Quantitative Research By Example

Version 1.0.0 {β}

Ioan Gelu Ionas
Ph.D, MBA

My Research Lab
MRL BETA

IOAN GELU IONAS, MBA, PH.D.

QUANTITATIVE RESEARCH BY EXAMPLE

$VERSION: \ 1.0.0 \, \{\beta\}$

Contents

Ioan Gelu Ionas, MBA, Ph.D.

not-not.net[1]

[1] http://www.not-not.net

[2] http://creativecommons.org/licenses/by-nc-sa/4.0/

Foreword

Today's world produces ever increasing amounts of data which, when used properly, can provide insights that were not possible before. While these insights are helpful, the process that leads to them, which often involves statistical analyses, can look so intimidating so much so that many will not even try to use this treasure trove of data available at their fingertips. Moreover, it can be even dangerous to use insights that were improperly derived from the data. I, for one, believe that quantitative analysis and statistics should be approachable, for which reason I have began developing this book, available for free to everyone, and the associated resources. I built these resources from the point of view that while deep understanding of how a statistical test works and what it does requires deep understanding of mathematics and probabilities, its use to understand a quantifiable phenomena does not have to be overwhelming.

This resource, comprised of a written book and an accompanying website and app[3], is designed as an accessible source of information and support to everyone interested in designing and conducting quantitative studies. It tries to hide the math as much as possible and focus on the application of statistical tests through real life examples, whenever available. Nevertheless, the book is not intended for absolute beginners as introductory knowledge of statistics and probabilities that helps readers make the most out of the content is not included. The material presented assumes the readers have at least a basic understanding of basic statistics principles.

[3] App will be available at a later time.

The the book and its associated ecosystem is intended from start to be under perpetual improvement through new and updated content, functionality, and tools. For this reason the My Research Lab website[4] will always provide the most up to date content, both online and in printable format.

[4] https://myrelab.com

The book covers a few concepts fundamental to the understanding of quantitative methods, offers guidance for the design of the study, and offers guided and annotated examples for statistical analysis tests. For

each of the analyses presented the focus is on application rather than on theory. These examples are not definitive and complete guides but rather a use case as the analysis is applied to a specific study. Because each study has its own specific characteristics, readers are strongly encouraged to consult as many resources specific to their application and situation as necessary to make informed decisions. The website offers access to a set of tools[5] to help design quantitative research studies, as well as access to the R code that powers the statistical analyses presented in the book and the associated curated datasets. In the future, as time and resources allow, the website will provide access to interactive Jupiter Notebooks[6] or similar tools in both R and Python for each of the example statistical analyses. Additional, more personalized, direct support may be available in the future. In the mean time, just drop me a line.

[5] Such as sample size calculators, helper tools for selecting the appropriate sample size, and so forth.

[6] https://jupyter.org/

It is my hope that this resource will grow with the help of other researchers willing to share their work by converting their quantitative studies into guided examples or case studies that can be then added to the book. Therefore, I invite everyone interested in sharing their research to help others to contact me. I will serve as guide and editor through the process and the example will be published under the original author's name.

The resource is under development and will be a work in progress for the foreseeable future. It is not intended to exhaustive, but rather serve as a guide to applied quantitive research.

The book is developed using the R software environment for statistical computing[7] augmented by additional packages and software applications. The RMarkdown[8] and Bookdown[9] packages are used to weave the text with the R code. This setup makes possible to update at the same time all written content, in all formats, from a single source of truth.

[7] https://www.r-project.org

[8] https://rmarkdown.rstudio.com/

[9] https://bookdown.org/

A big thank you to those who helped develop this resource this far. Special thanks to Dr. Mugur Geana and Dr. Dan Cernusca for their continued support.

A Few Words About Causality

Causality is pervasive and ubiquitous. It helps build intellectual understanding, supports deliberations, is involved in planning, in technology, and even in language[10]. It is part of everyday decisions, requiring implications and consequences to be considered in the process of evaluation and judgment. Causality is also a construct of intelligence. In Wesley C Salmon's words (Salmon 1998):

> "If there had never been any human or other intelligent beings, there never would have been causes and effects - that is to say, there never would have been causal relations - in the physical universe the events would occur, but the causal relations would not exist." (p. 8)

[10] Words like *break* or *move* can serve as indicators of causal relationships or events.

The understanding of causality is critical to the understanding of scientific research and everything that comes with it: research design, analysis, interpretations. This is because the drive to understand causality is deeply rooted in people's need to make sense of the world around them, or to come to terms with it.

More commonly understood, causality represents the relationship between two or more entities where the behavior of one or more of them determines the behavior of the other(s). Fundamentally, causality arises from the empirical relations of:

- *Contiguity*: in space and time;
- *Temporal succession*: temporal order determines causal priority and requires the cause to be readily present for the effect to occur. It is useful in determining which of the two variables that covary is the cause (the Independent Variable, or IV), and which is the effect (the Dependent Variable, or DV);
- *Constant conjunction*: covariation between two variables, which happen when two *objects, constructs, processes, etc.* are in constant (***repeat many times***) conjunction with each other. That is, for a *causal relation* to exist, they have to constantly, repeatedly, show the same behavior.

Conjunctive plurality, which states that an effect is rarely the result of a single cause, was later added as another important attribute to emphasize the complexity of real-life processes which are the target of many research studies.

Another way of looking at the fundamental conditions for a causal relation to exist is that causal relations are *durable*. They always (or most of the time) hold true. That is, they are *stable*, *consistent*, and *reliable* and remain true across time and space and across instances of the same system (Sloman 2005). Considering that our lives as intelligent beings are predicated on the ability to predict, we have to be *selective*[11] and *attend to what is stable*[12]. Finding these durable relations is, in essence, the purpose of scientific inquiry.

When reasoning about causality it may be helpful to think of it as two intersecting processes in space and time (Salmon 1998). The intersection, while not a causal construct or concept in itself, can help distinguish causal phenomena from non-causal ones. At that intersection we can expect two major types of events: *causal interactions* and *non-causal intersections*. If both processes exit the intersection in a changed state and that changed state persists beyond the place/time of intersection we have a *causal interaction*[13]. Alternatively, if either or none of the processes is changed, we are looking at a *non-causal intersection*[14].

While understanding causality is fundamental to both scientific and everyday reasoning there are many problems people have when dealing with causality and causally linked events. It is important for researchers to be aware that these issues and biases exist and to understand and make efforts to minimize their impact on a research study. For example, a few of these problems, relevant to designing research studies, are:

- People have a tendency to favor obvious, localized, simple, linear, and sequential causal relations;
- People tend to simplify otherwise more complex causal structures, a process which results in distorted understanding;
- For well-structured problems (for which both states and constraints are clearly defined) people tend to use strategies that convert the problem's elements into computations while, many times, missing the conceptual underpinnings of the problem and the domain. We should note though that real-life phenomena, in the social sciences are rarely well structured[15];
- When faced with discordant information people tend to hold onto their *old/existing schemas* (their existing understanding) rather than update them or build new ones. That is, people tend to show *resistance to change*.

[11] Most facts and information are useless for a specific decision. Taking everything possible into consideration overloads our minds and slows us down. Therefore, selectivity means that only those things that carry relevant information for the task or decision are to be considered to make the process fast and efficient.

[12] Stability, or invariance, is fundamental to the process of prediction because it highlights the variables that behave constantly across time and space and therefore make it easy to infer their future values or behavior from their current values or state.

[13] For example, when a bullet hits a target, both the bullet and the target are affected at the point of intersection and in the future. The target ends up with a whole and the bullet loses energy.

[14] Think of two beams of light that cross paths and intersect at a point. At that point they are just superimposed on each other, but neither affects the other. After they leave that point, they continue unaffected.

[15] Well structured problems can be easily converted into procedures that can be used without much thought and without understanding the underlying principles.

Understanding the concept of causality is valuable to the design of a research study, experimental or not. For example, for causality to exist, the cause and effect have to be observed (measured) many times (2 or 3 is not enough), proving that constant conjunction exists and therefore confirming the potential for the existence of a causal relationship. This is what, fundamentally, statistical analysis tries to do. Therefore, the study should be designed so that that is possible. The other attributes are helpful in guiding the design and interpretation. For example, *temporal succession* should guide not only the determination of which is the *independent* and which is the *dependent* variable, but also in the selection of the variables appropriate for the study. *Contiguity* is also helpful in the selection and definition of variables and can be used as a checkpoint in building the model to be tested. *Conjunctive plurality* helps by raising awareness about the fact that many events can participate in the onset of the effect and is useful in the selection and definition of the variables used to define the construct being studied.

For a causal relationship to be truly understood, one needs to answer three questions: *what?*, *why?*, *how?*. The *what?* questions is answered through statistical analysis and proves that constant conjunction happens while the other requirements for a causal relation to exist are met. The *how?* and *why?* questions provide an explanation of the mechanisms at work that underlie the causal relations as well as the reasoning for the purpose of the relationship and what it means for theory and practice.

Discovery, understanding, and documentation of causal relationships is, most of the time, the focus of most research studies that use quantitative approaches to understand concepts, constructs, and phenomena. This chapter just scratched the surface of a much larger conversation on causality as it relates to the scientific method reflected in quantitative research designs. Causality is both the starting point and the end goal of many statistical analysis methods and can be observed at work through the theory and practice of statistics. It is construct worth exploring and understanding and a conversation worth having as it will lead to better research designs, with higher quality findings. Do not stop here.

Variance and Error

Variability is an essential characteristic of the natural world. We see it everywhere around us. For example, an extreme case is the human fingerprints, unique to each individual, which makes them useful for identification purposes. Similarly, maybe not as extreme and dependent on what is being studied, research participants are different from each other, differences which introduce variability in the study[16]. Due to this variability, the values obtained from measuring a construct differ from research participant to research participant. In classical statistical inference the *variance* is a measure of how spread out these readings are from the average of the sample.

The *Variance* is related to *Standard Deviation (SD)*[17] which is an indication of how much variation of or dispersion is in the values of a sample. A large SD value indicates that the values are spread out from the mean over a wider range of values while a small SD value indicates that the values are close together around the mean.

Total variance can be thought of as the sum of two variances: *systematic (between-groups)*[18] variance and *error (within-group)* variance. The ratio of the two variances can serve as an indication if the differences between groups are systematic or due to chance.

Systematic (between-groups) variance is the result of the intervention and any additional confounding variables present in the study. It is the intent of an experiment to generate variability in the dependent variable (DV) by manipulating the independent variable (IV). This is the type of variance research is looking for.

Error (within-groups) or non-systematic variance is the unexplained variability in the DV. It is usually more of a nuisance and it can be lived with. It is determined by the random variability between subjects.

Considering that reality is usually described by more than two variables, there are other variables that can affect systematic variance as well. Of these, the variables that influence both the IV and the DV are called *confounding variables*. *Confounding* happens when the design

[16] For this reason in laboratory experiments researchers attempt to control as many of the factors that produce variability as possible. On the opposite side are studies that embrace variability and attempt to collect data in the participant's natural environment.

[17] In mathematical terms, the *Standard Deviation* is the square root of *Variance*.

[18] Do not confuse between-group/within-group variance with the between-subjects/within-subjects research designs. See Between-Subjects vs. Within-Subjects Designs for more information.

of the experiment (controls) makes difficult or impossible to eliminate alternative explanations for an observed cause-effect relationship[19]. In many situations confounding variables are variables that the experiment did not account for. They can cause two major issues: increase variance and introduce bias.

The variability generated by the confounding variables is impossible to separate from the variability due to the intervention, which makes the interpretation of the results difficult or impossible. Therefore, any experimental design should attempt to eliminate any confounding variables and attempt to produce an as small error variance as possible.

The most effective way to control confounding variables is to use random assignment of participants to the experimental groups, which forces all variables other than those studied to create only random (no-systematic) variance. Random assignment has the effect of transferring the variance due to confounding to error variance[20].

On the other hand, the smaller the error variance, the more powerful the design. Therefore, in addition to eliminating any confounding variables, it is recommended to try to reduce as much as possible the variance due to error. Here are a couple of accepted ways of reducing the error variance:

- If possible, hold constant some of the variables instead of randomizing all variables in the study.
- Increase the size of the sample as error variance is inversely proportional to the number of degrees of freedom[21] of the sample.

The concept of *variance* is closely related to that of *error*. The following are the most significant sources of error in a quantitative research design:

- *Random error* - occurs by chance and can be produced by anything that randomly interferes with measurement;
- *Systematic error* - is generated by consistent differences between the measured value and the true value[22];
- *Measurement error* - denoting the *validity* and *reliability* of an instrument:
  - *Validity* - the instrument is capable to accurately measure the construct it was designed to measure;
  - *Reliability* - or reproducibility, is the capacity of the instrument to perform consistently over time and across observers[23].

[19] For example you run an experiment that includes the same number of men and women. The treatment is not relevant; what is relevant is how the treatment is applied to the participants. Consider two groups, a treatment group and a control group. If only men are assigned to the control group and only women to the treatment group, when it comes the time to interpret the results, there is no way to know if the observed effects are accurate and are only due to the treatment. In this case, participant's gender is confounding with the treatment, preventing the researcher to determine if the effects are due to the treatment only or the participant gender has also something to do with it.

[20] The rationale for transferring variance from the confounding variable(s) to error variance is that, in most cases, it is better to try to deal with error variance than bias in interpreting the results.

