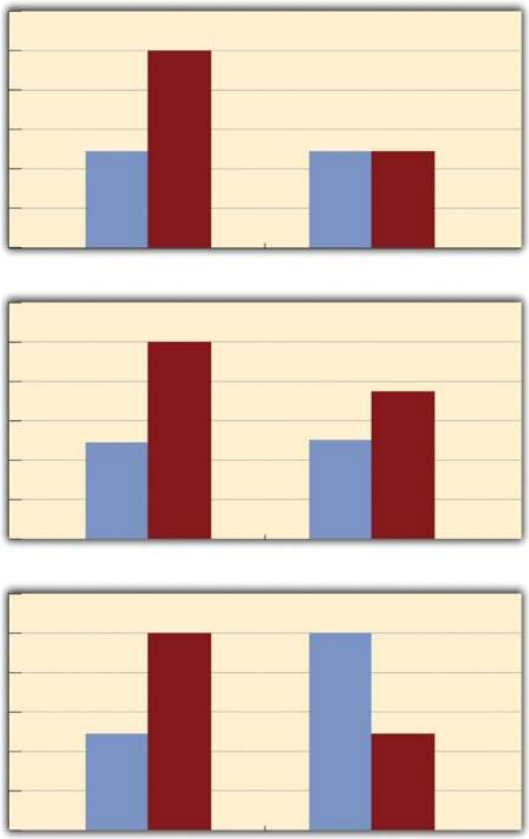


Figure 3.3.2: Bar Graphs Showing Three Types of Interactions. In the top panel, one independent variable has an effect at one level of the second independent variable but not at the other. In the middle panel, one independent variable has a stronger effect at one level of the second independent variable than at the other. In the bottom panel, one independent variable has the opposite effect at one level of the second independent variable than at the other.

Figure 3.3.3 shows examples of these same kinds of interactions when one of the independent variables is quantitative and the results are plotted in a line graph. Note that the top two figures depict the two kinds of spreading interactions that can be found while the bottom figure depicts a crossover interaction (the two lines literally “cross over” each other).

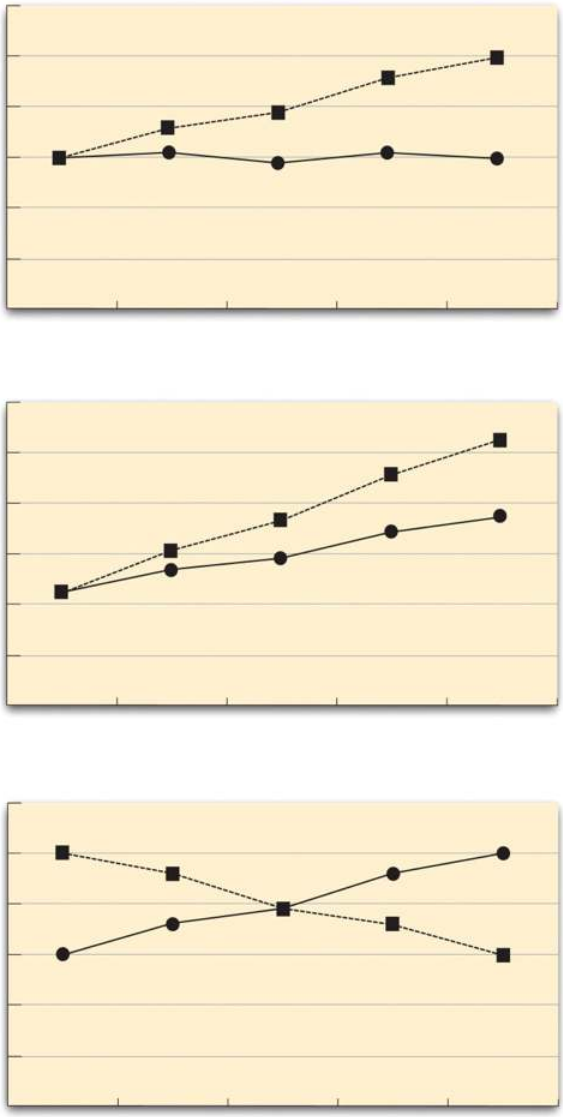


Figure 3.3.3: Line Graphs Showing Different Types of Interactions. In the top panel, one independent variable has an effect at one level of the second independent variable but not at the other. In the middle panel, one independent variable has a stronger effect at one level of the second independent variable than at the other. In the bottom panel, one independent variable has the opposite effect at one level of the second independent variable than at the other.

References

Gilliland, K. (1980). The interactive effect of introversion-extraversion with caffeine induced arousal on verbal performance. *Journal of Research in Personality, 14*, 482–492.

This page titled 3.3: Factorial ANOVA - Interaction Effects is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **9.3: Interpreting the Results of a Factorial Experiment** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.4: Factorial ANOVA - Simple Effects

Simple Effects

When researchers find an interaction it suggests that the main effects may be a bit misleading. Still using Gilliland's study (1980) mentioned on the last page, think of the example of a crossover interaction where introverts were found to perform better on a test of verbal test performance than extraverts when they had not ingested any caffeine, but extraverts were found to perform better than introverts when they had ingested 4 mg of caffeine per kilogram of body weight. To examine the main effect of caffeine consumption, the researchers would have averaged across introversion and extraversion and simply looked at whether overall those who ingested caffeine had better or worse verbal memory test performance. Because the positive effect of caffeine on extraverts would be wiped out by the negative effects of caffeine on the introverts, no main effect of caffeine consumption would have been found. Similarly, to examine the main effect of personality, the researchers would have averaged across the levels of the caffeine variable to look at the effects of personality (introversion vs. extraversion) independent of caffeine. In this case, the positive effects extraversion in the caffeine condition would be wiped out by the negative effects of extraversion in the no caffeine condition. Does the absence of any main effects mean that there is no effect of caffeine and no effect of personality? No of course not.

The presence of the interaction indicates that the story is more complicated, that the effects of caffeine on verbal test performance depend on personality. This is where simple effects come into play. **Simple effects** are a way of breaking down the interaction to figure out precisely what is going on. An interaction simply informs us that the effects of at least one independent variable depend on the level of another independent variable. Whenever an interaction is detected, researchers need to conduct additional analyses to determine where that interaction is coming from. Of course one may be able to visualize and interpret the interaction on a graph but a simple effects analysis provides researchers with a more sophisticated means of breaking down the interaction. Specifically, a simple effects analysis allows researchers to determine the effects of each independent variable at each level of the other independent variable. So while the researchers would average across the two levels of the personality variable to examine the effects of caffeine on verbal test performance in a main effects analysis, for a simple effects analysis the researchers would examine the effects of caffeine in introverts and then examine the effects of caffeine in extraverts. As we saw previously, the researchers also examined the effects of personality in the no caffeine condition and found that in this condition introverts performed better than extraverts. Finally, they examined the effects of personality in the caffeine condition and found that extraverts performed better than introverts in this condition. For a 2 x 2 design like this, there will be two main effects the researchers can explore and four simple effects.

As described previously, Brown and colleagues found an interaction between type of words (health related or not health related) and hypochondriasis (high or low) on word recall. To break down this interaction using simple effects analyses they examined the effect of hypochondriasis at each level of word type. Specifically, they examined the effect of hypochondriasis on recall of health-related words and then they subsequently examined the effect of hypochondriasis on recall of non-health related words. They found that people high in hypochondriasis were able to recall more health-related words than people low in hypochondriasis. In contrast, there was no effect of hypochondriasis on the recall of non-health related words.

Once again examining simple effects provides a means of breaking down the interaction and therefore it is only necessary to conduct these analyses when an interaction is present. When there is no interaction then the main effects will tell the complete and accurate story. To summarize, rather than averaging across the levels of the other independent variable, as is done in a main effects analysis, simple effects analyses are used to examine the effects of each independent variable at each level of the other independent variable(s). So a researcher using a 2×2 design with four conditions would need to look at 2 main effects and 4 simple effects. A researcher using a 2×3 design with six conditions would need to look at 2 main effects and 5 simple effects, while a researcher using a 3×3 design with nine conditions would need to look at 2 main effects and 6 simple effects. As you can see, while the number of main effects depends simply on the number of independent variables included (one main effect can be explored for each independent variable), the number of simple effects analyses depends on the number of levels of the independent variables (because a separate analysis of each independent variable is conducted at each level of the other independent variable).

This page titled 3.4: Factorial ANOVA - Simple Effects is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **9.3: Interpreting the Results of a Factorial Experiment** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.5: Partioning the Sum of Squares

While we are not going into calculating the ANOVA table in factorial ANOVA, we will look at how to **partition the sums of squares**. ANOVAs are all about partitioning the sums of squares. We already did some partitioning in the last chapter. What do we mean by partitioning?

Imagine you had a big empty house with no rooms in it. What would happen if you partitioned the house? What would you be doing? One way to partition the house is to split it up into different rooms. You can do this by adding new walls and making little rooms everywhere. That's what partitioning means, to split up.

The act of partitioning, or splitting up, is the core idea of ANOVA. To use the house analogy. Our total sums of squares (SS Total) is our big empty house. We want to split it up into little rooms. Before we partitioned SS Total using this formula:

$$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

Remember, the SS_{Effect} was the variance we could attribute to the means of the different groups, and SS_{Error} was the leftover variance that we couldn't explain. SS_{Effect} and SS_{Error} are the partitions of SS_{TOTAL}, they are the little rooms.

Now let's see how we can split up the house in factorial ANOVA. We will still use the example in section 3.1, the experiment on the effect of cell phone use (yes vs. no) and time of day (day vs. night) on driving ability. This is a 2 x 2 ANOVA. Cell phone use is IV1. Time of day is IV2. Remember the logic of the ANOVA is to partition the variance into different parts. The SS formula for the between-subjects 2 x 2 ANOVA looks like this:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

Now we split up the house into a lot more little rooms, $SS_{\text{Effect IV1}}$, $SS_{\text{Effect IV2}}$, $SS_{\text{Effect IV1xIV2}}$, and SS_{Error}.

SS Total

We calculate the grand mean (mean of all of the score). Then, we calculate the differences between each score and the grand mean. We square the difference scores, and sum them up. That is SS_{Total}, as always.

SS Cell Phone Use (Main Effect - IV1)

We need to compute the SS for the main effect for cell phone use. This step is essentially the same as how we calculated the SS effect in single-factor between-subjects ANOVA. the key is when we calculate the main effect of cell phone use, we ignore the other IV time of day. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two cell phone use conditions (yes vs. no). Then we treat each score as if it was the mean for it's respective cell phone use condition. We find the differences between each cell phone use condition mean and the grand mean. Then we square the differences and sum them up. That is $SS_{\text{Cell phone use}}$. Again, the key is when we calculate the main effect of one independent variable, we ignore the other independent variable or pretend it doesn't exist.