[21] Degrees of freedom is an estimate of the number of independent pieces of information that go into computing the estimate. It is the number of values that are free to vary in a data set. For one dataset it is calculated as the number of items in the set minus 1 (n-1). For two samples the degrees of freedom are computed considering that there are two n values, one for each sample, to consider. In this case the number of degrees of freedom is computed as df = n1 + n2 - 2.

[22] For example the measurement with an instrument that has a calibration issue or with a watch that is consistently ahead two minutes.

[23] That means that when the same instrument is used to measure something twice, the result of the measurement would be approximately the same.

- *Sampling error* - different samples generate different results, a fact which needs to be accounted for when making inferences from sample to population. This is measured by the standard error and it may result in:
  - *Type I Error* - occurs when the null hypothesis is rejected when it is true. The probability of occurrence of this error is called *significance level* and is denoted by the Greek letter alpha (α);
  - *Type II Error* - occurs when a false null hypothesis is accepted. The probability of this error not occurring is called *power* and is denoted by the Greek letter (β).

Sampling error cannot be completely eliminated but it can be reduced by increasing the sample size.

Design

To understand the world around us we use models we construct in our mind. Rooted in our education and our experience, these models may or may not be true representations of reality and how close that representation is. While for most people these models are "good enough" for their everyday needs, researchers take them one step further and attempt to devise ways to determine and understand if they are indeed true representations of reality. For this purpose, researchers *formalize* these *models* and *design* studies to test their validity. These formal models are based on the shared understanding of the phenomena under study at the time of the research[24] and the researcher's own insights. That is, scientific models are a representation of our evolving understanding of the physical world.

[24] For which reason a thorough literature review is necessary to building the most relevant model.

A research study is *designed* with the intent of finding or supporting a *model* representing a phenomenon or construct as defined by the relations between the various variables involved. The theoretical framework underlying this model is a major determinant for the choice of analytic technique, for how this technique is applied, and for how the results are interpreted. That is, the potential for success in a quantitative research study is determined, among other things, by the use of analytic techniques appropriate for the model involved.

When designing a research study one should be always aware that *all a statistically significant finding means is that the probability that there is nothing to be found*[25] *is small*. Therefore, a sound research design is based on a chain of decisions about the effect size that makes the relationships sought substantially meaningful for the study, the level of significance and the power of the statistical test, and a calculation of sample size. This approach helps avoid pitfalls such as findings that are meaningful but not statistically significant or findings that are statistically significant but not meaningful. That is, the focus of the research process should be on the *meaning*[26] of the findings from the perspective of the theory and existing research.

[25] The null hypothesis is true.

[26] This meaning, such as relations between variables or differences between means, cannot be established in the absence of other research.

Occam's Razor

Also known as the *Principle of Parsimony*, it states that simple explanations are better than more complicated ones. In statistics for example, using Occam's Razor means that an explanatory model with fewer variables is better than one with a larger number of variables.

Attributed to the English philosopher William of Ockham, the *Principle of Parsimony* states: *Given a set of equally valid equivalent explanatory models, the best explanation is the simplest one.* In statistical terms, it means:

- The model should have as few variables as possible.
- Linear models should be preferred to non-linear models.
- Models that rely on on fewer assumptions should be preferred to those that rely on many.
- Simpler explanations should be preferred to more complicated ones.

Used in the process of simplification of statistical models, the *Principle of Parsimony* advises that when a variable does not guarantee a significant increase in deviance[27] when removed, it should be excluded from the study.

OK, but what does this mean for a study? In essence, the *Principle of Parsimony* advises you to keep your research model as simple as possible. Many researchers, myself included, have at times a tendency to over complicate their studies[28], effectively overlooking this principle.

If you ever worked in one of the social sciences fields you probably know how difficult it can be to access and recruit participants for a study. Therefore, when opportunity arises, one might tend to include as many data collection items as feasibly possible. While having a lot of data is not a bad situation to be in, it also makes it tempting to fit it all in the model. This can also happen when a research tries to justify collecting more data just because access to participants is readily available.

Don't do it. That is not to say that you shouldn't collect the data if possible. By all means, if the opportunity presents itself, do so. Nevertheless, do not use more data than you actually need in your analysis. That is, let the *Principle of Parsimony* guide your decisions and ask yourself the following questions:

- Are all variables (data) collected valuable and relevant to the study?
- From a theoretical perspective, would the inclusion of a variable make a difference?

[27] *Deviance* is a statistic used to compare models. In this case, if the two models - with and without the variable in question - are not significantly different, the model without the variable, the simpler model, should be selected.

[28] By including as many variables as can be fit and collected, for examples.

The answer may not be obvious at first. To find out which model is more parsimonious, you should first start looking at the existing literature. Attempt to find similar studies and study the models they proposed as well as the data they used. You can then analyze each piece of data (variable) individually to assess if it, theoretically, adds value to your study, from both the perspective of prior research and based on your own understanding of the problem under study.

When you decide how complex to make your model and how much data you need to collect to validate it, think of the downsides increased complexity brings about. Here are a few:

- Increased difficulty in analyzing and interpreting data.
- Longer and more complicated research instruments[29].
- The time you will need to complete the study can also increase significantly[30].

[29] This could lead to a significant increase in the time the participants need to complete the task which, in turn, could significantly increase the chance for more participants to either decide to leave the study early or not to participate in it at all. That is, the longer an instrument is, the less likely is for the participants to complete it.

[30] If, for example, you are working towards a Doctoral Dissertation, increased complexity could bring delays in completion and graduation.

Experimental Versus Non-Experimental Research

A cursory search will show the wide variety of quantitative designes researchers use. Nevertheless, all of them can be included in one of two types: *experimental* or *non-experimental* studies. The most significant difference between the two types is the level of control researchers have over the environment in which the research is conducted.

Experimental research studies are designed to provide the researchers with the highest level of control possible over the experimental conditions. The intent of an experiment is to discover the relationships between the variables of interest while attempting to hold all other variables constant (or control them)[31]. For this purpose the experimenter usually manipulates a condition (the treatment) and attempts to assess its impact on one or more variables of interest. Because of the ability to manipulate the experimental conditions and the restrictive design aimed at controlling as much as possible other extraneous variables, experimental research offers the best chance of finding causal relationships between variables. Nevertheless, the controlled nature of the studies, makes them less capable of reflecting reality. Given the restrictive nature of the design, experiments offer a high level of reliability and control[32].

[31] For example, *random assignment* is such a way to control for differences between subjects in research involving human subjects.

[32] In STEM fields, studies conducted in the laboratory employ experimental designs. In non STEM fields experimental design is used, for example, to understand the differences between groups of participants (people) subjected to different experimental conditions, such as different visual stimuli, or different learning environments.

Non-Experimental research studies are designed to look at phenomena and contexts the researcher does not have control over. In this case the researchers cannot manipulate the conditions or variables of interest and they have to rely on observations and measurements of variables available to them and use those to seek an answer to the research question they pose. This lack of control renders non-experimental studies

less capable of identifying causal relationships due to the large number of variables that usually accompany real-life contexts. Therefore, while a relationship may be observed and inferences can be made, the researcher's ability to strongly suggest causality is limited by the potential interference in the process of other variables that were not accounted for[33]. Because non-experimental studies look at phenomena in their natural environment, they tend to have a higher level of external validity, which makes them much easier to generalize to larger populations[34].

Going a bit further, quantitative research studies can be grouped into descriptive, correlational, quasi-experimental, and experimental.

Descriptive studies are designed to describe the status of a phenomenon using mostly observational type data. They do not have hypotheses, though one may be developed after the data is examined.

Correlational studies use mostly observational data to explore relationships between variables without looking at cause-effect relationships.

Quasi-experimental studies are designed to recognize cause-effect relationships between variables in situations when no groups are assigned beforehand and no variables are manipulated to elicit a desired outcome. The groups for which variable statistical summary data are compared are identified after the data has been collected.

Experimental studies follow the guidelines of the scientific method and are specifically designed to verify the existence of a cause-effect relationship between variables describing a phenomenon. For this purpose all efforts should be made to control for as many variables as possible while manipulating the variable(s) of interest.

[33] The context in a non-experimental social sciences research study is so complex that the researchers cannot capture and measure every variable that may influence the phenomena they are studying. For example, in educational studies, prior knowledge has significant influence on how well a learner understands new concepts. In real life situations, such as during classroom instruction, there may not be time or capabilities to assess the learners' prior knowledge of a concept or construct.

[34] These types of studies are frequently encountered in the social sciences fields, where researchers attempt to study phenomena as they unfold in their normal environments.

Between-Subjects vs. Within-Subjects Designs

The *between-subjects* and *within-subjects* research designs are differentiated by the number of measurements done for every subject. In *between-subjects* designs only one measurement is performed for each participant while in *within-subjects* designs, there are multiple, successive, measurements, for which reason the *within-subjects* studies are many times called *repeated measures* studies[35].

In essence, the *between-subjects* research design allows researchers to study the differences between groups of participants at a given point in time. They usually involve comparing the groups on one or more summary or central tendency measures[36]. The participants are part of only one of the research groups and are exposed to only one intervention.

In *within-subjects* designs all participants are members of the same group and all are exposed to all treatments. The comparison usually

[35] *Repeated measures* studies are only a subgroup, probably the largest, of the broader category of *within-subjects* designs.

[36] E.g., mean or median.

happens between the successive values of central tendency measures of the same variable. These types of designs tend to have more power than the *between-subjects* designs and make possible to observe change over time, but tend to suffer from confounding issues[37].

[37] Confounding in *within-subjects* designs can be mitigated by counterbalancing. For example, participants can be grouped together in small groups and the order in which they are subjected to the various treatment conditions can be randomized across these groups. Or, the randomization of how the treatments are applied can be done for each individual subject.

Variables

Variable: A specific characteristic that can be measured and can assume different values.

Continuous or Quantitative variables: a variable that has numerical values, such as test scores, lengths, durations, etc.

Classification or Categorical variables: represent categories, usually used as grouping variables, such as gender or race.

Measurement Scales

One of the major ways of understanding variables is to look at how they are measured and their scale of measurement. From this perspective variables can be *nominal*, *ordinal*, *interval*, or *ratio*. This classification is important - because the statistical procedure to be used depends on the scale of measurement.

Nominal Scale: Classifies in mutually exclusive categories. The variable becomes a classification variable.

Ordinal Scale[38]: Rank order with respect to the variable being assessed. The values represent a hierarchy of levels. Provides limited information because the equal steps in the scale values do not necessarily have an equal real-life quantitative meaning.

[38] As an example, letter grades are ordinal because how much A is better than a B cannot be known. For a score range between 0 and 100 A is between 90 and 100 and B between 80 and 89.9. The difference between an A and a B is anything between 19.9 and 0.1. An F is for any score below 50, different from the others. Recoding to a scale of 1 to 5 is misleading because the numerical difference between 1 and 2 is not the same as the difference between F and D.

Interval Scale: Provides more information than the ordinal scale because equal differences between values have a real-life equal meaning/counterpart. The downside of the interval scale is that is has no true zero point[39].

[39] A value of zero on the scale is equal to a zero quantity of the variable being assessed. For example, the Celsius scale does not have a true zero point because the value of 0 does not mean that there is absolutely no heat present.

Ratio Scales: Are similar to interval scales in that equal differences between scale values have equal real-life quantitative meaning. However, ratio scales also have a true zero point which gives them an additional property. With ratio scales, it is possible to make meaningful statements about the ratios between scale values[40].

[40] For example, the system of inches used with a common ruler is an example of a ratio scale. There is a true zero point with this system in which zero inches does, in fact, indicate a complete absence of length. The Kelvin scale is, as opposed to the Celsius and Fahrenheit scales, a ratio scale because 0 degrees Kelvin means, by design, that there is absolutely no heat present.

Independent vs. Dependent Variables

Many research studies are primarily focused on two categories of variables: *dependent* and *independent.* This categorization is based on what

the variable measures and how it is intended to be used in the analysis. To understand the difference, let's look at an experimental study design.

An experimental study can be designed to find, for example, if two interventions or treatments offer different or similar outcomes. This implies the study needs two groups of participants, with one of the groups being subject to one of the treatments, while the second group is subject to the other. If the groups come from the same population and are homogeneous enough, the experiment should be able to recognize the effects of the treatments. To recognize the effects, the design should also include ways to measure them, such as test grades or scores.

In terms of variables, the one measuring the effects or outcomes is the *dependent variable (DV)*, while the one that places the participants in groups is the *independent variable (IV)*. The study will attempt to determine the influence of the *independent variable* (treatment group membership) on the *dependent variable*.

The *dependent variables* are those that measure an observed effect. Examples of dependent variable could be a test score used as proxy for students performance on a task.

The *independent variables* reflect either conditions specified by design to help single out the effect or determined by conditions that are outside the researcher's and, potentially, the participant's control. Examples of such variables are grouping variables (e.g., treatment vs. control groups) or demographic variables (e.g., age, gender, etc.).

With this knowledge lets look at an example. Consider we have a study in which the participants are assigned to one of two groups. This is under the researcher's control. More exactly, these groups have been formed by design to represent two different conditions or interventions. In the analysis, they will be represented by the independent variable. On the other side, the researcher has devised a way to measure the effects of each of the two conditions or interventions (e.g., test scores). As, by design, the researcher expects these test scores to be, on average, different for the two groups, one could say that they "depend" on which group the participants were assigned to. In the analysis, this measure translates into the *dependent variable*. All else equal the effects (measured by the dependent variable) depend on the group to which the participants were assigned to (represented as the independent variable). Let's consider a more concrete example.