SS Time of Day (Main Effect - IV2)

We need to compute the SS for the main effect for time of day. Similarly, when we calculate the main effect of time of day, we ignore the other IV cell phone use. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two time of day conditions (day vs. night). Then we treat each score as if it was the mean for it's respective time of day condition. We find the differences between each time of day condition mean and the grand mean. Then we square the differences and sum them up. That is $SS_{\text{Time of day}}$. Again, this step is essentially the same as how we calculated the SS effect in single-factor between-subjects ANOVA. The key is when we calculate the main effect of one independent variable, we ignore the other independent variable or pretend it doesn't exist.

SS Cell Phone Use by Time (Interaction Effect)

We need to compute the SS for the interaction effect between cell phone use and time of day. This is the new thing that we do in an ANOVA with more than one IV. How do we calculate the variation explained by the interaction?

The heart of the question is something like this. Do the individual means for each of the four conditions do something a little bit different than the group means for both of the independent variables.

For example, let's say the overall mean for all of the scores in the no cell phone group to be 6.6. Now, was the mean for each no cell phone group in the whole design a 6.6? For example, in the day group, was the mean for no cell phone condition also 6.6? Let's say the answer is no, it was 9.6. How about the night group? Was the mean for the night condition in the no cell phone group 6.6? Let's say the answer is no, it was 3.6. The mean of 9.6 and 3.6 is 6.6. If there was no hint of an interaction, we would expect that the means for the no cell phone condition in both levels of the time of day group would be the same, they would both be 6.6. However, when there is an interaction, the means for the no cell phone group will depend on the levels of the group from another IV. In this case, it looks like there is an interaction because the means are different from 6.6, they are 9.6 for the day condition and 3.6 for the night conditions. This is extra-variance that is not explained by the mean for the no cell phone condition. We want to capture this extra variance and sum it up. Then we will have measure of the portion of the variance that is due to the interaction between the cell phone use and time of day conditions.

What we will do is this. We will find the four condition means. Then we will see how much additional variation they explain beyond the group means for cell phone use and time of day. To do this we treat each score as the condition mean for that score. Then we subtract the mean for the cell phone use group, and the mean for the time of day group, and then we add the grand mean. This gives us the unique variation that is due to the interaction.

Here is a formula to describe the process for each score:

$$\bar{X}_{\text{condition}} - \bar{X}_{\text{IV1}} - \bar{X}_{\text{IV2}} + \bar{X}_{\text{Grand Mean}}$$

We would apply this formula to the calculation of each of the differences scores. We then square the difference scores, and sum them up to get $SS_{\text{Interaction}}$.

SS Error

The last thing we need to find is the SS Error. We can solve for that because we found everything else in this formula:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

Even though this textbook meant to explain things in a step by step way, you are probably tired from reading how to work out the 2x2 ANOVA by hand. I have already shown you how to compute the SS for error before, so we will not do the full example here. In essence, not every score in a particular condition group is the same. We subtract each score (from a particular condition) and from the condition mean, square the differences, and add them up. Then we do this same step for each condition group, and combined, we will get SS Error.

Like mentioned earlier, we are not going into details of ANOVA calculations here. Please refer to the lecture for those. The key is to know the difference between one-way ANOVA and factorial ANOVA. The advantage of factorial ANOVA over multiple one-way ANOVA is its ability to examine the potential Interaction effects.

This page titled 3.5: Partioning the Sum of Squares is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: Partioning the Sums of Squares** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.
- **9.6: 2x2 Between-subjects ANOVA** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

CHAPTER OVERVIEW

4: Between-Subjects Design with a Control Variable

4.1: Overview of the Control Variable
4.2: Radomized Block Design
4.4: Analysis of Covariance (ANCOVA)

This page titled 4: Between-Subjects Design with a Control Variable is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

4.1: Overview of the Control Variable

In the first chapter, we talked about many different threats to internal validity of a research design and one of the control techniques is to build the extraneous variable into our research design. In this chapter, we will extend between-subjects design by looking at different ways to add in an extraneous or a control variable. Why do we need to add control variables? And what criteria should we use when selecting control variables? The main reason we want to include control variables is that the control variables are having an effect on the dependent variable we are studying. Since control variables are not the independent variables in our research, they could potentially confound the results of the study if left unattended. In other words, they can impose threats to the internal validity of the research design. By taking some measures to include the control variables, we are minimizing their effect on the dependent variable, which gives us more confidence to claim it is the independent variable, not the control variable, that causes changes in the dependent variable.

Using the example from the previous chapter, let's say we are conducting an experiment on the effect of cell phone use (yes vs. no) on driving ability. The independent variable is cell phone use with two treatment conditions (yes or no) and the dependent variable is driving ability. A potential control variable would be driving experience as driving experience is most likely to have an impact on driving ability. In order to reduce the potential threat driving experience has on driving ability, we can add it into our study as a control variable. Although it is not the focus of the study, control variable IS a part of your study as we know it influences the outcome variable. By including driving experience into our study, we can minimizing its effect on our research design, and be more confident it is the cell phone use, not driving experience, that leads to changes in driving ability. Therefore adding control variables can increase the internal validity of the research design.

How do we select control variables? Any variables can be potential control variables as long as there is good theoretical or empirical evidence(s) to show they influence the outcome variables. The nature of the variable is not a concern. The control variable can be categorical or continuous. Using the same example above, to measure driving experience, we can ask participants to identify which following level of driving experience represents them the best - seasoned, intermediate, or inexperienced. Or we can ask participants to identify how many months they have driven. Or if you are concerned about the accuracy of participants' own estimation, you can ask participants the age they received their driving license and do the calculation yourself. Regardless how you measure it, as long as the control variable is solid, as in it indeed influences the outcome variable, it can be included in the research study.

Then how do we use different types of control variables? There are two major ways to use control variables. One is randomized block design, which uses control variables at the design stage when we actively set up the experiment. Randomized block design typically uses categorical control variables. The other one is analysis of covariance, which uses control variables at the data analysis stage when we analyze the statistical data. Analysis of covariance typically uses continuous variables. We will look at each of them closely in the following sections.

This page titled 4.1: Overview of the Control Variable is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

4.2: Radomized Block Design

In randomized block design, the control technique is done through the design itself. First the researchers need to identify a potential control variable that most likely has an effect on the dependent variable. Researchers will group participants who are similar on this control variable together into blocks. This control variable is called a blocking variable in the randomized block design. The purpose of the randomized block design is to form groups that are homogeneous on the blocking variable, and thus can be compared with each other based on the independent variable.

How to Carry Out a Randomized Block Design

Using the example from the last section, we are conducting an experiment on the effect of cell phone use (yes vs. no) on driving ability. The independent variable is cell phone use and the dependent variable is driving ability. A potential control variable would be driving experience as it most likely has an effect on driving ability. Driving experience in this case can be used as a blocking variable. We will then divide up the participants into multiple groups or blocks, so that those in each block share similar driving experiences. For example, let's say we decide to place them into three blocks based on driving experience - seasoned; intermediate; inexperienced. You may wonder how we decide on three blocks. We will get to this in a little bit.

Once the participants are placed into blocks based on the blocking variable, we would carry out the experiment to examine the effect of cell phone use (yes vs. no) on driving ability. Those in each block will be randomly assigned into either treatment conditions of the independent variable, cell phone use (yes vs. no). As we carry out the study, participants' driving ability will be assessed. We can determine whether cell phone use has an effet on driving ability after controlling for driving experience. Again, since seasoned drivers are randomly assigned into both cell phone use conditions, as well as those with intermediate driving experience, and little driving experience, we already took care of the effect of the blocking variable, driving experience; so we are confident varied driving experience is not competing with the independent variable, cell phone use, in explaining the outcome variable, driving ability.

No Blocking Variable vs. Having a Blocking Variable

Randomized block design still uses ANOVA analysis, called randomized block ANOVA. When participants are placed into a block, we anticipate them to be homogeneous on the control variable, or the blocking variable. In other words, there should be less variability within each block on the control variable, compared to the variability in the entire sample if there were no control variable. Again going back to the same example, seasoned drivers may still vary in their driving experiences, but they are more similar to each other, thus as a subgroup would have less variability in driving experience than that of the entire sample. This is the key advantage of randomized block design. Less within-block variability reduces the error term and makes estimate of the treatment effect more robust or efficient, compared to without the blocking variable.

Without the blocking variable, ANOVA has two parts of variance, SS intervention and SS error. All variance that can't be explained by the independent variable is considered error. By adding the blocking variable, we partition out some of the error variance and attribute it to the blocking variable. As a results, there will be three parts of the variance in randomized block ANOVA, SS intervention, SS block, and SS error, and together they make up SS total. In doing so, the error variance will be reduced since part of the error variance is now explained by the blocking variable. In F tests, we look at the ratio of effect and error. When the numerator (i.e., error) decreases, the calculated F is going to be larger. We will achieve a smaller P obtained value, and are more likely to reject the null hypothesis. In other words, good blocking variables decreases error, which increases statistical power.

While it is true randomized block design could be more powerful than single-factor between-subjects randomized design, this comes with an important condition. That is we must select good blocking variables. As you have seen from the procedure described above, it shouldn't come as a surprise that it is very difficult to include many blocking variables. For one, the procedure becomes cumbersome. Second, as the number of blocking variables increases, we need to create more blocks. Each block has to have a sufficient group size for statistical analysis, therefore, the sample size can increase rather quickly. The selection of blocking variables should be based on previous literature.

Furthermore, as mentioned early, researchers have to decide how many blocks should there be, once you have selected the blocking variable. We want to carefully consider whether the blocks are homogeneous. In the case of driving experience as a blocking variable, are three groups sufficient? Can we reasonably believe that seasoned drivers are more similar to each other than they are

to those with intermediate or little driving experience? It is a subjective decision left up to the researchers. If the blocks aren't homogeneous, their variability will not be less than that of the entire sample. In that situation, randomized block design can decreases the statistical power and thus be worse than a simple single-factor between-subjects randomized design. Again, your best bet on finding an optimal number of blocks is from theoretical and/or empirical evidences.

Assumptions of Randomized Block Design/ANOVA

Randomized block ANOVA shares all assumptions of regular ANOVA. There are two additional assumptions unique to randomized block ANOVA.

First, the blocking variable should have an effect on the dependent variable. Just like in the example above, driving experience has an impact on driving ability. This is why we picked this particular variable as the blocking variable in the first place. Even though we are not interested in the blocking variable, we know based on the theoretical and/or empirical evidence that the blocking variable has an impact on the dependent variable. By adding it into the model, we reduce its likelihood to confound the effect of the treatment (independent variable) on the dependent variable. If the blocking variable (or the groupings of the block) has little effect on the dependent variable, the results will be biased and inaccurate. We are less likely to detect an effect of the treatment on the outcome variable if there is one.

Second, the blocking variable cannot interact with the independent variable. In the example above, the cell phone use treatment (yes vs. no) cannot interact with driving experience. This means the effect of cell phone use treatment (yes vs. no) on the dependent variable, driving ability, should not be influenced by the level of driving experience (seasoned, intermediate, inexperienced). In other words, the impact of cell phone use treatment (yes vs. no) on the dependent variable should be similar regardless of the level of driving experience. If this assumption is violated, randomized block ANOVA should not performed. One possible alternative is to treat it like a factorial ANOVA where the independent variables are allowed to interact with each other.

This page titled 4.2: Radomized Block Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

4.4: Analysis of Covariance (ANCOVA)

As mentioned earlier this chapter, there are two ways to add control variables into a research study. One is through design, such as randomized block design. The other is through statistical control, known as analysis of covariance. The control variables are called covariates. Covariates usually have an impact on the dependent variable and thus can be included into an ANOVA analysis. We can this technique analysis of covariance (ANCOVA).

Analysis of Covariance (ANCOVA)

Recall in randomized block design/ANOVA, we would utilize the control variable in the design stage by sorting participants and grouping them into different blocks based on the control variable. By doing that, the participants will be similar on the control variable before they are assigned into different treatment or intervention groups. However, in ANCOVA, we don't do anything with the control variable, or the covariate, in the design stage. Instead, we simply collect the data on the covariate along with other data we would collect from participants, and analyze the covariate during the data analysis stage. That's why this is considered a statistical control technique.

Still using the same example as used in randomized block design, we are conducting an experiment on the effect of cell phone use (yes vs. no) on driving ability and include driving experience (as measured by months of driving) as a control variable. Instead of grouping participants into different blocks based on their driving experience, in ANCOVA, we would treat driving experience as a covariate and simply collect data on it and analyze it using ANCOVA technique. Or another example is, say we want to test the impact of different teaching methods on students' performance in an introductory calculus course. Previous research has established that math scores on college entrance test impacts students' performance in calculus courses. Therefore, we would include ACT math scores as the covariate. Notice that, in both cases, driving experience (as measured by months of driving) and ACT math scores, are intervally scaled variables. This is typically the expectations of ANCOVA, to have the covariate on an interval scale. As a matter of fact, we would expect there to be a linear relationship between the covariate and the dependent variable. More on this in a little bit.

ANCOVA is an extension of ANOVA. The main advantage of using ANCOVA over using ANOVA is that by adding covariates into the study/model, we are minimizing the effect of the covariates on the dependent variable. Recall, the covariates are known to have an influence on the dependent variable, which is why they are included in the study in the first place. By controlling for the effect of covariate, we are reducing its threat to confound the results, and this gives us more confidence to establish that the intervention, or the independent variable, causes the change in the dependent variable. In the example above, by controlling for the effect of driving experience on driving ability, we are more certain it is the cell phone use that causes the change in driving ability.

How to Use ANCOVA

The analyses of ANCOVA is fairly complex. Without getting into the details of computations, this section provides a brief overview of how to use ANCOVA.

Just like ANOVA, ANCOVA uses Fisher's F test. Therefore, the key to understand ANCOVA is still partitioning of the variance. When adding a covariate into the study, it essentially becomes another predictor in the model, even though it is not the researchers' main focus or interest. So we will partition out the variance that can be explained by this variable. As a results, there will be three parts of the variance in ANCOVA, SS Intervention, SS Covariate, and SS Error, and together they make up SS Total. If we were to compare this to ANOVA, the difference is SS Covariate. Without the covariate, ANOVA has two parts of variance, SS Intervention and SS Error. By adding the covariate, we partition out some of the error variance and attribute it to the covariate. In doing so, the error variance will be reduced. As we have seen many times, in F tests, we look at the ratio of effect and error. When the numerator (i.e., error) decreases, the calculated F is going to be larger. We will achieve a smaller P obtained value, and are more likely to reject the null hypothesis. In other words, good covariates decreases error, which increases statistical power. This is another main advantage of ANCOVA (besides control mentioned above), assuming the covariate we selected is a decent one based on theoretical/empirical evidences.

As mentioned above, covariates should be either interval or ratio. That's because ANCOVA essentially uses a linear regression model. With the linear model, the computation can be rather complicated. But conceptually, by including the covariate into the model, ANCOVA adjusts each group mean on the outcome variable. Using the same example of studying the effect of cell phone

use (yes vs. no) on driving ability with driving experience as a covariate, it is possible that one of the treatment group (the no cell phone use) happens to be higher on the covariate, that is, have more driving experience than the other treatment group (cell phone use). Accounting for that, ANCOVA will lower the no-cell-phone-use group's average score on the dependent variable, driving ability. You probably have guessed, since the other group (cell phone use) is lower on the covariate, that is, less driving experience, ANCOVA will increase its group average score on the dependent variable, driving ability. Mathematically, this allows us to compare the means of the treatment groups at the mean/average value of the covariate. In other words, the treatment groups in the study will be "adjusted" for the linear model, so that the "playing field is leveled". By doing so, ANCOVA allows us to find the best estimates of how different treatment groups would have scored on the dependent variable if they all had statistically equivalent means on the covariate.

Assumptions of ANCOVA

ANCOVA shares the assumptions of ANOVA. In addition, there are three assumptions that are unique to ANCOVA.

First, in ANCOVA, the independent variable and the covariate must be independent from each other. In other words, the levels or groups of the intervention/treatment should have no influence on the covariate. In the example above, this means cell phone use treatment groups (yes vs. no) should be independent of, or have no influence on, driving experience. When this assumption is violated, the effect of the independent variable (treatment) and the effect of the covariate overlaps. Translate that into statistical calculations, the treatment and the covariate would share some of the variance. This will skew the analysis and make the results biased. When covariate driving experience is affected by the independent variable cell phone use treatments (yes vs. no), adding driving experience into the model as a covariate does not control for the differences between treatment groups on the dependent variable driving ability. The ANCOVA results will be inaccurate.

Second, the relationship between the covariate and the dependent variable must be linear. In the example above, this means driving experience and driving ability is expected to have a linear relationship. It is critical to first examine the nature of the relationship between the covariate and the dependent variable, for example through scatter plots, before performing ANCOVA. If the relationship is not linear, the adjustments ANCOVA makes will be biased and the results will be inaccurate.

In addition, these regression lines (on the covariate and the dependent variable) from different treatment groups must be parallel to each other. In other words, different treatment groups should have similar slopes. In the above example, this mean for both groups (cell phone use and no cell phone use), the slope for the relationship between driving experience and driving ability should be similar. This assumption is called homogeneity of regression slopes. This is one of the most important assumptions of ANCOVA as it allows us to "adjust" for the group means. If this assumption is violated, it means there is an interaction between the independent variable and the covariate. In this case, ANCOVA will be biased and the results will be inaccurate.

References

Field, A. (2013). *Discovering Statistics Using IBM Statistics.* London: Sage Publications.

This page titled 4.4: Analysis of Covariance (ANCOVA) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

CHAPTER OVERVIEW

5: Within-Subjects Design

5.1: Within-Subjects Design
5.2: Partioning the Sums of Squares
5.3: Repeated Measures ANOVA
5.4: More Notes on Repeated Measures ANOVA

This page titled 5: Within-Subjects Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

5.1: Within-Subjects Design

The designs we have covered so far in the book are all between-subjects designs, meaning each participant is assigned into one condition and being tested under one condition only. When we compare whether the treatments yield different outcomes, we are compare between participants or subjects from different treatment conditions, thus the name between-subjects design. In this chapter, we will look at a different type of design, where each participants will be assigned into multiple treatments. It is called within-subjects design.

Within-Subjects Design

In a **within-subjects design**, each participant is tested under all conditions. Consider an experiment on the effect of a defendant's physical attractiveness on judgments of his guilt. Again, in a between-subjects experiment, one group of participants would be shown an attractive defendant and asked to judge his guilt, and another group of participants would be shown an unattractive defendant and asked to judge his guilt. In a within-subjects experiment, however, the same group of participants would judge the guilt of both an attractive and an unattractive defendant.

The primary advantage of this approach is that it provides maximum control of extraneous participant variables. Participants in all conditions have the same mean IQ, same socioeconomic status, same number of siblings, and so on—because they are the very same people. Within-subjects design also makes it possible to use statistical procedures that remove the effect of these extraneous participant variables on the dependent variable and therefore make the data less "noisy" and the effect of the independent variable easier to detect. We will look more closely at this idea later shortly. However, not all experiments can use a within-subjects design nor would it be desirable to do so.

Carryover Effects and Counterbalancing

Remember in the first chapter, we talked about different common threats that can jeopardize the internal validity of our research designs. There are certain threats that are associated with within-subject design. The primary one in within-subjects designs is order effects. An **order effect** occurs when participants' responses in the various conditions are affected by the order of conditions to which they were exposed. One type of order effect is a carryover effect. A **carryover effect** is an effect of being tested in one condition on participants' behavior in later conditions. For example, participants may perform a task better in later conditions because they have had a chance to practice it. Or it could be the opposite where participants may perform a task worse in later conditions because they become tired or bored. Being tested in one condition can also change how participants perceive stimuli or interpret their task in later conditions. This type of effect is called a **context effect (or contrast effect)**. For example, an average-looking defendant might be judged more harshly when participants have just judged an attractive defendant than when they have just judged an unattractive defendant. Within-subjects experiments also make it easier for participants to guess the hypothesis. For example, a participant who is asked to judge the guilt of an attractive defendant and then is asked to judge the guilt of an unattractive defendant is likely to guess that the hypothesis is that defendant attractiveness affects judgments of guilt. This knowledge could lead the participant to judge the unattractive defendant more harshly because he thinks this is what he is expected to do. Or it could make participants judge the two defendants similarly in an effort to be "fair."

Carryover effects can be interesting in their own right. (Does the attractiveness of one person depend on the attractiveness of other people that we have seen recently?) But when they are not the focus of the research, carryover effects can be problematic. Imagine, for example, that participants judge the guilt of an attractive defendant and then judge the guilt of an unattractive defendant. If they judge the unattractive defendant more harshly, this might be because of his unattractiveness. But it could be instead that they judge him more harshly because they are becoming bored or tired. In other words, the order of the conditions is a confounding variable. The attractive condition is always the first condition and the unattractive condition the second. Thus any difference between the conditions in terms of the dependent variable could be caused by the order of the conditions and not the independent variable itself.