Let's say that we are studying the influence off the skill and drill practice on mathematics performance in high school students. For this purpose we select two groups of students. One group of students will do math as usual and will not engage in any skill and drill practice. The other group will continue to do math as usual but, in addition,

will have a few extra sessions during which they will do math drills. In this case, the first group is the *control group* and the second group is the *treatment group*. The variable which defines the group a student is member of, is the *independent variable*. All this said and done, the next step is to find a way to assess the students' performance. Let's consider that we chose a specific math test to be administered to all students after the treatment has ended[41]. The score the student obtains for this test could be considered a measure of their performance. Therefore, for further analysis, this score is considered to be the *dependent variable*. That is, the score the student obtains depends on which group he was member of and therefore dependent on the type of training (treatment) the students engaged in.

[41] Our expectation might be, for example, for the students in the skill and drill group to perform better than the students that did not do any math drills on this test.

The variables in a research design, while serving the same purpose and representing the same components of relationship, may be found under different names. For example, in experimental research studies the use of *independent*[42] / *dependent*[43] variables is preferred. For non-experimental studies, the preference is for using *predictor*[44] / *criterion/response*[45].

[42] Presumed cause in an experimental study.

[43] Studied effect in experimental studies.

[44] Presumed cause in an non-experimental study.

[45] Studied effect in a non-experimental study.

Descriptive vs. Inferential Designs

Descriptive statistical analysis is focused on measuring population characteristics. For the purpose of these analyses a *population* is defined as the entire collection of subjects or things that are being studied[46].

[46] For example, all the students in a course.

Inferential analysis is a *statistic*, a numerical value calculated using a sample (or subset) of people, objects, events, etc. that can be used to describe the characteristics of the sample[47] and/or used to make inferences/estimates about the population from which the sample was extracted.

[47] For example, the mean of some value.

Most statistical tests included in this resource will probably fall in the inferential category. Overall, inferential tests can be categorized in two basic types: *tests of group differences* and *tests of association*.

Test of group differences - are designed to help determine if there are differences between the mean scores of one or more dependent variables[48] between two populations. One of the best known examples is the *one-tailed t-test*.

[48] Or criterion variables in non-experimental studies.

Tests of association - for a single population, to determine if there is a relationship between two or more variables that describe this population. Best known example is the *correlation coefficient*.

A third, more involved, class of inferential analyses allows to study if the association between two variables is the same across two or more populations. An example of such type of analysis is ANCOVA.

Research Questions

In quantitative studies research questions ask, in essence, if a relationship exists between two events. In most cases this relationship is causal in nature. That is, a research question asks if the onset of an event has an impact on some other event. Let's use the well known butterfly effect[49] as an example.

[49] Term used in chaos theory, coined by Edward Lorenz.

> *Would the flap of the wings of a butterfly in the Amazonian jungle influence the number of hurricanes in Japan?*

A closer look shows that this is indeed a question. The first part introduces the originating event or the cause (*butterfly flaps wings*). The second part describes the effect (*number of hurricanes in Japan*). Further analysis shows that this question only asks if a relationship exists, but does not include any indication of how strong the relationship is and in which direction the effect will be. This is called a *non-directional* research question.

If the literature supports or suggests a direction for the causal relation, then a *directional* research question would be more appropriate as it includes an indication of how the relation is thought to behave. Let's transform the question above to a directional research question.

> *Would the flap of the wings of a butterfly in the Amazonian jungle significantly increase the number of hurricanes in Japan?*

This time the question suggests both the direction and the strength of the relationship, which is achieved by replacing the word *influence* with the words *significantly increase*. In this case, *significantly* is an indication of the strength of the relation and *increase* is an indication oft the direction[50].

[50] If the existing literature cannot provide any guidance as to what the strength of the relation may be, removing *significantly* will not affect the type of question.

Hypotheses

Hypotheses are questions worded as statements to be tested using statistical tests. They are derived from the study's research questions and describe the causal relation(s) between events and/or variables. In the most basic format hypotheses are bi-variate in that they state the influence of one independent variable (IV) on one dependent variable (DV).

Null Hypothesis (H_0) - statement saying that nothing is different. For studying group differences the null hypothesis states that there are no differences between group means of some variable of interest. For

the study of association the null hypothesis states that there are no relationships between the variables of interest.

Alternative Hypothesis (H_1) - is the opposite of the null hypothesis and states that there is a significant[51] difference between the means or that a relationship exists between the variables.

[51] A difference between the means is likely to exist; the question is if that difference is significant so that an inference can be made based on the results.

Alternative hypotheses can be further classified into *non-directional* and *directional.*

Non-directional alternative hypotheses - predict that the means of the population differ significantly but do not make a specific prediction about the direction of the difference (which one is higher or lower). These types of hypotheses can be answered using two-sided (two-tailed) statistical tests[52].

[52] If the statistics computed by the test follows a symmetrical distribution, there are three possible alternatives for defining hypotheses to test, two for one-sided tests and one for a two-sided test. The one-sided tests look only to one side, left or right of the distribution curve of the statistic, effectively testing for one direction of the relationship while ignoring the other. A two-sided test tests both tails of the statistic distribution, but with less resolution. Two-sided statistical tests are considered less powerful than one-sided tests, which are used to test the directionality of the hypothesis in addition to the significance of the difference.

Directional alternative hypotheses - are more specific in that in addition to predicting that the means of the groups differ on some variable it also predicts which of the means will be higher and which lower. These hypotheses can be tested using more powerful one-sided (one-tailed) statistical tests.

Because the one-sided statistical tests are more powerful than the two-sided variety, using directional alternative hypotheses is preferred to the use of non-directional alternative hypotheses.

A good hypothesis includes three elements:

- A clear statement of the causal relationship to be tested;
- A clear indication of the direction of that causal relationship, if known;
- A clear indication of the variables between which the causal relation occurs.

Hypothesis Testing

Testing a hypothesis means to determine if the null hypothesis (H_0) can be rejected with (acceptable) confidence. For this reason, statistical tests compute the *p-value* as the probability that the presently computed value of the statistic will be obtained if the null hypothesis is true[53]. Therefore, if the value of p is very small, the null hypothesis (H_0) can be rejected and the alternative hypothesis (H_1) should be accepted.

[53] **Important**: p does ***NOT*** provide the probability that the null hypothesis is true.

One commonly accepted cutoff value for p is 0.05. If the computed p value is > 0.05, the test indicates that the null hypothesis should be accepted (or that the test fails to reject the null hypothesis). If the

computed p value is < 0.05 the null hypothesis may be rejected and the alternative hypothesis is accepted, indicating that the differences or relationships found seem to be statistically significant.

The reference p-value is selected for each analysis individually based on the level of confidence in the predictive power of the test necessary to generalize the findings from the sample to the population from which the sample was drawn.

The p-*Value Controversy*

In recent years scientists have voiced concerns about potential misuse of the p-value in research (Arnheim, Greenland, and Blake 2019). There seem to be a few more widely accepted explanations for this: misunderstanding of what the p-value is, tradition as reflected by the education researchers receive in their formative years, and journal reliance on p-values in accepting submissions for publication.

According to the American Statistical Society (ASA), an informal explanation of the p-value is (Wasserstein and Lazar 2016):

> "A p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

Acknowledging this controversy, the American Statistical Association (ASA) recommends researchers consider and follow a few guiding principles in designing, conducting, and reporting their studies (Wasserstein and Lazar 2016):

- The smaller the p-value is, the more incompatible the data is with the null hypothesis, given a set of assumptions hold true. That is, p-values are an indication of how compatible or incompatible the data are with a hypothesized statistical model.
- The p-value is an indicate about how the data relates to a hypothetical explanation, but not about the explanation itself. That is, the p-value does not represent the probability of the hypothesis being true or false.
- The results of an analysis should not be interpreted as a hard yes or no as a statistical finding is not automatically true or false depending on where it falls related to the p-value threshold. That is, scientific conclusions or business decisions should not be based on the p-value alone.

- *P*-hacking or cherry-picking the results tends to generate a body of research skewed towards significant findings. This can be avoided through transparency and open and full reporting of a study and its findings.
- Even weak treatments can produce small *p*-values if the sample is large enough. Or, alternatively, strong treatments may produce irrelevant *p*-values if the sample is not adequate or the measurements are incorrect. That is, the *p*-value cannot measure the size of an effect because statistical significance is not the same thing as scientific or human significance.
- Because it provides limited information, the *p*-value requires a context in which to be interpreted. That is, the *p*-value is irrelevant by itself.

The proposed solution for issues raised by the inadequate use *p*-values in research studies is to use other approaches instead of or in addition to it. Because they are easier to reason about, the most common suggestion is to use confidence intervals[54] instead of the *p*-value.

[54] Confidence intervals describe the variability surrounding the sample point estimate. The wider the interval, the less confident one can be about the estimate of the population mean. In general, the larger the sample size, the more precise the estimate is.

Despite all these issues scientists raise, the *p*-value remains a valuable tool in the researcher's toolbox. It just needs to be used with care, not treated as a binary, definitive answer, and the research using it should observe the appropriate guidelines for design, collection, and reporting. This resource uses the *p*-value in its traditional acception and attempts to follow, as much as possible, the above mentioned guiding principles.

How to Choose the Appropriate Statistical Test

Two simple criteria, *type of variable's scale* and *number of variables of each type* (dependent and independent), can be used as the starting point for determining which statistical analysis would be more appropriate[55]. Of course, once a possible analysis is selected, it should be carefully considered, as not all analyses work in all instances. If we were to build a table or diagram of all possible alternatives, for each specific case, it would quickly become unusable. Therefore, tables 1, 2, and 3 offer some guidelines for what general type of analysis one should start from (adapted from Hatcher & Stepanski (1994)).

[55] The MyReLab website (https://www.myrelab.com) offers a tool that helps with the selection an appropriate statistical test.

Multiple types of analyses can be applied for the same combination of variables/scales. The final selection depends the specifics of the analysis as it applies to the actual data. For example, Table 1 shows that three statistical tests can be used for an analysis with one nominal independent variable and an interval or ratio dependent variable. The Kruskall-Wallis test is usually used for ordinal dependent variables, but can be

used with interval/ratio dependent variables when these show significant departures from normality. Similarly, the t-Test is applicable only if the independent variable has only two possible values. Therefore, when deciding which analysis to use, the requirements and assumptions of each statistical test should be carefully considered.

The advice for how to choose when to apply the most commonly used statistical analyses presented in Tables 1, 2, and 3 has been adapted from Hatcher & Stepanski (1994).

ONE Independent Variable	ONE Dependent Variable	Statistical Analysis
Nominal	Nominal	Chi-Square
Nominal	Ordinal/Interval/Ratio	Kruskal-Wallis
Nominal	Interval/Ratio	t-Test, One-Way ANOVA
Ordinal/Interval/Ratio	Ordinal/Interval/Ratio	Spearman Correlations Coefficient
Interval/Ratio	Interval/Ratio	Pearson Correlations Coefficient

Table 1: ONE DV x ONE IV

MANY Independent Variables	ONE Dependent Variable	Statistical Analysis
Nominal/Interval/Ratio	Nominal/Ordinal	Logistic Regression
Interval/Ratio	Nominal	Discriminant Analysis
Nominal	Interval/Ratio	Factorial ANOVA
Nominal/Interval/Ratio	Interval/Ratio	ANCOVA, Multiple Regression

Table 2: ONE DV x MANY IVs

Independent Variable(s)	MANY Dependent Variables	Statistical Analysis
Nominal (ONE)	Interval/Ratio	One-Way ANOVA
Nominal (MANY)	Interval/Ratio	Factorial MANOVA
Nominal/Interval/Ratio (MANY)	Interval/Ratio	MANCOVA
Interval/Ratio (MANY)	Interval/Ratio	Canonical Correlations

Table 3: MANY DV x ONE or MANY IVs

There are many resources available to help with deciding what statistical test to use for data analysis. For example, Bruce Frey's (2016) book *There's a Stat for That! What to Do and When to Do It*, provides a thorough overview and guides through the selection process, which makes it a worthy addition to any researcher's toolbox.

Effect Size and Power

An experiment, or a study in general, should be designed to be sufficiently sensitive to be able to detect any differences the population may exhibit. The most direct ways to increase the sensitivity is to increase the sample size, by choosing treatments expected to produce large effects, and by reducing unexpected variance.

Relative Treatment Magnitude

The most popular measure of treatment magnitude is called *omega squared* (ω^2)[56]. It is a relative measure that reflects the portion (proportional amount) of population variance that can be attributed to the experimental treatment. That is, the proportion of variability *explained* by the treatment or, more commonly, *explained variance*. Its value is 0 if the treatment effects are absent in the population and has values between 0 and 1 if the effect is present.

[56] Another measure used is the squared multiple-correlation coefficient, which represents how much of the total variation is associated with the variation in treatment.

Based on the value of ω^2, in the behavioral sciences field, the effect size can be interpreted as (Cohen 1977):

- *Small*, for a value of .01;
- *Medium*, for a value of .06;
- *Large*, for a value of .15 or greater.

But how would one know whether the treatment is weak or not? Effects size, as measured by ω^2 is, basically, the ratio between the *variance due to treatment* and *total variance (treatment + error)*. First, the actual effect size can only be estimated after the data is known. So, how would one estimate the treatment effect size at design stage? There are a few possibilities:

1. Deep knowledge of theory should be the primary source of information when estimating the potential strength of an intervention.

2. Search literature for independent variables that seem to produce large effects; use similar research published by others.