There is a solution to the problem of order effects that can be used in many situations. It is **counterbalancing**, which means testing different participants in different orders. Using counterbalancing, the researcher(s) can have an equal or similar number of participants complete each possible order of conditions. For example, half of the participants would be tested in the attractive defendant condition followed by the unattractive defendant condition, and others half would be tested in the unattractive condition followed by the attractive condition. With three conditions (A, B, C), there would be six different orders (ABC, ACB, BAC, BCA, CAB, and CBA), so some participants would be tested in each of the six orders. With four conditions, there would be 24 different

orders; with five conditions there would be 120 possible orders. With counterbalancing, participants are assigned to orders randomly, using the techniques we have already discussed. Here, instead of randomly assigning to conditions, they are randomly assigned to different orders of conditions.

There are two ways to think about what counterbalancing accomplishes. One is that it controls the order of conditions so that it is no longer a confounding variable. Instead of the attractive condition always being first and the unattractive condition always being second, the attractive condition comes first for some participants and second for others. Likewise, the unattractive condition comes first for some participants and second for others. Thus any overall difference in the dependent variable between the two conditions cannot have been caused by the order of conditions. A second way to think about what counterbalancing accomplishes is that if there are carryover effects, it makes it possible to detect them. One can analyze the data separately for each order to see whether it had an effect.

Between-Subjects or Within-Subjects?

Now you have seen both between-subjects designs and within-subjects designs. Which one should we use? Almost every experiment can be conducted using either a between-subjects design or a within-subjects design. This possibility means that researchers must choose between the two approaches based on their relative merits for the particular situation.

Between-subjects designs have the advantage of being conceptually simpler and requiring less testing time per participant. They also avoid carryover effects without the need for counterbalancing. Within-subjects designs have the advantage of controlling extraneous participant variables, which generally reduces noise in the data and makes it easier to detect any effect of the independent variable upon the dependent variable. Within-subjects designs also require fewer participants than between-subjects designs to detect an effect of the same size.

A good rule of thumb, then, is that if it is possible to conduct a within-subjects experiment (with proper counterbalancing) in the time that is available per participant—and you have no serious concerns about carryover effects—this design is probably the best option. If a within-subjects design would be difficult or impossible to carry out, then you should consider a between-subjects design instead. For example, if you were testing participants in a doctor's waiting room or shoppers in line at a grocery store, you might not have enough time to test each participant in all conditions and therefore would opt for a between-subjects design. Or imagine you were trying to reduce people's level of prejudice by having them interact with someone of another race. A within-subjects design with counterbalancing would require testing some participants in the treatment condition first and then in a control condition. But if the treatment works and reduces people's level of prejudice, then they would no longer be suitable for testing in the control condition. This difficulty is true for many designs that involve a treatment meant to produce long-term change in participants' behavior (e.g., studies testing the effectiveness of new teaching/advising/counseling techniques). Clearly, a between-subjects design would be necessary here.

Remember also that using one type of design does not preclude using the other type in a different study. There is no reason that a researcher could not use both a between-subjects design and a within-subjects design to answer the same research question. In fact, professional researchers often take exactly this type of mixed methods approach.

This page titled 5.1: Within-Subjects Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

- **5.3: Experimental Design** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

5.2: Partioning the Sums of Squares

Time to partition the sums of squares again. Remember the act of partitioning, or splitting up, the variance is the core idea of ANOVA. To continue using the house analogy, our total sums of squares (SS Total) is our big empty house. We want to split it up into little rooms. Before in the between-subjects ANOVA, we partitioned SS Total using this formula:

$$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

The SS_{Effect} was the variance we could attribute to the means of the different groups, and SS_{Error} was the leftover variance that we couldn't explain. SS_{Effect} and SS_{Error} are the partitions of SS_{TOTAL}, they are the little rooms.

In the between-subjects ANOVA above, we got to split SS_{TOTAL} into two parts. What is most interesting about the repeated-measures design, is that we get to split SS_{TOTAL} into three parts, there's one more partition. Can you guess what the new partition is? Hint: whenever we have a new way to calculate means in our design, we can always create a partition for those new means. What are the new means in the repeated measures design?

Here is the formula for partitioning SS_{TOTAL} in a repeated-measures ANOVA:

$$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Subjects}} + SS_{\text{Error}}$$

We've added SS_{Subjects} as the new idea in the formula. What's the idea here? Well, because each subject or participant was measured in each condition, we have a new set of means. These are the means for each subject or participant, collapsed across the conditions. For example, subject 1 has a mean (mean of their scores in conditions A, B, and C); subject 2 has a mean (mean of their scores in conditions A, B, and C); and subject 3 has a mean (mean of their scores in conditions A, B, and C). There are three subject means, one for each subject, collapsed across the conditions. And, we can now estimate the portion of the total variance that is explained by these subject means.

Before we go into the calculations, it's important to pause and compare the differences of how the sum of squares are partitioned in between-subjects ANOVA vs. within-subjects ANOVA.

Recall, in between-subjects ANOVA, we use different words to describe parts of the ANOVA (which can be really confusing). For example, we described the SS formula for a between-subjects ANOVA like this:

$$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

The very same formula is often written differently, using the words between and within in place of effect and error, it looks like this:

$$SS_{\text{TOTAL}} = SS_{\text{Between}} + SS_{\text{Within}}$$

Here, SS_{Between} (which we have been calling SS_{Effect}) refers to variation **between** the group means, that's why it is called SS_{Between}. Second, and most important, SS_{Within} (which we have been calling SS_{Error}), refers to the leftover variation within each group mean. Specifically, it is the variation between each group mean and each score within that group. Remember, for each group mean, every score is probably off a little bit from the mean. So, the scores within each group have some variation. This is the within group variation, and it is why the leftover error that we can't explain is often called SS_{Within}.

Perhaps a picture will help to clear things up.

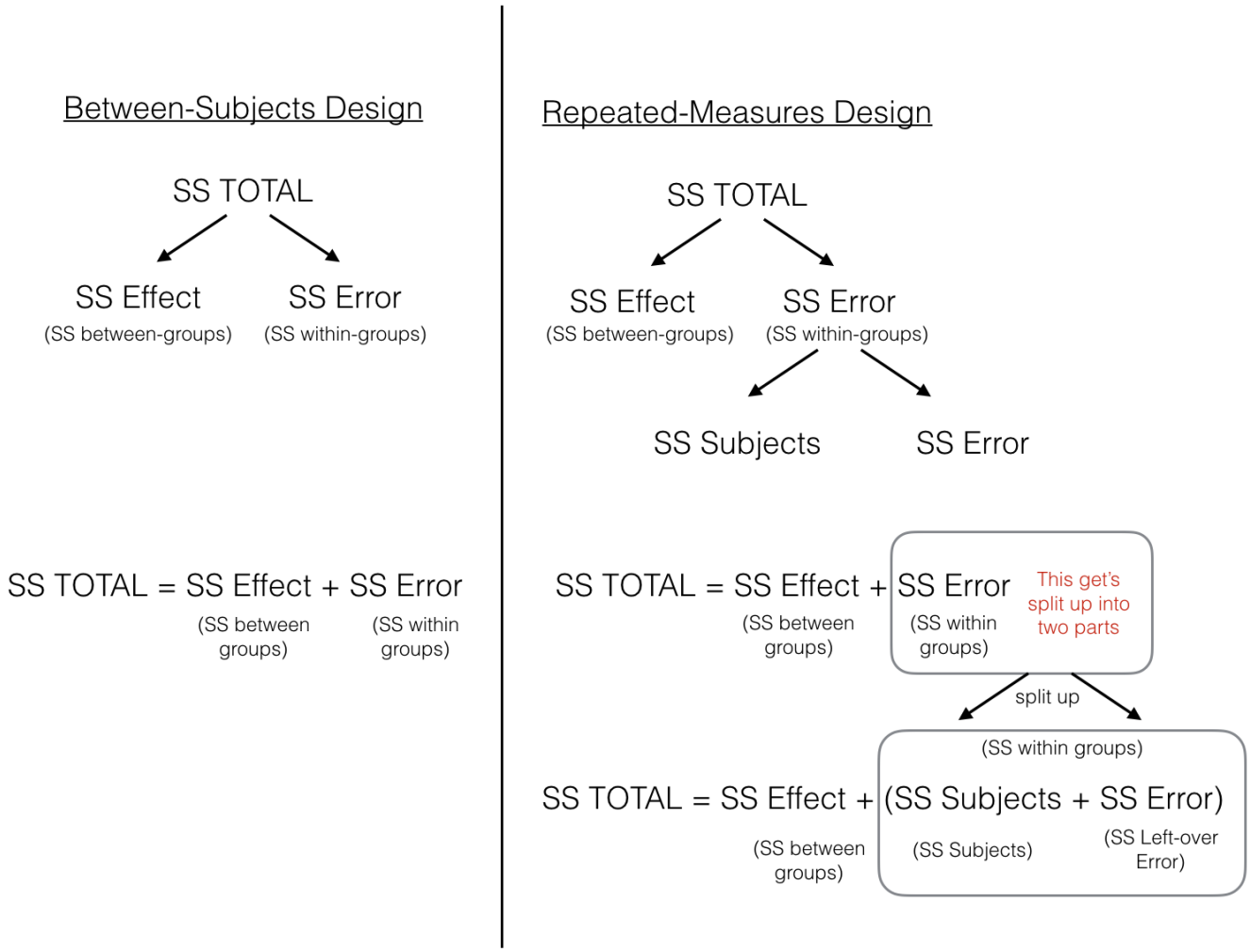


Figure 5.2.1: Illustration showing how the total sums of squares are partitioned differently for a between-subjects versus repeated-measures design.

The Figure 5.2.1 lines up the partitioning of the Sums of Squares for both between-subjects and repeated-measures designs. In both designs, SS_{Total} is first split up into two pieces $SS_{\text{Effect (between-groups)}}$ and $SS_{\text{Error (within-groups)}}$. At this point, both ANOVAs are the same. In the repeated measures case we split the $SS_{\text{Error (within-groups)}}$ into two more littler parts, which we call $SS_{\text{Subjects (error variation about the subject mean)}}$ and $SS_{\text{Error (left-over variation we can't explain)}}$.

The critical feature of the repeated-measures ANOVA, is that the SS_{Error} that we will later use to compute the MS (Mean Squared) in the denominator for the F-value, is smaller in a repeated-measures design, compared to a between subjects design. This is because the $SS_{\text{Error (within-groups)}}$ is split into two parts, $SS_{\text{Subjects (error variation about the subject mean)}}$ and $SS_{\text{Error (left-over variation we can't explain)}}$.

To make this more clear, here is Figure 5.2.2:

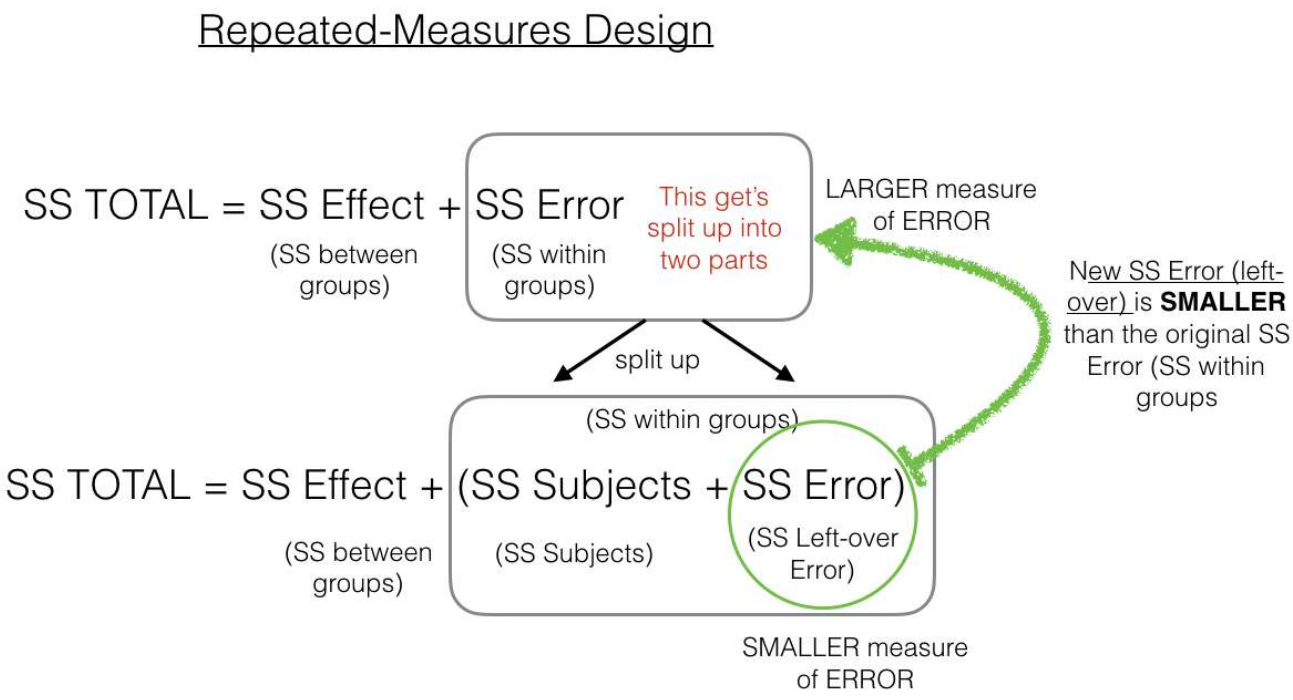


Figure 5.2.2: Close-up showing that the Error term is split into two parts in the repeated measures design.

As we point out, the $SS_{\text{Error (left-over)}}$ in the green circle will be a smaller number than the $SS_{\text{Error (within-group)}}$. That's because we are able to subtract out the SS_{Subjects} part of the $SS_{\text{Error (within-group)}}$. This can have the effect of producing larger F-values when using a repeated-measures design compared to a between-subjects design, which is more likely to yield smaller P obtained values and allow us to reject the null hypothesis.

This page titled 5.2: Partioning the Sums of Squares is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: Partioning the Sums of Squares** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

5.3: Repeated Measures ANOVA

Let's take a look at the things we need to find out to make the ANOVA table. What we need to do is calculate all the SSes that we did before for the between-subjects ANOVA. That means the next three steps are identical to the ones you did before. In fact, I will just basically copy the next three steps to find SS_{TOTAL}, SS_{Effect}, and $SS_{\text{Error (within-conditions)}}$. After that we will talk about splitting up $SS_{\text{Error (within-conditions)}}$ into two parts, this is the new thing for this chapter. Here we go!

SS Total

The total sums of squares, or $SS\text{Total}$ measures the total variation in a set of data. All we do is find the difference between each score and the grand mean, then we square the differences and add them all up.

subjects	conditions	scores	diff	diff_squared
1	A	20	13	169
2	A	11	4	16
3	A	2	-5	25
1	B	6	-1	1
2	B	2	-5	25
3	B	7	0	0
1	C	2	-5	25
2	C	11	4	16
3	C	2	-5	25
Sums		63	0	302
Means		7	0	33.5555555555556

The mean of all of the scores is called the **Grand Mean**. It's calculated in the table, the Grand Mean = 7.

We also calculated all of the difference scores **from the Grand Mean**. The difference scores are in the column titled `diff`. Next, we squared the difference scores, and those are in the next column called `diff_squared`.

When you add up all of the individual squared deviations (difference scores) you get the sums of squares. That's why it's called the sums of squares (SS).

Now, we have the first part of our answer:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

$$SS_{\text{total}} = 302$$

and

$$302 = SS_{\text{Effect}} + SS_{\text{Error}}$$

SS Effect

SS_{Total} gave us a number representing all of the change in our data, how they all are different from the grand mean.

What we want to do next is estimate how much of the total change in the data might be due to the experimental intervention/treatment. For example, if we ran an experiment that causes change in the measurement, then the means for each group will be different from other, and the scores in each group will be different from the other groups. As a result, the

intervention/treatment forces change onto the numbers, and this will naturally mean that some part of the total variation in the numbers is caused by the intervention.

The way to isolate the variation due to the treatment conditions (also called effect) is to look at the means in each group, and the calculate the difference scores between each group mean and the grand mean, and then the squared deviations to find the sum for SS_{Effect}.

Consider this table, showing the calculations for SS_{Effect}.

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	4	16
2	A	11	11	4	16
3	A	2	11	4	16
1	B	6	5	-2	4
2	B	2	5	-2	4
3	B	7	5	-2	4
1	C	2	5	-2	4
2	C	11	5	-2	4
3	C	2	5	-2	4
Sums		63	63	0	72
Means		7	7	0	8

Notice we created a new column called `means`, these are the means for each condition, A, B, and C.

SS_{Effect} represents the amount of variation that is caused by differences between the means. The `diff` column is the difference between each condition mean and the grand mean, so for the first row, we have 11-7 = 4, and so on.

We found that $SS_{\text{Effect}} = 72$.

SS Error (within-conditions)

Great, we made it to SS Error. We already found SS Total, and SS Effect, so now we can solve for SS Error just like this:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error (within-conditions)}}$$

switching around:

$$SS_{\text{Error}} = SS_{\text{total}} - SS_{\text{Effect}}$$

$$SS_{\text{Error (within conditions)}} = 302 - 72 = 230$$

Or, we could compute $SS_{\text{Error (within conditions)}}$ directly from the data:

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	-9	81
2	A	11	11	0	0
3	A	2	11	9	81
1	B	6	5	-1	1
2	B	2	5	3	9
3	B	7	5	-2	4
1	C	2	5	3	9
2	C	11	5	-6	36
3	C	2	5	3	9
Sums		63	63	0	230
Means		7	7	0	25.5555555555556

When we compute $SS_{\text{Error (within conditions)}}$ directly, we find the difference between each score and the condition mean for that score. This gives us the remaining error variation around the condition mean, that the condition mean does not explain.

SS Subjects

Now we are ready to calculate new partition, called SS_{Subjects}. We first find the means for each subject. For subject 1, this is the mean of their scores across Conditions A, B, and C. The mean for subject 1 is 9.33 (repeating). Notice there is going to be some rounding error here, that's OK for now.

The `means` column now shows all of the subject means. We then find the difference between each subject mean and the grand mean. These deviations are shown in the `diff` column. Then we square the deviations, and sum them up.

subjects	conditions	scores	means	diff	diff_squared
1	A	20	9.33	2.33	5.4289
2	A	11	8	1	1
3	A	2	3.66	-3.34	11.1556
1	B	6	9.33	2.33	5.4289
2	B	2	8	1	1
3	B	7	3.66	-3.34	11.1556
1	C	2	9.33	2.33	5.4289
2	C	11	8	1	1
3	C	2	3.66	-3.34	11.1556
Sums		63	62.97	-0.0299999999999994	52.7535
Means		7	6.99666666666667	-0.00333333333333326	5.8615

We found that the sum of the squared deviations SS_{Subjects} = 52.75. Note again, this has some small rounding error because some of the subject means had repeating decimal places, and did not divide evenly.

We can see the effect of the rounding error if we look at the sum and mean in the `diff` column. We know these should be both zero, because the Grand mean is the balancing point in the data. The sum and mean are both very close to zero, but they are not zero because of rounding error.

SS Error (left-over)

Now we can do the last thing. Remember we wanted to split up the $SS_{\text{Error (within conditions)}}$ into two parts, SS_{Subjects} and $SS_{\text{Error (left-over)}}$. Because we have already calculate $SS_{\text{Error (within conditions)}}$ and SS_{Subjects}, we can solve for $SS_{\text{Error (left-over)}}$:

$$SS_{\text{Error (left-over)}} = SS_{\text{Error (within conditions)}} - SS_{\text{Subjects}}$$

$$SS_{\text{Error (left-over)}} = SS_{\text{Error (within conditions)}} - SS_{\text{Subjects}} = 230 - 52.75 = 177.25$$

We have finished our job of computing the sums of squares that we need in order to do the next steps, which include computing the MS (mean squared) for the effect and the error term. Once we do that, we can find the F-value, which is the ratio of the two MS's.

Compute the MS

Calculating the MS's (mean squares) that we need for the F-value involves the same general steps as last time. We divide each SS by the degrees of freedom for the SS.