3. Choose the treatment and then run preliminary or pilot studies. Use the data to estimate the effect size for the main study. Eventually adjust treatment if needed.

In social sciences it is unlikely to observe large effect sizes[57]*. It is often the case that if a study has an IV that has large effects, that study is just the first step. Further refinements, looking as components of that first IV, will observe theoretical relevance for increasingly smaller effect sizes.*

[57] Perception of effect size for the same treatment may differ between fields, researchers, and studies. This is why it is advisable to base any preliminary (at design time) estimates of effect size on theory and prior research as close to the desired field as possible, followed by pilot testing.

Standardized Effect Size[58]

[58] The MyReLab website (https://www.myrelab.com) offers a more comprehensive power analysis tool.

Known as Cohen's d, the standardized effect size represents the difference between the means of two groups divided by the Standard Deviation (SD), in absolute values.

$$d = \frac{|mean_{s1} - mean_{s2}|}{SD}$$

Fundamentally, Cohen's d expresses the difference between two means in term of Standard Deviation units. It can be interpreted as an equivalent to a z-score for the standard normal distribution. Therefore, if the effect size is 0.6 (SDs above average) between group 1 and group 2, with mean of group 1 > mean of group 2, then group 1, on average, exceeds the values of 59% of group 2. While unlikely to observe standardized effect sizes this large in real life, if its value is > 1, the difference between the two means is > 1 SD, while for d > 2, the difference between means is > 2 SDs.

According to Cohen and later Sawilowsky (Sawilowsky 2009), the standardized effect sizes can be thought of as:

- .01 - Very Small
- 0.2 - Small
- 0.5 - Medium
- 0.8 - Large
- 1.2 - Very Large
- 2.0 - Huge

Note: For independent groups, the Standard Deviation used to compute Cohen's d is the pooled[59] Standard Deviation.

[59] Pooled/combined/composite variance is based on the variance of multiple populations when the variance of the population is the same while the means may be different.

Controlling Type I and Type II Errors

In statistical analysis *power* is the probability the findings will reject a false null hypothesis. That is, when an effect is present, *power* is the likelihood that the effect is detected. So, why should the power of an experiment be controlled? First, because an experiment's power represents the degree to which it can detect differences in treatment and the chances that the experiment can be replicated. Second, a power analysis will help avoid wasting resources when not necessary[60].

[60] For example, adding more participants to a study can be costly in both time and money.

Overall, the statistical power of a test is determined by three factors:

- How large is the difference between the variables measured for the two or more groups involved in the study. A small difference produced by the treatment or cause will require for the study to have more power.
- What level of significance (p-value) is sought[61]. The lower the p-value the higher the power necessary to confirm the difference.
- How often the effects occur in the study groups. A study's power peaks when about half of the population exhibits the effect.

[61] For example, 0.05, 0.01, or 0.001

Controlling Power Through Sample Size

Effect Size ω^2	Power								
	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
				$\alpha = 0.05$					
0.01	21	53	83	113	144	179	219	271	354
0.06	5	10	14	19	24	30	36	44	57
0.15	3	5	6	8	10	12	14	17	22
				$\alpha = 0.01$					
0.01	70	116	156	194	232	274	323	385	478
0.06	13	20	26	32	38	45	53	62	77
0.15	6	8	11	13	15	18	20	24	29

Figure 1: Relationship between Power, Effect Size (ω^2), Significance, and Group Sample Size

Power, *Effect Size*, and *Significance* influence the number of participants that are necessary to be able to observe differences between groups or variables of interest.

Figure 1 (adapted from Keppel (1991), p. 72) illustrates the relationship between *Power*, *Effect Size*, *Significance*, and *Group Size*. The number of participants (sample size) is directly proportional with the *Power* of the design and inversely proportional with the *Effect Size* and

Significance level. For example, for an *Effect Size* of .01 and an expected *Power* of .50 and α of .05, the minimum number of participants in each group of the design would be 144. Therefore, if the study includes two groups, a control group and a treatment group, the entire sample size should be at least 288 participants. That is, the weaker the treatment, the more participants are needed to be able to observe the effects.

As a rule of thumb, a study should be designed for at least a medium *Effect Size* ($\alpha = 0.6$) and a relatively high *Power* (.70 or .80) for a *Significance Level* (ω^2) of .05. A small effect size (weak treatment) requires considerably more resources to be able to observe the effect. Therefore, if possible, the intervention and/or variable(s) should be chosen to avoid weak treatments. Lower power is also to be avoided because it wastes resources (e.g., time, energy) to produce a significant result[62]. Experiments with low power do not produce reliable findings. The *sweet spot* for the design of a study would be at the intersection of the highlighted columns and rows in the figure above.

[62] For example, for a power of .50 there is a 50-50 chance of observing significance.

Population vs. Sample

Population: Is the entire set or pool of similar individuals, items, or events of interest to a researcher. For example, all freshmen at a two-year college can represent a population. Another example would be all the wolves in the Yellowstone National Park. That is, the entire collection to be studied. Can be large or small, depending on the researcher's interests.

Sample: A subset of individuals, items, or events, drawn from the population of interest. To continue the example above, 200 freshmen constitutes a sample. Or the 50 wolves researchers may have tagged with geo locators to follow their behavior. Because in many instances it may be impossible to cover the entire population, a sample allows researchers to use manageable numbers of subjects as representatives of the population to be studied. If the size and characteristics of the sample are appropriate, judgment calls or estimates can be made about the entire population.

Practical Advice on Sample Size

The purpose of computing the size of the minimum necessary sample at design time is to make sure the data is collected from enough participants so that the results can be generalized back to the population the sample was drawn from.

The table in figure 1 can be, for example, used for such a purpose. At the design phase, use prior studies, on the same subject or on similar or related topics, to estimate *effect size* (ω^2) of the treatment intended to be administered. Then choose the level of *significance* (α) you want (e.g., 0.05 or 0.01, or something else) and choose the *power* you wish your experiment to have. Based on the table you can then determine how many participants you should have, at the minimum, in each study group. So, if the experiment has, let's say, two groups, if the table indicates that 30 participants are needed, at the minimum, per group, to observe the values you chose, you would need at least 60 participants in total, equally distributed between the two groups. But this is just a *theoretical* number.

This estimate is just the first step of the process. You should then consider the possibility that not all responses you receive (or measurements of the DV) will be usable, so it would be advisable to adjust upwards the value you calculate so that it is more likely to get the minimum number of *usable* responses.

Besides prior research and immersion in theory, peers working in the same field or one close to it may the best resources to reach out to to determine a meaningful sample size. They are likely to have worked with the same or similar participant pools and have insights into how potential participants may respond to the proposed treatment.

Many of the tests work better if the groups are balanced (in number of participants). Therefore, the procedure for the selection of participants and assignment to the experimental groups should attempt to make that happen. And this is not just a matter of, say, assigning incoming participants (they come in randomly) alternatively to each study groups (treatment condition). You should also consider the type of treatment you are applying and the likelihood for the participant to drop early or to not complete the entire study.[63]

[63] For example, the more complex and cognitively involved the task is, the more likely is for the participant to drop early, skip responses, or just guess, situations in which the experimenter ends up with missing or unusable data, or incomplete records.

In the end, the sample size determination is based both on numerical computation and the researcher's understanding of the field and his or her grasp of how prior research fared. Usually, power estimates are based on the *minimum* effect size the researcher *wishes* to detect. A *realistic* estimate is usually based on prior research.

Below I listed resources that can help determine power and sample size. The first one is a software application that allow you to make the necessary computations. The second one is a resource that helps you understand how to use the R statistical computing language for the same purpose.

G*Power (http://www.gpower.hhu.de)[64]

Power analysis in R (https://www.statmethods.net/stats/power.html)[65].

MyReLab website (https://www.myrelab.com)[66].

[64] http://www.gpower.hhu.de

[65] https://www.statmethods.net/stats/power.html

[66] https://www.myrelab.com

Sample Size

A sample is a subset of a population selected to represent the entire population in a study. It is used because while the study is interested in learning more about the population, it may not always be feasible to study every member of that population. The reasons can be, for example, unfeasible incurred costs[67] or, it may just be impossible due to geographical distribution or availability.

[67] For example in the form of time and money.

Factors Affecting the Size of the Sample

For any study, the sample size depends on a few elements:

- Level of significance (what is acceptable as an error rate). This is the p-value, such as 95% ($\alpha = 0.05$) which indicates the researcher's readiness to accept a certain probability that the obtained result is due to chance and not to the intervention or researcher's intention.
- Power, discussed in the context of Type II error, the failure to detect a difference when one doesn't exist, or the chance of false negatives. The power of the study increases with the decrease in the chance of committing a Type II error. Usually 80% is an acceptable level for the power of a study. It means that the researcher is accepting the study misses a real difference in one in five times. For more strict studies, power can be increased to 90% or more.
- Expected effect size, represents the difference between a variable's value in one groups and its value in another group. It is inversely proportional with the sample size. There is no formula to determine the effect size. Most often is determined based on prior studies reported in the literature.
- Effect prevalence in the population, estimated from previous studies.
- Population standard deviation, a measure of dispersibility.

When estimating sample size a researcher should consider other elements as well, such as administrative issues, costs, possible participant

response rate, and so forth. Each study should be considered from all angles and all potential elements that could participate in determining the sample should be studied carefully.

Methods of Determining the Sample Size

A cursory review of the literature shows that sample size can be determined in many ways using formulas and/or tables and that there is no universal "formula" for sample size calculations. Each of the methods has a recommended use.

For example, an approach to make a rough determination of a sample size for an experimental design using effect size and power was discussed in the Controlling Power Through Sample Size section. You will also find many sample size calculators available online, many of them based on Cochran's Sample Size Formula.

Cochran's Sample Size Formula

Used to compute an ideal sample size for a desired level of precision, it is recommended to be used for studies with infinite populations (Cochran 1977).

$$n_0 = \frac{z^2 \cdot p \cdot (1-p)}{e^2}$$

e: desired level of precision, the margin of error

p: the fraction of the population (as percentage) that displays the attribute

z: the z-value, extracted from a z-table[68].

Let's consider an example. Think of a study of students in a large university campus for which we don't know the campus size[69]. We are interested in finding the percentage of students who eat lunch at the campus dinner halls but we do not have insider information. The question is how many students would we need to ask that question to be able to determine, with reasonable confidence, what percentage of students conform to the sought behavior. Given the lack of information we start by considering that 50% of the students eat lunch at the school dining halls, which provides the largest variability. Then we consider a 95% confidence level (leading to an α=0.05) and a $\pm$5% precision. From the z-tables, the value for z is 1.96. Therefore, the theoretical sample would be:

$$n_0 = \frac{1.96^2 \cdot 0.5 \cdot (1-0.5)}{0.05^2} = 384.16 \approx 385$$

How to find the value of z from a *z-table*. The procedure is:

[68] The entry for z in a z-table represents the area under the normal distribution curve to the left of z (Figure 2).

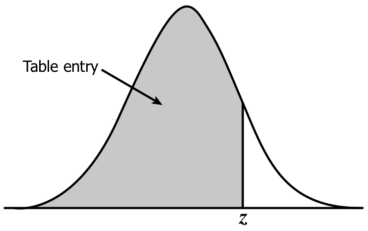

Figure 2: Area represented by the z-value.

[69] For example a large campus may have 10 - 15 K students

1. Convert the confidence level from percent form to decimal form as value between 0 and 1. (95% $\rightarrow$ 0.95)
2. Subtract the value from 1 and divide by 2 to find out how much is half (1 - 0.95 = 0.05; 0.05/2 = 0.025)
3. Add the value from 2) to the value from 1) (0.95 + 0.025 = 0.975)
4. Look for the value obtained in step 3) in table values. In Table 4 the value sits at the intersection of row labeled 1.9 and column labeled 0.06.
5. Determine the value of z by adding the value for the column with the value for the row obtained in step 4 (1.9 + 0.06 = 1.96).

Cochran's Modified Formula for Finite Populations

A slightly modified formula can be used if the size of the population is known.

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

n_0: Cochran's sample size computed using the formula for ideal sample size;

N: the size of the population[70].

As an example, let's look at the same problem as before but for a much smaller campus of N = 600 students. While we can still use the theoretical sample of 385 participants computed before, do we need to? The necessary sample size may be smaller.

$$n = \frac{385}{1 + \frac{385 - 1}{600}} = 234.76 \approx 235$$

The result of this computation indicates that for smaller populations the number of subjects (sample size) can be smaller (235 vs. 385) for the researchers to be reasonably confident of the findings.

[70] The sample size is dependent on the size of the population until the population reaches about 40-50 K, after which the increase is almost none. Therefore, if the estimated population is this large or larger, the theoretical sample size, as computed for an unknown population, is about equal to the one generated by the modified formula.

Yamane's Simplified Formula for Sample Size

To make it simpler to compute the sample size without over estimating it when the population is known Yamane (1967) proposed the following formula:

$$n = \frac{N}{1 + N \cdot e^2}$$

N - population size

e - level of precision

Using the same example as before, Yamane's formula would suggest a sample size of 240 subjects for a student population of 600.

$$n = \frac{600}{1 + 600 \cdot 0.05^2} = 240$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table 4: z-table

Assumptions and Outliers

Most statistical tests rely on a set of assumptions about the data they are used to analyze. These assumptions ensure that the test performs as expected and that the results can be interpreted with a high degree of confidence. Understanding when and how these assumptions lead to biases in analysis and what the consequences are is essential to meaningful data analysis.

For example, many tests[71] assume that the data is normally distributed. While these statistical tests vary in how well they can deal with departures from normality, if the data does not follow a normal distribution, the results provided may not have sufficient power to explain the phenomenon properly and therefore is important to confirm before hand that if the analysis is conducted the interpretation is valid. In situations in which normality is not observed it is usually possible to find a different statistical test that is less sensitive to departures from normality.

[71] E.g., Linear Regression, ANOVA.

What is important to remember is that each statistical test functions better when certain conditions, established by those who developed the test, are met. Therefore, assumptions need to be tested to make sure that the statistical test is appropriate to be used for the data set.

The most common parametric statistical analyses expect the data to be *normally distributed* and *homogeneous*.

Normality

Compares the sample's distribution with the normal distribution (Figure 3).

To study the normality of a data sample we use the *skewness* and *kurtosis* of the sample's distribution to determine its departure from the theoretical normal distribution.

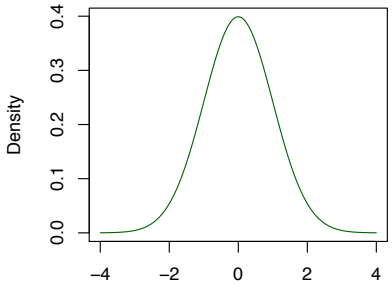

Figure 3: The normal distribution

Skewness

Measures the deviation from symmetry as compared to normal distribution, which has a *skewness* of 0. A value other than 0 means that the data is either skewed to the left or to the right of the corresponding normal distribution. A positive skewness value indicates that the sample's distribution is skewed towards higher values, to the right. A negative value indicates a sample distribution skewed towards smaller values, to the left.

Although largely arbitrary, in most situations a simple rule of thumb of ± 1 can be used to interpret the sample's distribution skewness. If skewness is either greater then 1 or smaller than -1, the distribution it is computed from shows a significant departure from symmetry.

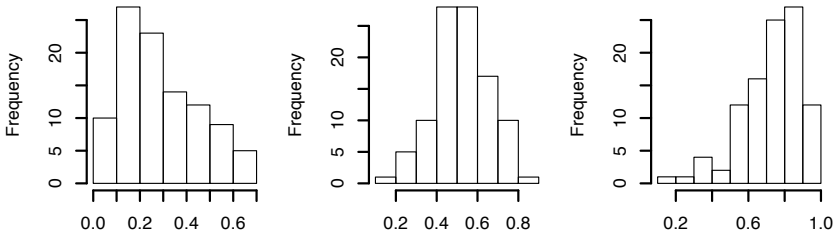

Figure 4: Left skewed, normal, and right skewed distributions

A more detailed interpretation (Bulmer 1979):

- Highly skewed if skewness < -1 or $> +1$;
- Moderately skewed, if skewness is between (-1 and -1/2) or between (+1/2 and +1);
- Approximately symmetric, if skewness is between -1/2 and +1/2.

Kurtosis

The normal distribution has a "balanced" shape, not too peaked and not too flat. *Kurtosis* is a measure of how much and in which way the "peakness" of the described distribution differs from the theoretical normal distribution. The *kurtosis* of the normal distribution is 3. A value other than 3 means that the distribution is either flatter or more peaked than the normal distribution. If the value is positive (> 3), the distribution is more peaked and is called to be *leptokurtotic*, with longer and fatter tails and higher and sharper central peak. If the value is negative (< 3), the distribution is flatter and is called *platykurtic*, which shorter and thinner tails and lower and broader central peak.

Tests of Normality

Visual Checks

Visual checks can be used to assess, roughly, how close the sample's distribution is to the normal distribution. This can be accomplished using *Quantile-Quantile plots (qq-plots)* that help visualize if a set of data comes from a theoretical distribution, the normal distribution in this case. In essence a *qq-plot* is just a scatter plot of two sets of quantiles, theoretical and observed, as they relate to each other[72]. If the sample came from the desired distribution, the plot will roughly approximate a straight diagonal line. Any departure from that shape is an indication of departures from normality (Figure 5).

[72] Quantiles, also known as percentiles, are points in the data that divide the observations in intervals with equal probabilities. In essence, quantiles are just the data sorted in ascending order.

Visually inspecting a *qq-plot* does not offer an exact and definitive proof that the sample data comes from a normal distribution. Nevertheless, it can be used to determine if further testing is necessary. For example, if the *qq-plot* has an S-shaped appearance, the sample data may be skewed.

More Specific Tests

Visual checks, while helpful, cannot be relied upon in most situations. Therefore, there are more specific tests that can be used to perform a more formal evaluation.

D'Agostino-Pearson omnibus test - Uses test statistics that combine skewness and kurtosis to compute a single p-value. This test has a tendency to hyper actively reject normality for small samples for which reason is not recommended to test normality of samples less than 20.

Kolmogorov-Smirnov test - Non-parametric test used to compare two samples that can serve as a goodness of fit test. When testing for normality for example, the sample data is standardized and compared with the theoretical normal distribution. It is less powerful than some other tests, such as Shapiro-Wilk.

Shapiro-Wilk's W test - Tests the null hypothesis that the data in the sample is part of a normally distributed population. The test computes the value of the W statistic and a p-value probability. Considering the commonly accepted 0.05 value for p, any computed p-value greater than 0.05 indicates that the null hypothesis cannot be rejected and therefore it should be true and that the assumption of normality is upheld. Does not work well for samples with many identical values.

Chi-square test of goodness-of-fit - Looks at a single categorical variable from a population and attempts to assess how close to or consistent

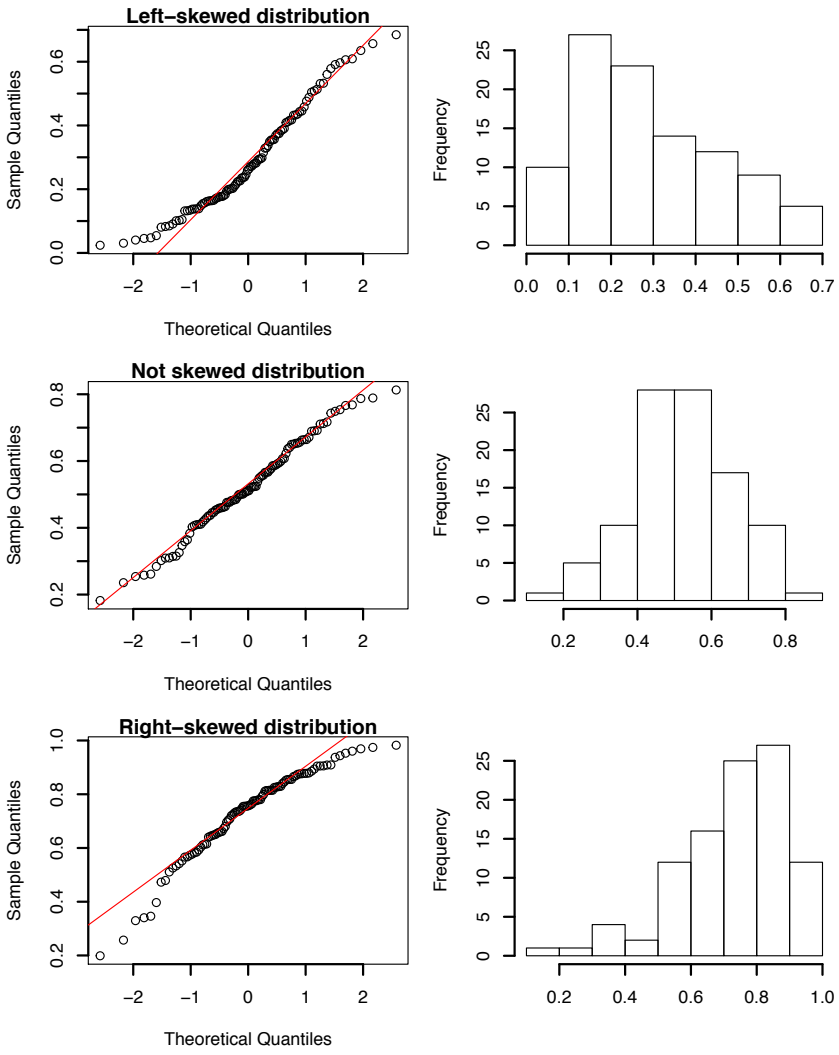

Figure 5: qq-plot behavior as function of distribution shape

with a hypothesized distribution the actual distribution of that varialbe is. See Goodness-of-Fit Test for more details.

Homogeneity of Variances or Homoscedasticity

The assumption of *homogeneity of variances* expects the variances in the different groups of the design to be identical. The homogeneity of variances is a standard assumption for many statistical tests and therefore it needs to be assessed so that the test results can be interpreted with confidence.

So, why is it important to test it? Many of the most common tests in statistical analysis[73] are part of a category of tests called *general linear models*. These models are *linear* in the sense that they add *things* together. To be able to add things together, these tests assume that the distributions of the things being added are the same. Otherwise, if distributions are not the same, the results/estimate and the conclusion could be biased (usually overestimation of goodness of fit), therefore effectively rendering the test results unusable.

[73] For example, ANOVA assumes that observations are randomly selected from the population and that all observations come from the same population (or underlying group), with the same degree of variability, following the same distribution.

In case of ANOVA, for example, if the variance of separate groups is significantly different, the p-values the test computes are no longer accurate because they are calculated based on the fact that certain results occur if the null hypothesis is true.

Let's look at a few of the more common tests one can use to investigate the homogeneity of variances.

Bartlett test for homogeneity of variances - Tests the hypothesis (null hypothesis) that the variances in each sample groups are the same. This is the preferred test if the data is normally distributed, but it has a higher likelihood to produce *false positive* results when the data is non-normal. Therefore, Bartlett's is preferred when the data comes from a known normal or close to normal distribution.

Levene's test - A more robust inferential statistic alternative to the Bartlett test used to evaluate or assess the equality of variances for a variable that has two or more groups. If the computed p-value is less then the set significance level[74], the sample variance is unlikely due to random sampling from a population with equal variances. Therefore, the null hypothesis that the variances are the same is rejected. The *Levene's test* is less sensitive to departures from normality than Bartlett's.

[74] E.g., the usual 0.05.

Fligner-Killeen test - Non-parametric test, robust for data sets with departures from normality. Can be useful if the normality of the sample data for groups is not observed.

Outliers

Outliers, extreme values data depart significantly from the majority of the values in the data set, can have substantial influence on the results of a statistical analysis. For example, they can skew the sample distribution to such extent that its departure from normality makes a large number of parametric statistical tests that require the data to be normal no longer applicable. Or, when present in many real-life time series data sets can significantly alter the outcomes of the analysis. For example, when analyzing an economic phenomenon, a hurricane affecting some link of the supply chain can produce a specific and significant change for a short period of time, which can influence the understanding of the phenomenon in average conditions.

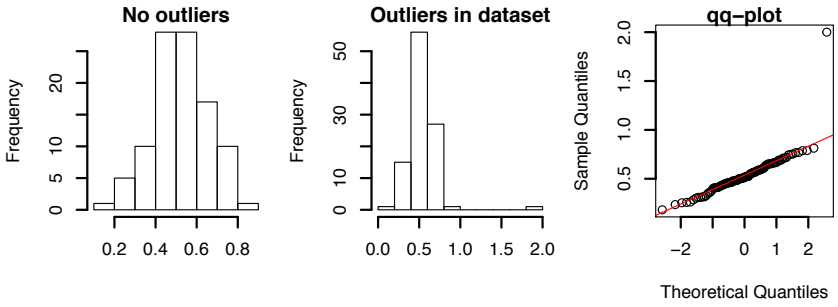

Figure 6: How outliers influence distribution shape

Figure 6 shows how outliers can influence the shape of the dataset's distribution. The graph on the left shows a set of normally distributed data[75]. The graph in the middle represents the same data set to which two data points were added as outliers[76]. A visual inspection shows how the addition of the two outlying data points "pushes" the distribution to the left, away from the shape of normally distributed data. The qq-plot in the graph on the right shows how the two outliers influence dataset normality.

Most of the times outliers can be identified visually, in graphical representation such as scatterplots (Figure 7 left), box plots (Figure 7 right), or mahalanobis distances[77]. In scatterplots outliers are represented as points plotted away from the "cloud" formed by majority of the data set. In box plots[78] they would be indicated by individual points plotted on the graph. Mahalanobis distances[79], when represented as a histogram, will indicate the possible existence of an outliers by the presence of bars at the right side of the graph.

Linear regression presents itself as one of the better ways to explain how outliers can influence the outcomes of the analysis because these extreme values can substantially influence the slope of the regression

[75] The same data used previously in this section to discuss skewness. For this graph the data ranges between 0 and 1.

[76] For exemplification, one value, 2, was added to simulate a possible outlier. The point is located outside the range of the original dataset.

[77] The use of mahalanobis distance is better understood in context. See the outliers section in multiple regression.

[78] Also known as box-and-whisker diagrams, are used to represent data graphically based on quartiles

[79] Mahalanobis distance (MD) is the distance between two points in multivariate space. It measures the distance relative to a base or central point considered as an overall mean for multivariate data (centroid). The centroid is a point in multivariate space where the means of all variables intersect.