The degrees of freedom for SS_{Effect} are the same as before, the number of conditions - 1. We have three conditions, so the df is 2. Now we can compute the MS_{Effect}.

$$MS_{\text{Effect}} = \frac{SS_{\text{Effect}}}{df} = \frac{72}{2} = 36$$

The degrees of freedom for SS_{Subjects} are the number of subjects - 1. We have three participants, so the df is 2. Now we can compute the MS_{Subjects}.

$$MS_{\text{Subjects}} = \frac{SS_{\text{Subjects}}}{df} = \frac{52.75}{2} = 26.375$$

The degrees of freedom for $SS_{\text{Error (left-over)}}$ are different than before, they are the (number of subjects - 1) multiplied by the (number of conditions -1). We have 3 subjects and three conditions, so $(3-1)*(3-1) = 2*2 = 4$. Or you can go the easy way, it is the difference of degrees of freedom between Total, Effect/Treatment, and Subjects.

$$df_{\text{Error (left-over)}} = df_{\text{Total}} - df_{\text{Effect}} - df_{\text{Subjects}} = 8 - 2 - 2 = 4$$

Regardless, now we can compute the $MS_{\text{Error (left-over)}}$.

$$MS_{\text{Error (left-over)}} = \frac{SS_{\text{Error (left-over)}}}{df} = \frac{177.33}{4} = 44.33$$

Compute F

We just found the two MS's that we need to compute F. We went through all of this to compute F for our data, so let's do it:

$$F = \frac{MS_{\text{Effect}}}{MS_{\text{Error (left-over)}}} = \frac{36}{44.33} = 0.812$$

And, there we have it!

p-value

We already conducted the repeated-measures ANOVA using statistical program and found the p-value associated with our F-value is 0.505. We might write up the results of our experiment and say that the main effect condition was not significant, F(2, 4) = 0.812, p = 0.505.

What does this statement mean? Remember, that the p-value represents the probability of getting the F value we observed or larger under the null (assuming that the samples come from the same distribution, the assumption of no differences). So, we know that an F-value of 0.812 or larger happens fairly often by chance (when there are no real differences), in fact it happens 50.5% of the time. As a result, we do not reject the idea that any differences in the means we have observed could have been produced by chance.

This page titled 5.3: Repeated Measures ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **8.3: Calculating the RM ANOVA** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

5.4: More Notes on Repeated Measures ANOVA

Repeated Measures ANOVAs have some special properties that are worth knowing about. The main special property is that the error term used to for the F-value (the MS in the denominator) will always be smaller than the error term used for the F-value the ANOVA for a between-subjects design. We discussed this earlier. It is smaller, because we subtract out the error associated with the subject means.

This can have the consequence of generally making F-values in repeated measures designs larger than F-values in between-subjects designs. When the number in the bottom of the F formula is generally smaller, it will generally make the resulting ratio a larger number. That's what happens when you make the number in the bottom smaller.

Because big F values usually let us reject the idea that differences in our means are due to chance, the repeated-measures ANOVA becomes a more sensitive test of the differences (its F-values are usually larger). This is a major advantage of using repeated-measures ANOVA, and within-subjects designs. The other advantage we have already mentioned before is that within-subjects design requires a smaller sample size.

This page titled 5.4: More Notes on Repeated Measures ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **8.4: Things worth knowing** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

CHAPTER OVERVIEW

6: Multiple Linear Regression

6.1: Simple Linear Equation Refresher
6.2: Multiple Regression
6.3: Regression Coefficients
6.4: Effect Size
6.5: Multicollinearity

This page titled 6: Multiple Linear Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang.

6.1: Simple Linear Equation Refresher

This chapter is all about regression. If you recall, linear regression for two variables is based on a linear equation with one independent variable. Both variables are intervally scaled. The equation has the form:

$$y = a + bx$$

where a and b are constant numbers.

The variable **x is the independent variable, and y is the dependent variable.** Another way to think about this equation is a statement of cause and effect. The X variable is the cause and the Y variable is the hypothesized effect. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable. Be aware that the words of "cause and effect" here are used strictly in a statistical sense. It by no means indicates a causal relationship from the research design perspective.

Example 6.1.1

The following examples are linear equations.

$y = 3 + 2x$

$y = -0.01 + 1.2x$

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation

Example 6.1.2

Graph the equation $y = -1 + 2x$.

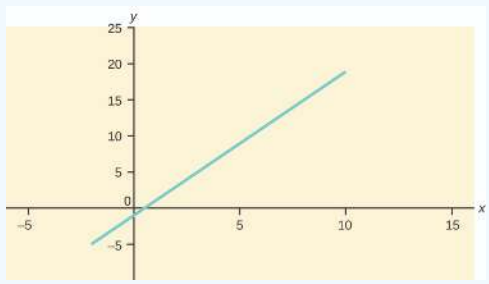


Figure **6.1.3**

Example 6.1.3

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Answer

Solution 13.3

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

? Example 6.1.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the *y*-intercept and what is the slope?

Answer

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is $25(a = 25)$. At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is $15(b = 15)$. For each session, Svetlana earns $15 for each hour she tutors.

This page titled 6.1: Simple Linear Equation Refresher is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **13.3: Linear Equations** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

6.2: Multiple Regression

Different Methods of Multiple Regression

Before we go into the statistical details of multiple regression, I want to first introduce three common methods of multiple regression: forced entry regression, hierarchical regression, and stepwise regression. The differences between these methods of multiple regression lies in how the variables are entered into a regression model. Let's look at each of them.

In forced entry regression, we choose independent variables, or predictors, based on theories and/or empirical literature to include in the regression model. Like the name suggests, we will force enter all chosen independent variables into the regression model simultaneously and study them altogether. In this case, all the predictors are treated equally. In other words, we don't have a preference or hold more interest in one predictor over the other predictors.

In hierarchical regression, just like in forced entry regression, we rely on previous theoretical and/or empirical evidence to select the independent variables to be included. But unlike forced entry regression, the predictors are not entered simultaneously. Instead, we, the researchers, determine in which order the predictors are entered. This is where the hierarchy comes in. It is essentially the order how the predictors are entered. Each step is considered a block, and one or multiple predictors can be included/entered in each block. Each block is considered a model. So hierarchical regression typically include multiple models. For example, if you enter two predictors (IV1, IV2) in block 1, and then enter another predictor (IV3) in block 2, then you will have two models, model 1 from block 1, and model 2 from block 2. In model 1, there are two predictors: IV1 and IV2. In model 2, there are three predictors (IV1 and IV2 from model 1, plus IV3 we just added). Keep in mind, each predictor can be entered only once. Once a predictor is entered into a block, it stays there for all the following blocks and you can't take the predictor(s) out once they are in. That's why IV1 and IV2 remain in model 2 above.

So how do we decide which order to enter the predictors? Generally speaking, in the first model, we would include participant background variables (e.g., self-esteem, gender, ethnicity, education levels, and so on). These predictors likely will influence the dependent variables even though they may not be the focus of our research study. In the next model (model 2), we would include any variables that are known predictors for the dependent variable(s). In the next model (model 3), we will add in new predictors we are particularly interested in. Often times, our goal is to determine if the newly added independent variables could better explain the dependent variable(s), in other words, whether the newly added independent variables could explain significantly more variance in the dependent variable above and beyond the other independent variables already included in the models.

Lastly, unlike the first two methods of regression, stepwise regression doesn't rely on theories or empirical literature at all. It is a purely mathematically based model. All you need to do is throwing in a bunch of IVs, and the software program will sift through all the IVs you entered to identify the ones that best predict the dependent variable(s) by selecting the predictor(s) that has the highest correlation with the dependent variable. It can be done using either forward method or backward method. Regardless, the decision is purely based on mathematical criterion, not on theories. If you know one thing about stepwise regression, that is to avoid it at all cost. As researchers, we want to make sure we choose the predictors based on theories and/or empirical literature.

Multiple Regression

Regression analysis is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further, regression analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. This last feature, of course, is all important in predicting future values.

Regression analysis is based upon a functional relationship among variables and further, assumes that the relationship is linear. This linearity assumption is required because, for the most part, the theoretical statistical properties of non-linear estimation are not well worked out yet by the mathematicians and statisticians. There are techniques for overcoming some of these difficulties, exponential and logarithmic transformation of the data for example, but at the outset we must recognize that standard ordinary least squares (OLS) regression analysis will always use a linear function to estimate what might be a nonlinear relationship.

The general linear regression model can be stated by the equation:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

where b_0 is the intercept, b_i's are the slope between Y and the appropriate X_i, and e, is the error term that captures errors in measurement of y and the effect on y of any variables missing from the equation that would contribute to explaining variations in y.

This model works only if certain assumptions hold. We'll look at this next.

Assumptions of the Ordinary Least Squares Regression Model

There are several assumptions of OLS regression. If one of these assumptions fails to be true, then it will have an effect on the quality of the estimates. Some of the failures of these assumptions can be fixed while others result in estimates that quite simply provide no insight into the questions the model is trying to answer or worse, give biased estimates.

1. The error term is a normally distributed with a mean of zero and a constant variance. The meaning of this is that the variances of the independent variables are independent of the value of the variable. Consider the relationship between personal income and the quantity of a good purchased, which is an example of a case where the variance is dependent upon the value of the independent variable, income. It is plausible that as income increases, the variation around the amount purchased will also increase simply because of the flexibility provided with higher levels of income. The assumption is for constant variance with respect to the magnitude of the independent variable called homoscedasticity. If the assumption fails, then it is called heteroscedasticity. Figure 6.2.1 shows the case of homoscedasticity where all three distributions have the same variance around the predicted value of Y regardless of the magnitude of X.
2. The independent variables are all from a probability distribution that is normally distributed. This can be seen in Figure 6.2.1 by the shape of the distributions placed on the predicted line at the expected value of the relevant value of Y.
3. The independent variables are independent of Y, but are also assumed to be independent of the other X variables, or other independent variables. The model is designed to estimate the effects of independent variables on some dependent variable in accordance with a proposed theory. The case where some or more of the independent variables are correlated is not unusual. There may be no cause and effect relationship among the independent variables, but nevertheless they move together. For example, you have two variables, household income and socio-economic status (SES), and they are theoretically related to each other. If you want to use both of them as predictors in one model, it would violate this assumption of regression analysis. This condition is called multicollinearity, which will be taken up in detail later.

Figure 6.2.1 does not show all the assumptions of the regression model, but it helps visualize these important ones.