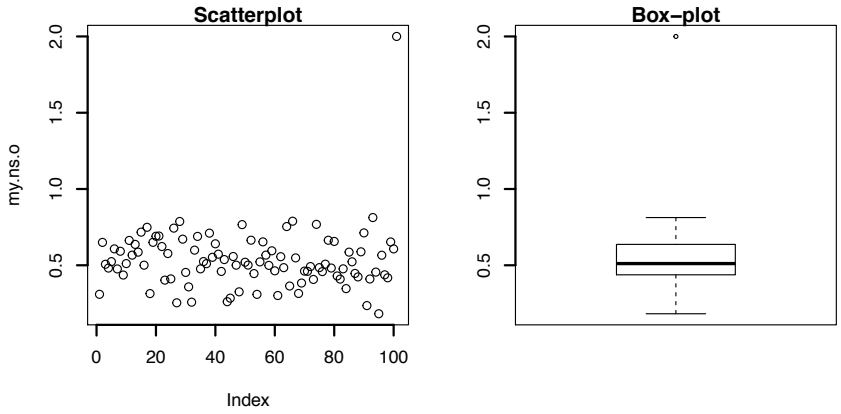

Figure 7: Scatterplot, Box plot, and mahalanobis distances

line[80]. In figure 8, the left side shows the data set without outliers and the right side shows the data set with one added outlier. Visually, the slope of the regression line (in red) changes from negative to positive when the outliers are not removed from the dataset. The change in slope may look small in the image below, but it could be significant in the analysis.

[80] This, in turn, will affect how well the regression equation will fit the data and the correlations coefficient. A single outlier can decrease the value of a correlation coefficient to the extent that the analysis rejects the existence of a real phenomenon.

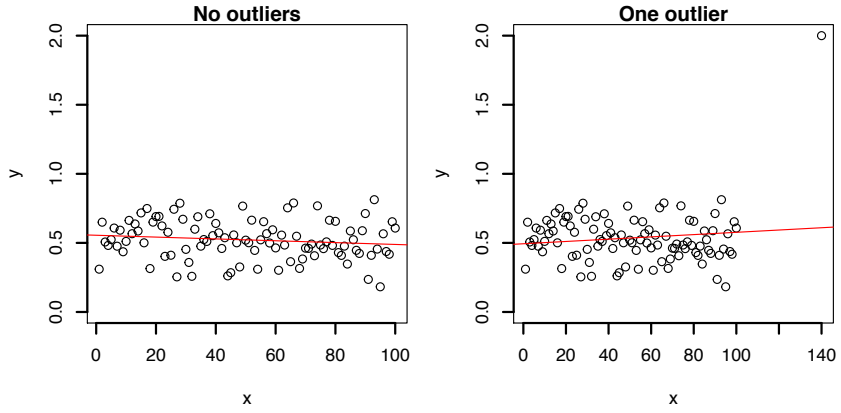

Figure 8: Outlier influence on regression outcomes

The way outliers are approached depends largely on the type of analysis being conducted and the type of data being analyzed. So far a widely accepted numerical method to deal with outliers does not seem to exist. Therefore, while in this resource approaches to outliers will be discussed when relevant to the statistical test and example, there are few guidelines that may be helpful across tests.

Outliers investigation is highly contextual and the criteria used for analysis ranges from theoretical necessity to common sense. There are legitimate situations when automatic (mathematical) removal of outliers makes sense. For example, when studying people reaction times to

some stimulus and the vast majority of data points are in the milliseconds range, values in the seconds range most likely indicate distracted participants and therefore they could be safely removed. Alternatively, in situations where outliers consistently show up across groups, design cells, or successive measurement, they could indicate the existence of a different phenomenon than those under direct study. When their presence is detected, each outlier should be investigated individually and a decision made if it should be corrected[81], removed[82], or retained[83] in the analysis.

[81] For example, in situations when the existence of an outlier is due to improper data input.

[82] In cases when the data point is determined to be an aberrant measurement.

[83] When the outlier proves to be a true measurement and there are practical and/or theoretical reasons to be included in the analysis.

Normalization, Standardization, and Data Transformation

In statistics the term *normalization* can have multiple meanings. For example, it can mean to adjust values of variables to be analyzed to a common scale when these variables are measured using different scales. Alternatively, normalization may mean to transform the values of a sample to get it closer to the normal distribution, to fulfill the assumption of normally distributed data underlying many common statistical tests. Or, in other situations, normalization could be an attempt to eliminate the effects of known influences.

Sometimes the term *normalization* is used interchangeably with the term *standardization*. While both serve the same general purpose, there is a difference:

- *Normalization*: intended to scale a variable to a range of values between 0 and 1;
- *Standardization* (e.g., z-score): intended to transform a variable to have its mean = 0 and SD[84] = 1.

[84] Standard Deviation

Some of the processes that fall under the concept of *normalization* may be called *data transformation*[85]. While all this may be confusing, the process always involves a purposeful conversion of the data of a data set from its current format to a different format. In most situations, the process also involves eliminating the units of measure with the intent of making the comparison easier.

[85] *IMPORTANT:* A transformed value, such as a log of the data, has little informational value when interpreted and presented. Therefore the once the analysis has been performed, when discussing the findings, that data should be converted back to its original format by applying the reverse sequence of formulae that was used to transform it in the first place.

Scaling/Rescaling

A very simple process which aims to change the spread of the data and/or the position of the data points. The transformation uses a simple linear equation of the form $y = a \times x + b$, but leaves unchanged the shape of the distribution or the z-scores[86]. It will change the data *median*,

[86] Because they are calculated as a ratio of the difference between the actual value and the sample's mean and the sample's standard deviation. In essence, the z-socre indicates how many standard deviations from the mean the data point is.

mean (μ), and *standard deviation* (σ).

The more common are:

- *Range scaling*: change the data from one range to another (magnification or reduction)[87];
- *Mean centering*: changes the data by subtracting the mean of the sample from each data point, resulting in a shifting of the data towards the mean, effectively making the mean of the scaled data set to be 0[88];
- *Standardization*: is intended to make data samples comparable. The process does not change the shape of the distribution, only the mean and the standard deviation. The most common method is the *z-transform* used to convert the data to *z*-scores. Also called *autoscaling*, the *z-transform* makes the data comparable by transforming observed data into multiples of its standard deviation (SD). The mean of the *z-transformed* sample is equal to 0. If the original distribution of the data is normal, the transformed data will also follow a normal distribution with a *mean* of 0 and a *standard deviation* of 1 (Figure 9).

[87] Often times done to bring the variables in an analysis to the same scale.

[88] Often done to center the analysis on the variation part of the data rather than, for example, a center tendency value.

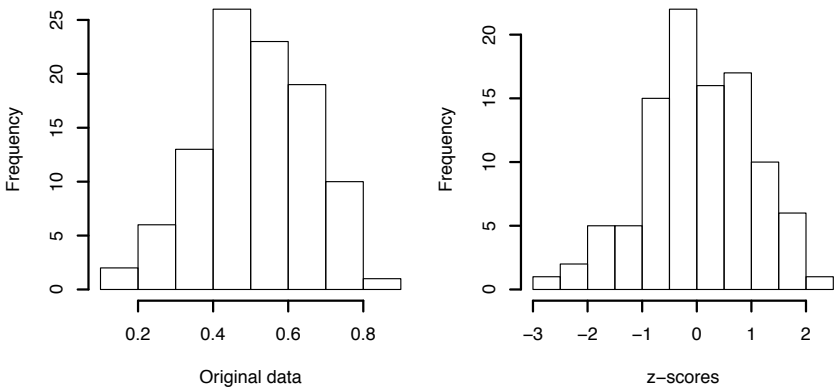

Figure 9: z-transformation of data

Distribution Normalization

There are many real world processes that generate data that do not follow the normal, Gaussian, distribution. Previous sections already discussed how to determine if the data are normally distributed or not. Non-normal data usually fits in one of two categories: 1) follows a different distribution or 2) is a mixture of distributions and data generation processes.

For the data that follows a known distribution[89], one can either iden-

[89] E.g., binomial, log-normal, exponential, etc.

tify the theoretical distribution it follows and analyze the data using tests appropriate for that distribution or somehow transform it to a normal distribution before analysis and then use tests that assume data distribution to be normal.

Dealing with the 2nd category of non-normal data, which displays a mixture of distributions, is more complicated because few, if any, transformations would be able to deal with all that variability for the entire data set. In this case, the data needs to be studied and refined before analysis. For example, an attempt to break down the data can produce sub-samples or categories for which, individually, the distributions are known or recognizable. In this case, transformations can be applied to each sub-sample or category. If the observed data is produced by multiple processes, such as business data produced by complex work activities, multiple shifts, locations, customers, seasonality, etc., an attempt to review the individual process that produced data points and find a common denominator to transform the data to can sometimes help bring the data close enough to the normal distribution so that analysis is possible[90].

[90] The process always works better for larger samples.

Some of the common options are:

- *Box-Cox transformation*: Uses a family of power functions to transform data to a more normal distribution form. The formula used for transformation are simple but computationally intensive. For this reason most statistical analysis packages (e.g., SPSS) offer an option to run Box-Cox transformations on the data set.
- *Log transformation*: The value of each observation is transformed by applying the base 10 or natural logarithm to the observed value. The reverse process is to raise the values at the power of 10 or e[91], depending of what type of logarithm was initially applied. It is especially useful if the original variable follows a log distribution; after transformation the resulted values will be normally distributed.
- *Square-root transformation*: The value of each observation is transformed by taking its square root. To reverse the process is to square the values. Usually used when the variable is a count of something. If the sample includes negative values, the sample should be first rescaled to have all positive values[92].
- *Arcsine transformation*: The value of each observation is transformed by taking the arcsine of the square root of the number. The numbers to be transformed should be between 0 and 1. The resulted unit of measure is radians and the resulting range is $-\pi/2$ to $\pi/2$. It is usually useful for proportions or ratio type data that ranges between 0 and 1.

[91] For the natural logarithms.

[92] For example by adding a certain amount to each observed value.

Let's look at an example of a log transformation. The origin and

meaning of the data set is irrelevant for this transformation example.

A histogram and qq-plot of the original sample data (Figure 10):

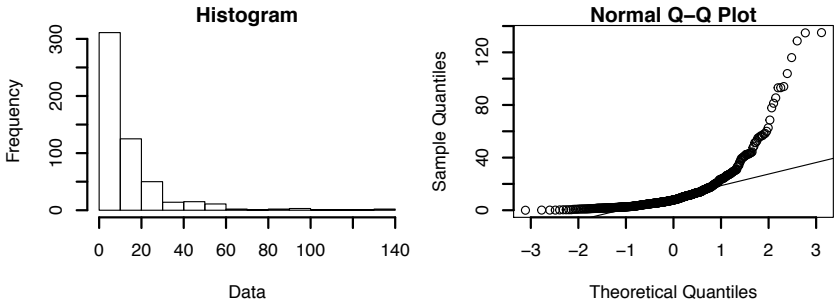

Figure 10: Original data

A visual inspection of the data shows that it is significantly different from a normal distribution and most likely closer to an exponential distribution. So, let's do a log transformation of the data. Once the data has been transformed, it becomes normally distributed (Figure 11).

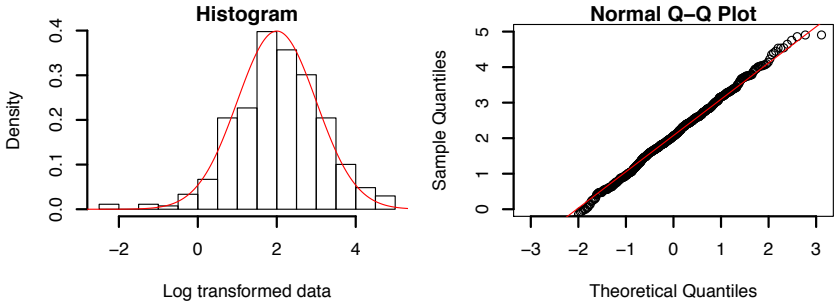

Figure 11: Log transformed data

Much better. Now that the assumption of normality is verified (alongside all other assumptions), the analysis can continue.

Once the analysis is completed the data should be transformed back to its original format for interpretation and reporting.

Correlations

Correlations[93] explore how two or more variables are related to each other. It attempts to assess if the changes in one variable systematically vary with the changes in another. There is no dependence relationship between variables (e.g., IV/DV). Most often correlations are used to look at how variables are correlated to each other in a data set, usually with a focus on the variable(s) of interest.

[93] *Remember*, Correlation does not imply causation!

This section explores the three most common correlation tests, one *parametric* (Pearson) and two *non-parametric* (Spearman and Kendall). All three tests compute a correlation coefficient that can range between -1 and 1. The closer the value is to the extreme (-1 or 1) the stronger the relationship is.

- < 0 - indicates a negative correlation, meaning that the as the value of x increases, the value of y decreases.
- 0 - indicates no association.
- > 0 - indicates a positive correlation, meaning that as the value of x increases, the value of y increases as well.

Null hypothesis (H_0): There is no correlation between the two variables. In this case the correlation coefficient (which depending on test can be r, ϱ, or τ) is zero or close to zero.

The data set used in the examples below is called *mtcars* and is available in R example datasets. The data, covering 11 variables describing cars, was extracted from the 1974 Motor Trend US magazine (Table 5).

The first few rows of the data set are shown in Table 6.

The question asked is: *Is there any relationship between mpg and wt*? That is, is there any correlation between the car's weight and its fuel efficiency?

Variable	Description
mpg	Miles/gallon (US)
cyl	Number of cylinders
disp	Displacement (in cubic inches)
hp	Horsepower
drat	Rear axle ratio
wt	Weight (in 1000 lb)
qsec	1/4 mile time
vs	Engine (0-V, 1-Line)
am	Transmission (0-automatic, 1-manual)
gear	Number of forward gears
carb	Number of carburators

Table 5: Description of variables in the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Table 6: First few rows of the mtcars dataset

Pearson Correlations Test

Parametric test used for two *interval* or *ratio* variables. The test requires the following assumptions to be met:

- Data is bi-variate normal;
- The relationship between variables is linear.

Before running the test, we verify the assumptions for the *Pearson Correlations*[94] test.

To test normality we use the *Shapiro-Wilk* test.

[94] An issue with the Pearson correlations test is the fact that outliers can negatively impact the test results. Therefore, an analysis of outliers should be performed for the sample before the test.

```
    Shapiro-Wilk normality test

data:  my.cor$mpg
W = 0.95, p-value = 0.1
```

```
    Shapiro-Wilk normality test

data:  my.cor$wt
W = 0.94, p-value = 0.09
```

The results show that the distribution of data for both variables is not significantly different from the normal distribution (both p-values are > 0.05), thus verifying the assumption of normality.

Besides using Shapiro-Wilks, data normality can be analyzed using *histograms*, *q-q plots*, and the values of *skewness* and *kurtosis* for each data set. These alternative ways of studying a data set's normality are exemplified for other analyses.

The linearity assumption can be visualized by generating a scatter plot representation with one variable on the X axis and the other variable on the Y axis.

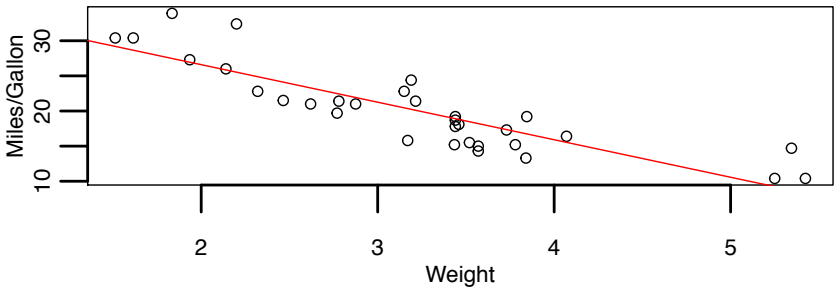

Figure 12: Scatterplot of miles/gallon (mpg) vs. weight (wt)

Looking at the scatter plot (Figure 12) the assumption of linearity seems to hold because the relationship between the two variables seems to be linear along the red line (regression line). Should the pattern of points show a different trend (e.g., curve), the relationship between the two variables is not linear and therefore other correlation tests should be used to analyze it.

With the assumptions verified, let's run the Pearson Correlations test.

```
    Pearson's product-moment correlation

data:  my.cor$mpg and my.cor$wt
t = -9.6, df = 30, p-value = 1e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9338 -0.7441
sample estimates:
    cor
-0.8677
```

The p-value < 0.05 suggests that there is a significant correlation between *mpg* (fuel efficiency) and *wt* (car's weight).

Spearman Correlations Test

Non-parametric test used for two *interval*, *ratio*, or *ordinal* type variables.

Because the *Spearman Correlations* test does not have any assumptions about the data, it can be run directly.

```
    Spearman's rank correlation rho

data:  my.cor$mpg and my.cor$wt
S = 10000, p-value = 1e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
    rho
-0.8864
```

Based on the computed p-value < 0.05 it can be concluded that the two variables are significantly correlated to each other.

Kendall Correlations Test

Non-parametric test used for two *interval*, *ratio*, or *ordinal* type variables.

Because the *Kendall Correlations* test does not have any assumptions about the data, it can be run directly.

```
    Kendall's rank correlation tau

data:  my.cor$mpg and my.cor$wt
z = -5.8, p-value = 7e-09
alternative hypothesis: true tau is not equal to 0
sample estimates:
    tau
-0.7278
```

Based on the computed p-value < 0.05 it can be concluded that the two variables are significantly correlated to each other.

Chi-Square Test

The *Chi-Square Test* applies to *nominal/categorical* variables. It has two forms:

- A one-way classification to estimate how close an observed distribution is to an expected distribution. This type of test is called a *goodness-of-fit test*.
- A two-way classification to estimate whether two variables from the same population are related or not. This type of test is called a *contingency test* or *test of independence*.

The dataset described below will be used to exemplify both types of analyses. It contains all hospital discharges in New York State in 1993 (12,844 records) for patients admitted with an Acute Myocardial Infarction (heart attack) who did not have surgery. The data is defined as shown in Table 7.

Let's take a look at how the data looks like by listing the first few rows of the dataset (Table 8).

Goodness-of-Fit Test

Looks at a single categorical variable from a population and attempts to assess how close to or consistent with a hypothesized distribution the actual distribution of that variable is.

Null hypothesis (H_0): The data follows the *theoretical* distribution.

Before running the analysis, let's look at a basic descriptive statistic, frequency analysis.

Running the chi-square goodness-of-fit test for the *gender* variable will attempt to test for equal counts in every cell of the design.

```
    Chi-squared test for given
    probabilities

data:  table(my.ha$gender)
X-squared = 570, df = 1, p-value <2e-16
```

Variable	Explanation
age	Patient age in years
gender	Patient's gender, coded M for males and F for females
diagnosis	Code based on ICD classification
drg	Diagnosis Related Group (121 - with complications who did not die; 122 - without complications who did not die; 123 - patients who died)
los	Hospital length of stay
died	1 if the patient died in the hospital; 0 if not
charges	$ amount of hospital charges

Table 7: Variables in the heart attack data set

diagnosis	gender	drg	died	charges	los	age
41041	F	122	0	4752	10	79
41041	F	122	0	3941	6	34
41091	F	122	0	3657	5	76
41081	F	122	0	1481	2	80
41091	M	122	0	1681	1	55
41091	M	121	0	6379	9	84

Table 8: First few rows of the heart attack data set

Given the small *p-value*, the null hypothesis can be rejected and it can be considered that the alternative is true.

Example of how to write it up:

The null hypothesis stating that patients are equally distributed across genders, $\chi^2(df = 1) = 573.48, p < 0.001$ *is rejected. Therefore, there are significant differences between the expected number of patients for each gender and the observed number.*

Observed frequencies can help explain the direction of the relationship for the possible underlying explanation or reason. This explanation can be added to the text if relevant for the study.

Test of Independence

In this case the *Chi-Square Test* is used to determine if a significant relationship exists between two categorical variables. The test compares the frequency of each category of one of the variables to the frequency for each of the categories of the other variable.

In this example we will try to find if there number of patients deaths is in some way related to their gender. Therefore, we will be testing the null hypothesis:

The null hypothesis (H_0): There is no relationship between *gender* and the *patient dying* in the hospital.

Gender	Freq
F	5065
M	7779

Table 9: Gender variable frequencies

Let's look at some cell counts first. An idea about how these look like will be useful later when compared with expected cell counts computed from the statistical test.

	0	1
F	4298	767
M	7136	643

Table 10: Gender count and Died (0 = no, 1 = yes) by gender variable counts

Now let's run the test:

```
    Pearson's Chi-squared test with Yates'
    continuity correction

data:  table(my.ha$gender, my.ha$died)
X-squared = 150, df = 1, p-value <2e-16
```

The obtained *p-value* $< .05$ indicates that the null hypothesis (indicating independence) can be safely rejected and the alternative hypothesis (indicating the existence of a relationship between *gender* and *died*) should be accepted.

If the null hypothesis were true, the expected counts are shown in Table 11, compared to the observed frequencies shown in Table 10.

	0	1
F	4509	556
M	6925	854

Table 11: Expected frequencies

A mosaic plot (Figure 13) can help visualize the relative cell sizes.

Now let's look at a chi-square analysis when at least one of the variables has more than one level. In this case we will be using the *drg* variable[95].

[95] The *drg* (Diagnosis Related Group) has three levels: 121 - with complications who did not die; 122 - without complications who did not die; 123 - patients who died.

The cell counts in this case are shown in Table 12.

	121	122	123
F	2328	1970	767
M	3059	4077	643

Table 12: Gender count and Diagnostic Related Group variable counts

Running the test returns the results shown below.

```
    Pearson's Chi-squared test

data:  table(my.ha$gender, my.ha$drg)
X-squared = 280, df = 2, p-value <2e-16
```

Figure 13: Mosaic plot of gender vs. age

And the mosaic plot is presented in Figure 14.

Table 13 shows the expected frequencies when the null hypothesis (H_0) is true. Compare them with the observed frequencies shown in Table 12.

	121	122	123
F	2124	2385	556
M	3263	3662	854

Table 13: Expected frequencies

Example of how to write it up:

The null hypothesis stating that there is no relationship between gender and diagnosis related group classification, $\chi^2(df = 2) = 283.43, p < 0.001$ *is rejected. Therefore, the differences between the number of males and females in the different diagnosis related group are significant.*

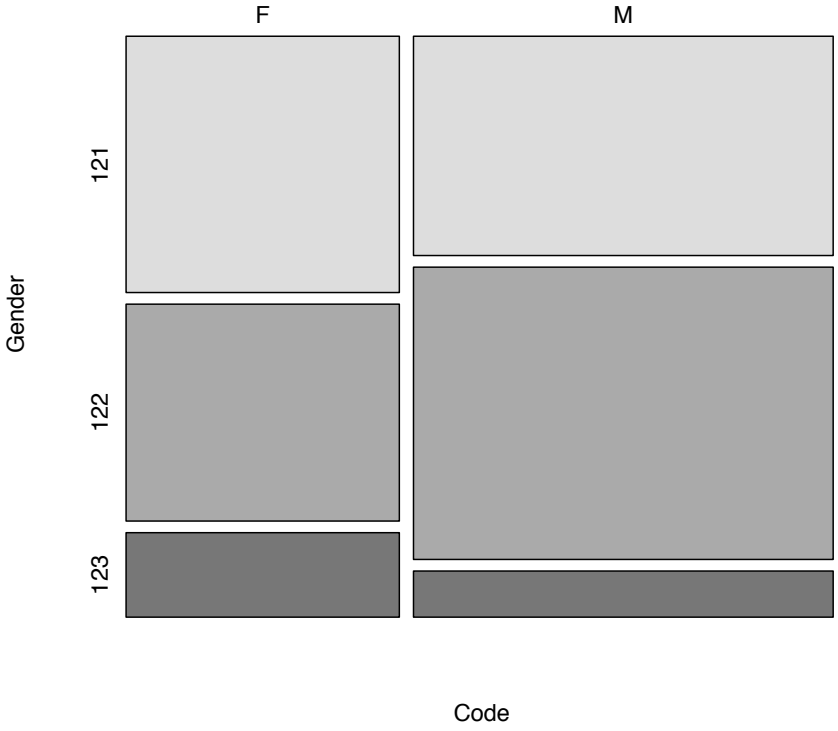

Figure 14: Relationship between Gender and Diagnostic Related Group (drg)

The t-Test

Parametric test of group differences. It is used when:

- There is a single predictor/independent variable (IV);
- The predictor/independent variable is measured on a nominal scale and can have only two values;
- The criterion/dependent variable is measured on an interval or ratio scale.

Independent Samples t-Test

Used when the observations collected under one treatment condition are not related to the observations collected under the other treatment condition. For example, the independent samples *t-Test* is used when randomly selected subjects are exposed to two different experimental conditions (IV) and we want to know if the two groups are or are not different on some characteristic (DV).

The following assumptions must be met to use the independent samples t-Test:

- Normally distributed data.
- Randomly selected samples;
- Samples are independent;

Data normality can be verified by looking at the distribution of the data (histograms) or through a test of normality. *F* or *Levene's* tests can be used to test equality of variances.

The test computes the probability of error for rejecting the null hypothesis of no difference between the two means. Therefore, the *p-value* reported by the *t-Test* represents the probability of being wrong in accepting the research hypothesis (alternative hypothesis) that a difference in means exists.

Gender	Height
F	67
F	67
F	60
M	72
M	71
M	67

Table 14: Excerpt from the height data set

An example: consider studying the differences in height between males and females of the homo sapiens species. We measure 100 individuals, 50 females (F) and 50 males (M). An excerpt of the height measurement data is presented in Table 14.

The null hypothesis (H_0): In the population there is no difference between male and female heights.

The first step in data analysis is to familiarize oneself with the data. So, let's take a look at the some summary statistics.

```
 Gender     Height
 F:50   Min.   :60.0
 M:50   1st Qu.:67.0
        Median :69.0
        Mean   :69.4
        3rd Qu.:72.2
        Max.   :79.0
```

Testing Assumption of Normality

As an example, we will work through the most common options to test the normality of a dataset.

Distributions

Based on the histogram plots (Figure 15) the data seems to be relatively close to a bell-shaped distribution[96]. In this case, while the resemblance to a bell shape may be more difficult to observe, it is because of the small number of data points available for analysis. The more data points in the sample, if the sample distribution is close to the normal distribution, the closer the histogram would be to the bell-shaped curve of the normal distribution. While the visual checks are fine, let's take a closer look at sample data normality.