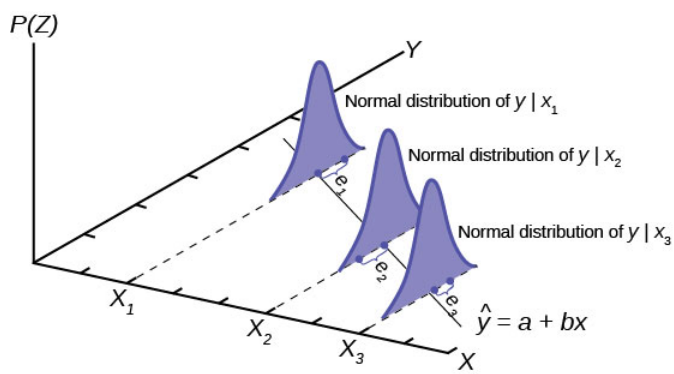
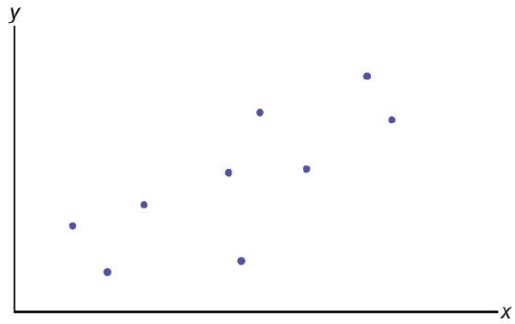


Figure 6.2.1



$$y = \beta_0 + \beta_1 X + \varepsilon$$

Figure 6.2.2

Going back to the general linear regression model stated earlier:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

This is the general form that is most often called the multiple regression model. So-called "simple" regression analysis has only one independent variable rather than many independent variables. Simple regression is just a special case of multiple regression. There is some value in beginning with simple regression: it is easy to graph in two dimensions, difficult to graph in three dimensions, and impossible to graph in more than three dimensions. Consequently, our graphs will be for the simple regression case. Figure 6.2.2 presents the regression problem in the form of a scatter plot graph of the data set where it is hypothesized that Y is dependent upon the single independent variable X.

Let's look at an example. The theoretical relationship states that as a person's income rises, their consumption rises, but by a smaller amount than the rise in income. If Y is consumption and X is income in the equation below Figure 6.2.3, the regression problem is, first, to establish that this relationship exists, and second, to determine the impact of a change in income on a person's consumption. Each "dot" in Figure 6.2.3 represents the consumption and income of different individuals at some point in time.

Regression analysis is often called "ordinary least squares" (OLS) analysis because the method of determining which line best "fits" the data is to minimize the sum of the squared residuals or errors of a line put through the data.

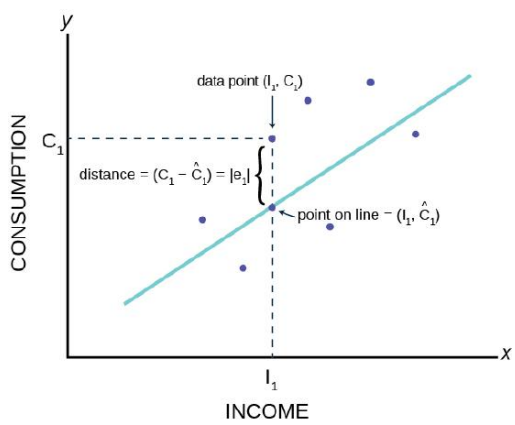


Figure 6.2.3

Estimated Equation: $C = b_0 + b_1 \text{Income} + e$

Figure 6.2.3 shows the assumed relationship between consumption and income based on the theory. Here the data are plotted as a scatter plot and an estimated straight line has been drawn. From this graph we can see an error term, e_1. Each data point also has an error term. Again, the error term is put into the equation to capture effects on consumption that are not caused by income changes. Such other effects might be a person's savings or wealth, or periods of unemployment. We will see how by minimizing the sum of these errors we can get an estimate for the slope and intercept of this line.

Consider the graph Figure 6.2.4 below. The notation has returned to that for the more general model rather than the specific example of the consumption and income.

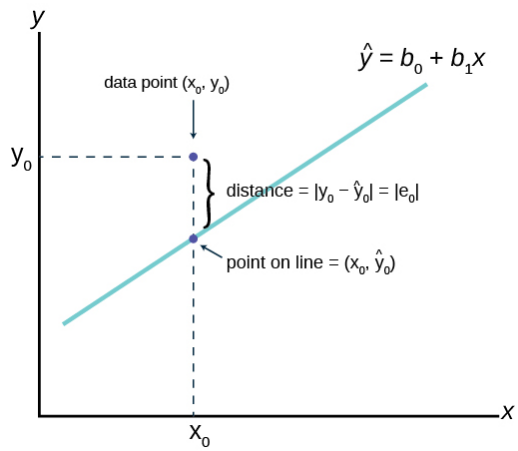


Figure 6.2.4

The $\hat{y}$ is read **"y hat"** and is the **estimated value of y**. (In Figure 6.2.3 $\hat{C}$ represents the estimated value of consumption because it is on the estimated line.) It is the value of y obtained using the regression line. $\hat{y}$ is not generally equal to y from the data.

The term $y_0 - \hat{y}_0 = e_0$ is called the **"error" or residual**. The error term was put into the estimating equation to capture missing variables and errors in measurement that may have occurred in the dependent variables. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line as can be seen on the graph at point X_0.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y.

If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the graph, $y_0 - \hat{y}_0 = e_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive. For each data point the residuals, or errors, are calculated $y_i - \hat{y}_i = e_i$ for $i = 1, 2, 3, ..., n$ where n is the sample size. Each $|e|$ is a vertical distance.

The sum of the errors squared is the term obviously called **Sum of Squared Errors (SS Error)**.

Using calculation, you can determine the straight line that has the parameter values of b_0 and b_1 that minimizes the **SS Error.** When you make the **SS Error** a minimum, you have determined the points that are on the line of best fit. We can further calculate the variance of the squared errors, e^2:

$$s_e^2 = \frac{\Sigma\left(y_i - \hat{y}_i\right)^2}{n-k} = \frac{\Sigma e_i^2}{n-k}$$

where $\hat{y}$ is the predicted value of y and y is the observed value, and thus the term $\left(y_i - \hat{y}_i\right)^2$ is the squared errors that are to be minimized to find the regression line. One important note is that here we are dividing by $(n-k)$, which is the degrees of freedom. The degrees of freedom of a regression equation will be the number of observations, n, reduced by the number of estimated parameters, which includes the intercept as a parameter.

The variance of the errors is fundamental in testing hypotheses for a regression. It tells us just how “tight” the dispersion is about the line. The greater the dispersion about the line, meaning the larger the variance of the errors, the less probable that the hypothesized independent variable will be found to have a significant effect on the dependent variable. In short, the theory being tested will more likely fail if the variance of the error term is high. Upon reflection this should not be a surprise. As we tested hypotheses about a mean we observed that large variances reduced the calculated test statistics and thus it failed to reach the tail of the distribution. In those cases, the null hypotheses could not be rejected. If we cannot reject the null hypothesis in a regression problem, we must conclude that the hypothesized independent variable has no effect on the dependent variable.

A way to visualize this concept is to draw two scatter plots of x and y data along a predetermined line. The first will have little variance of the errors, meaning that all the data points will move close to the line. Now do the same except the data points will have

a large estimate of the error variance, meaning that the data points are scattered widely along the line. Clearly the confidence about a relationship between x and y is affected by this difference between the error variances.

This page titled 6.2: Multiple Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **13.4: The Regression Equation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

6.3: Regression Coefficients

The whole goal of the regression analysis was to test the hypothesis that the dependent variable, Y, was in fact dependent upon the values of the independent variables as asserted by some theory, such as the consumption and income example. Looking at the estimated equation under Figure 13.8, we see that this amounts to determining the values of b_0 and b_1.

The regression analysis output provided by the computer software will produce an estimate of b_0 and b_1, and any other b's for other independent variables that were included in the estimated equation. The issue is how good are these estimates? In order to test a hypothesis concerning any estimate, we have found that we need to know the underlying sampling distribution. It should come as no surprise by now in the course that the answer is going to be the normal distribution. This can be seen by remembering the assumption that the error term in the population is normally distributed. If the error term is normally distributed and the variance of the estimates of the equation parameters, b_0 and b_1, are determined by the variance of the error term, it follows that the variances of the parameter estimates are also normally distributed. And indeed this is just the case.

To test whether or not Y does indeed depend upon X, or in our example, that whether consumption depends upon income, we need only test the hypothesis that b_1 equals zero. This hypothesis would be stated formally as:

$$H_0 : b_1 = 0$$

$$H_a : b_1 \neq 0$$

If we cannot reject the null hypothesis, we must conclude that our theory has no validity. If we cannot reject the null hypothesis that $b_1 = 0$ then b_1, the coefficient of Income, is zero and zero times anything is zero. Therefore the effect of Income on Consumption is zero. There is no relationship as our theory had suggested.

Notice that as before, we have set up the presumption, the null hypothesis, as "no relationship". This puts the burden of proof on the alternative hypothesis. In other words, if we are to validate our claim of finding a relationship, we must do so with a level of significance greater than 95 percent typically. No relationship exists, and to be able to make the claim that we have actually added to our body of knowledge we must do so with significant probability of being correct.

The test statistic for this test comes directly from our old friend, the t-test formula:

$$t_c = \frac{b_1 - \beta_1}{S_{b_1}}$$

where b_1 is the estimated value of the slope of the regression line, β_1 is the hypothesized value of slope of the regression line, which is always zero, and S_{b_1} is the standard deviation of the estimate of b_1. In this case we are asking how many standard deviations is the estimated slope away from the hypothesized slope. This is exactly the same question we asked before with respect to a hypothesis about a mean: how many standard deviations is the estimated mean, the sample mean, from the hypothesized population mean?

The decision rule for acceptance or rejection of the null hypothesis follows exactly the same form as in all our previous test of hypothesis. Namely, if the calculated value of t (or Z) falls into the tails of the distribution, where the tails are defined by α, the required significance level in the test, we have enough evidence to reject the null hypothesis. If on the other hand, the calculated value of the test statistic is within the critical region, we fail to reject the null hypothesis.

If we conclude that we reject the null hypothesis, we are able to state with $(1 - \alpha)$ level of confidence that the slope of the line is given by b_1. This is an extremely important conclusion. Regression analysis not only allows us to test if a cause and effect relationship exists, we can also determine the magnitude of that relationship, if one is found to exist. It is this feature of regression analysis that makes it so valuable. If models can be developed that have statistical validity, we are then able to simulate the effects of changes in variables that may be under our control with some degree of probability, of course. For example, if intentional advising is demonstrated to affect student retention, we can determine the effects of changing to intentional advising and decide if the increased retention are worth the added expense.

This page titled 6.3: Regression Coefficients is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **13.4: The Regression Equation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

6.4: Effect Size

In the last section we concerned ourselves with testing the hypothesis that the dependent variable did indeed depend upon the hypothesized independent variable or variables. It may be that we find an independent variable that has some effect on the dependent variable, but it may not be the only one, and it may not even be the most important one. Remember that the error term was placed in the model to capture the effects of any missing independent variables. It follows that the error term may be used to give a measure of the "goodness of fit" of the equation taken as a whole in explaining the variation of the dependent variable, Y.

The effect size is given by the formula:

$$R^2 = \frac{\text{SSReg}}{\text{SSTotal}}$$

where SS Reg (or SSR) is the regression sum of squares, the squared deviation of the predicted value of y from the mean value of $y(\hat{y} - \overline{y})$, and SS Total (or SST) is the total sum. Figure 6.4.1 shows how the total deviation of the dependent variable, y, is partitioned into these two pieces.

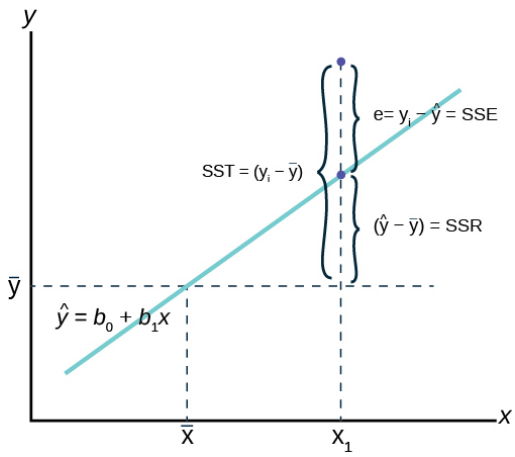


Figure 6.4.1

Figure 6.4.1 shows the estimated regression line and a single observation, x_1. Regression analysis tries to explain the variation of the data about the mean value of the dependent variable, y. The question is, why do the observations of y vary from the average level of y? The value of y at observation x_1 varies from the mean of y by the difference $(y_i - \overline{y})$. The sum of these differences squared is SST, the sum of squares total. The actual value of y at x_1 deviates from the estimated value, $\hat{y}$, by the difference between the estimated value and the actual value, $(y_i - \hat{y})$. We recall that this is the error term, e, and the sum of these errors is SSE, sum of squared errors. The deviation of the predicted value of y, $\hat{y}$, from the mean value of y is $(\hat{y} - \overline{y})$ and is the SS Reg, sum of squares regression. It is called “regression” because it is the deviation explained by the regression. (Sometimes the SS Reg is called SS Model for sum of squares model because it measures the change from using the mean value of the dependent variable to using the model, the line of best fit. In other words, it measures the deviation of the model from the mean value of the dependent variable, y, as shown in the graph.).

Because SS Total = SS Reg + SS Error, we see that the effect size is the percent of the variance, or deviation in y from its mean value, that is explained by the equation when taken as a whole. R^2 will vary between 0 and 1, with 0 indicating that none of the variance in y was explained by the equation and a value of 1 indicating that 100% of the variance in y was explained by the equation.

While a high R^2 is desirable, remember that it is the tests of the hypothesis concerning the existence of a relationship between a set of independent variables and a particular dependent variable that was the motivating factor in using the regression model. The goal of choosing the regression analysis is to validate a statistical relationship developed by some theory. Increasing the number of independent variables will have the effect of increasing R^2. But the goal is not to add as many independent variables as you possibly can; instead it is to select robust independent variables as informed by theories and/or empirical literature.

This page titled 6.4: Effect Size is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **13.4: The Regression Equation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

6.5: Multicollinearity

Multicollinearity

Our discussion earlier indicated that like all statistical models, the OLS (Ordinary Least Squares) regression model has important assumptions attached. Each assumption, if violated, has an effect on the ability of the model to provide useful and meaningful estimates. Here we will look at the effects on OLS estimates if the independent variables are correlated. We take up multicollinearity because it is so often prevalent in social sciences studies and it often leads to frustrating results.

The OLS model assumes that all the independent variables are independent of each other. This assumption is easy to test for a particular sample of data with simple correlation coefficients. Correlation, like much in statistics, is a matter of degree: a little is not good, and a lot is terrible.

The goal of the regression technique is to tease out the independent impacts of each of a set of independent variables on some hypothesized dependent variable. If two independent variables are interrelated, that is, correlated, then we cannot isolate the effects on Y of one from the other. In an extreme case where x_1 is a linear combination of x_2, correlation equal to one, both variables move in identical ways with Y. In this case it is impossible to determine the variable that is the true cause of the effect on Y.

The correlation has the same effect on the regression coefficients of both these two variables. In essence, each variable is “taking” part of the effect on Y that should be attributed to the collinear variable. This results in biased estimates.

Furthermore, multicollinearity often results in failing to reject the null hypothesis that the X variable has no impact on Y when in fact X does have a statistically significant impact on Y. Said another way, the large standard errors of the estimated coefficient created by multicollinearity suggest statistical insignificance even when the hypothesized relationship is strong.

This page titled 6.5: Multicollinearity is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Yang Lydia Yang via source content that was edited to the style and standards of the LibreTexts platform.

- **13.4: The Regression Equation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-business-statistics.

Index

A

anova
- 3.5: Partioning the Sum of Squares

B

block randomization
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

C

carryover effect
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

coefficient of determination
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

coefficient of multiple determination
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

cohort effect
- 1.6: Non-Experimental Research

complete counterbalancing
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

confounds
- 1.1: Research Designs

context effect
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

contrast effect
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

correlation
- 1.1: Research Designs

correlational research
- 1.6: Non-Experimental Research

counterbalancing
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

D

dependent variable
- 1.1: Research Designs

E

estimate of the error variance
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

Experimenter expectations
- 1.1: Research Designs

F

factorial design
- 3.1: Factorial Designs

factorial design table
- 3.1: Factorial Designs

fatigue effect
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

I

independent variable
- 1.1: Research Designs

interaction
- 3.2: Factorial ANOVA - Main Effects
- 3.3: Factorial ANOVA - Interaction Effects
- 3.4: Factorial ANOVA - Simple Effects

L

Longitudinal studies
- 1.6: Non-Experimental Research

Longitudinal study
- 1.1: Research Designs

M

main effect
- 3.2: Factorial ANOVA - Main Effects
- 3.3: Factorial ANOVA - Interaction Effects
- 3.4: Factorial ANOVA - Simple Effects

mixed factorial design
- 3.1: Factorial Designs

multiple correlation coefficient
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

N

nonequivalent groups design
- 1.5: Common Quasi-Experimental Designs

O

Observational research
- 1.6: Non-Experimental Research

operational definitions
- 1.1: Research Designs

order effect
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

P

Participant demand
- 1.1: Research Designs

placebo effect
- 1.1: Research Designs

posttest only nonequivalent groups design
- 1.5: Common Quasi-Experimental Designs

practice effect
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

R

random assignment
- 1.1: Research Designs
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

random counterbalancing
- 2.1: Between-Subjects Design
- 5.1: Within-Subjects Design

regression equation
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

S

Simple effects
- 3.2: Factorial ANOVA - Main Effects
- 3.3: Factorial ANOVA - Interaction Effects
- 3.4: Factorial ANOVA - Simple Effects

spreading interactions
- 3.2: Factorial ANOVA - Main Effects
- 3.3: Factorial ANOVA - Interaction Effects
- 3.4: Factorial ANOVA - Simple Effects

standard error of the estimate
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

sum of squared errors (SSE)
- 6.2: Multiple Regression
- 6.3: Regression Coefficients
- 6.4: Effect Size
- 6.5: Multicollinearity

switching replication with treatment removal design
- 1.5: Common Quasi-Experimental Designs

Detailed Licensing

Overview

Title: EDCEP 917: Experimental Design (Yang)

Webpages: 43

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 97.7% (42 pages)
- Undeclared: 2.3% (1 page)

By Page

- EDCEP 917: Experimental Design (Yang) - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *CC BY-NC-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY-NC-SA 4.0*
  - 1: Introduction to Research Designs - *CC BY-NC-SA 4.0*
    - 1.1: Research Designs - *CC BY-NC-SA 4.0*
    - 1.2: Internal and External Validity - *CC BY-NC-SA 4.0*
    - 1.3: Threats to Internal Validity and Different Control Techniques - *CC BY-NC-SA 4.0*
    - 1.4: Common Experimental Designs - *CC BY-NC-SA 4.0*
    - 1.5: Common Quasi-Experimental Designs - *CC BY-NC-SA 4.0*
    - 1.6: Non-Experimental Research - *CC BY-NC-SA 4.0*
  - 2: Between-Subjects Single Factor Design - *CC BY-NC-SA 4.0*
    - 2.1: Between-Subjects Design - *CC BY-NC-SA 4.0*
    - 2.2: One-Way ANOVA - *CC BY-NC-SA 4.0*
    - 2.3: Computing the F statistics - *CC BY-NC-SA 4.0*
    - 2.4 Post Hoc Tests - *CC BY-NC-SA 4.0*
  - 3: Between-Subjects Factorial Design - *CC BY-NC-SA 4.0*
    - 3.1: Factorial Designs - *CC BY-NC-SA 4.0*
    - 3.2: Factorial ANOVA - Main Effects - *CC BY-NC-SA 4.0*
    - 3.3: Factorial ANOVA - Interaction Effects - *CC BY-NC-SA 4.0*
    - 3.4: Factorial ANOVA - Simple Effects - *CC BY-NC-SA 4.0*
    - 3.5: Partioning the Sum of Squares - *CC BY-NC-SA 4.0*
  - 4: Between-Subjects Design with a Control Variable - *CC BY-NC-SA 4.0*
    - 4.1: Overview of the Control Variable - *CC BY-NC-SA 4.0*
    - 4.2: Radomized Block Design - *CC BY-NC-SA 4.0*
    - 4.4: Analysis of Covariance (ANCOVA) - *CC BY-NC-SA 4.0*
  - 5: Within-Subjects Design - *CC BY-NC-SA 4.0*
    - 5.1: Within-Subjects Design - *CC BY-NC-SA 4.0*
    - 5.2: Partioning the Sums of Squares - *CC BY-NC-SA 4.0*
    - 5.3: Repeated Measures ANOVA - *CC BY-NC-SA 4.0*
    - 5.4: More Notes on Repeated Measures ANOVA - *CC BY-NC-SA 4.0*
  - 6: Multiple Linear Regression - *CC BY-NC-SA 4.0*
    - 6.1: Simple Linear Equation Refresher - *CC BY-NC-SA 4.0*
    - 6.2: Multiple Regression - *CC BY-NC-SA 4.0*
    - 6.3: Regression Coefficients - *CC BY-NC-SA 4.0*
    - 6.4: Effect Size - *CC BY-NC-SA 4.0*
    - 6.5: Multicollinearity - *CC BY-NC-SA 4.0*
  - Back Matter - *CC BY-NC-SA 4.0*
    - Index - *CC BY-NC-SA 4.0*
    - Glossary - *CC BY-NC-SA 4.0*
    - Detailed Licensing - *CC BY-NC-SA 4.0*