[96] The histograms for Males and Females may look different because of possible differences between the number of bars used to represent the data. This is due to the way R computes the number of bins to use for data representation of a data set.

qq-Plots

As Figure 16 shows, the data points form a fairly straight line for both males and females, which suggests the conclusion that the sample data for the two groups is normally distributed. The slight S-shaped curve is an indication of some departures from normality, but a visual inspection

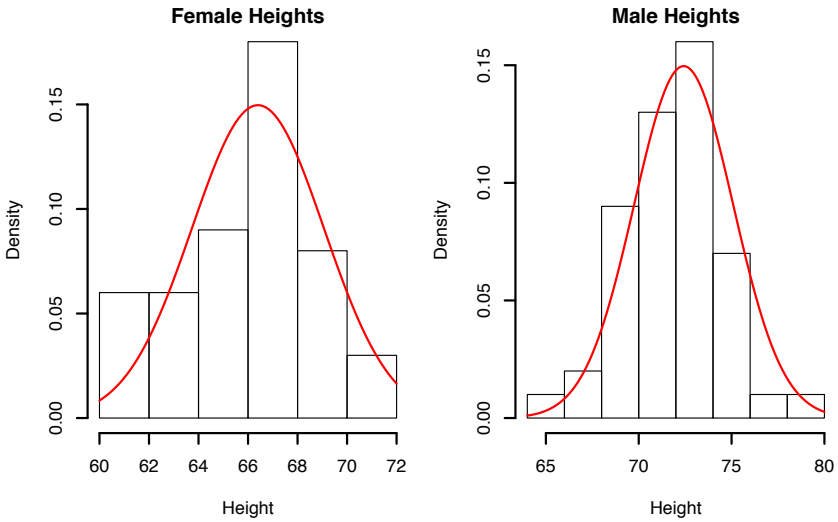

Figure 15: Histogram of male and female heights

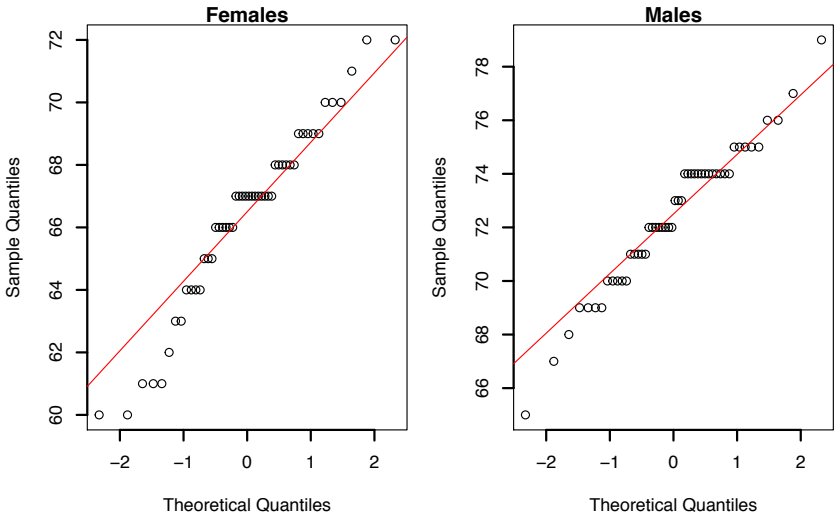

Figure 16: qq-plots of height data by gender.

	Skewness	Kurtosis
Females	-0.4309	2.828
Males	-0.3150	3.284

Table 15: Skewness and kurtosis of height data

comparing the data points curves with the red straight diagonal lines suggests that these departures are not significant. Nevertheless, to be sure, let's not rely on visual clues alone.

Skewness and Kurtosis

The data in Table 15 shows that for females the data is slightly skewed to the left and playtkurtic (kurtosis < 3). The sample of males shows a similar situation. While numbers offer a bit more detail than visual representation, let's go a step further and test data normality using a more specific statistical test.

Shapiro-Wilk Test

```
    Shapiro-Wilk normality test

data:  my.F
W = 0.96, p-value = 0.06
```

```
    Shapiro-Wilk normality test

data:  my.M
W = 0.97, p-value = 0.2
```

A *p-value* > 0.05 indicates that the null hypothesis cannot be rejected and therefore it should be concluded that the samples follow a normal distribution.

The t-Test

```
    Welch Two Sample t-test

data:  Height by Gender
t = -11, df = 97, p-value <2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.13 -4.91
sample estimates:
mean in group F mean in group M
          66.40           72.42
```

The output shows the probability p *(p-value)* that the t statistics would be this large or larger in absolute value (without the - sign) if the null hypothesis is true. For this analysis the value of p is so small[97] that there is practically no chance that the null hypothesis is true. Therefore, the conclusion is that, on average, there is a significant difference in height between males and females.

[97] The value of the *p-value* is so small that R considers it, for the purpose of this test, indistinguishable from 0 and thus represented as 2.2e-16 (0.00000000000000022). R can represent smaller numbers, but for all intents and purposes, unless one needs to work with very, very small p-values, this value can be effectively considered equal to 0.

Finally, here is an example of how to write up the results in a publication.

An independent-samples t-Test used to analyze the data revealed a significant difference in height between males and females, t(97.193) = -10.765; p < 0.001. The sample means show that males are significantly taller (mean = 72.42) than females (mean = 66.40).

Paired Samples t-Test

The paired-samples *t-Test* is used when the observation collected from one group is related in some way to the corresponding observation in the second group. Some examples of studies where this test would be appropriate:

- A study using repeated measures in which case some characteristic is measured for the same participants before and after a treatment. In this case, because the participants are the same, an observation before the treatment will have a corresponding observation after the treatment, with the correspondence being provided by the participant.
- A variant of the above is a pretest-postest study, where there is a test given to the participants before the intervention, and then another test is given after the intervention.
- A study in which the participants are assigned to the treatment groups using some type of *matching procedure or process*. That is, one participant is exposed to one of the experimental conditions while another participant, selected using a matching procedure is exposed to the other experimental treatment.
- A study in which each participant is exposed to both treatment conditions.

Assumptions underlying paired-samples t-test:

- DV - interval or ratio.
- IV - nominal with only two categories.
- Observations should be paired in some meaningful way.
- Independent observations, a participant's score in one treatment should not be affected by anther participant's score(s).
- Random sampling drawn from population.
- Normal distribution of difference scores.
- Homogeneity of variance.

Paired-samples t-tests problems:

- The designs that utilize this type of t-test are many times fairly weak.
- The experiment can have many confounding variables, such as the order of treatments for a study in which each participant is exposed to both treatment conditions, in which case it could be difficult to determine if the outcomes are because of the treatment conditions themselves, due to the order of treatments, or a combination of both.
- Repeated measures tests can also show significant confounding, especially for those that span a significant amount of time (e.g., a semester) over which things that happen outside the treatment conditions can influence the outcome.

ANOVA: Analysis of Variance

Used to compare the means of a *criterion (dependent)* variable between two or more groups defined by the *predictor (independent variable)*. The test computes an F value indicative of the ratio between the variation between groups to the variation within groups to determine if the observed differences between groups on the criterion variable represent differences between populations from which the samples are drawn (alternative hypothesis, H_1). Or, if the observed differences are due purely to chance (null hypothesis, H_0).

The analysis below will use the same data set used for the Independent Samples t-Test. Therefore, the experiment and description is the same.

Assumptions

- Sample data is normally distributed:
  - Check frequency distributions, q-q plots;
  - If the data is not normally distributed, proceed, but with caution. The F test is robust to departures from normality because the chances of Type I error are not increased by deviations from normality;
- The groups are homogeneous (have equal variances):
  - Check homogeneity of variances (Levene's test);
  - If NOT homogeneous:
    - For between-subjects variables, do not worry. F is robust and violations are not likely to increase the chance of *Type I* error significantly;
    - For within-subjects variables, ***worry***, as the results can suggest false significant effects;
- Independence of observations seems a reasonable assumptions since each participant was measured individually, independent of the others.

	Df	Sum of Squares	Mean Square	F	p (sig)
Gender	1	906.0	906.0	115.9	<.000
Residuals	98	766.2	7.8		

Table 16: ANOVA results in APA format

Testing Normality

Normality is tested similarly to the Analysis of normality presented in The t-Test chapter. Please follow the guidance provided in that section.

The analysis of normality presented in the t-Test chapter, using both visual and computation methods, suggests that despite some departures from normal distribution, the data follows a normal distribution. Therefore, the assumption of normality is verified.

Homogeneity of Variances

Let's use Levene's test to look at the homogeneity of variances.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1       0   0.96
      98
```

The test shows ($p > 0.05$) that the null hypothesis (H_0) that error variances of the criterion variable (Height) are equal across groups (Females & Males) should be accepted.

The ANOVA Test

```
            Df Sum Sq Mean Sq F value Pr(>F)
Gender       1    906     906     116 <2e-16
Residuals   98    766       8

Gender      ***
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

A significance value of $p < .001$ suggests that the null hypothesis can be rejected safely and the alternative hypothesis accepted. For reporting the output should be converted to the appropriate format for the publication venue. As an example, the American Psychological Association Manual of Style recommends the following table format (Table 16).

MANOVA: Multiple Analysis of Variance

Used to compare means of multiple *criterion (dependent)* variables between two or more groups defined by the *predictor (independent)* variable. A different data set is going to be used as an example for this analysis. To better understand the example, the study scenario is introduced below.

A research to study the effect of race stereotypical crimes was conducted with 105 participants. Three identical versions of a scenario about a crime, referencing one of three ethnicities (African-American, Hispanic, Caucasian) were randomly presented to the participants. The study measured[98]:

[98] Column labels are: *ethnicity*, *punish*, *repeats*, and *dispos*.

- Dependent variables:
  - Perceived level of punishment ("punish");
  - Perceived likelihood of committing the same crime again ("repeat");
  - Perceived likelihood that the reason the defendant committed the crime was due to his/her character, or disposition ("dispos");
- Independent variable:
  - Ethnicity[99].

[99] Labels used for the independent variable, ethnicity, are: 0 = African-American, 1 = Hispanic, and 2 = Caucasian.

Null Hypothesis (H_0): In the population, there is no difference in perceived level of punishment to be administered to the defendant, perceived likelihood the defendant will commit the same crime again, and the perceive likelihood that the reason the defendant committed the crime was due to his or her character for African-American, Hispanic, and Caucasian ethnicities.

The analysis is a *one-way MANOVA* for a *between groups design*.

Outliers

To determine whether the data set has outliers let's use a graphical representation. The primary plot represents the ordered squared ro-

bust Mahalanobis distances against the empirical distribution function MD_i^2. Of the three additional graphs the first shows the data, the second attempts to highlight potential outliers as detected using the $chisq_p$ distribution, and the third presenting the outliers by the adjusted quantile (Filzmoser, Garret, and Reimann 2005).

```
Projection to the first and second robust principal components.
Proportion of total variation (explained variance): 0.9381
```

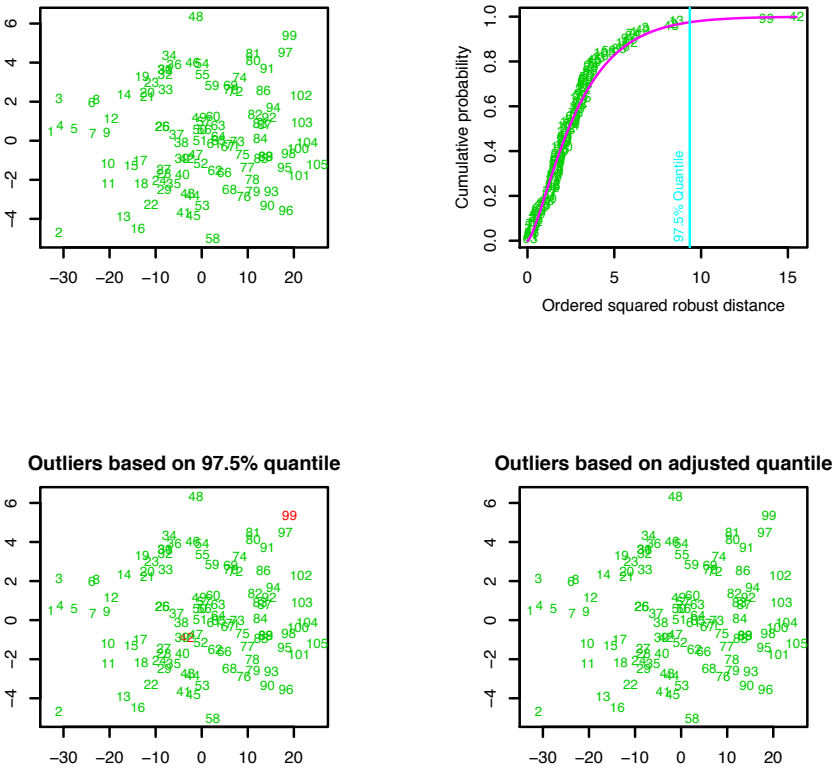

Figure 17: Outliers in the dataset

Two of plots in Figure 17 show two cases, 42 and 99, as potential outliers while the third, the outliers based on adjusted quantiles, suggests that the dataset has no outliers. In this case, the decision is left to the researcher. Each outlier should be analyzed individually and a decision should be made if it is to be removed or not. The criteria used for such decisions can range from theoretical necessity of including the possible outliers in the analysis to the point highlighted as outlier to be a valid data point based on an in-depth analysis of the data. In this example, because the analysis does not conclusively highlight the two cases as outliers, they will not be removed[100].

[100] A more extensive analysis, outside the scope of this chapter, focused on the two possible outliers conducted on the dataset led to the same conclusion.

Assumptions

Before conducting the analysis, assumptions of *normality*, *homogeneity of covariances*, *homogeneity of variances*, and *independence of observations* need to be tested.

Summary statistics including skewness and kurtosis[101].

[101] In the output, X1 refers to the dependent variable being analyzed: *punish, repeats, dispos.*

DV: *punish*

```
 Descriptive statistics by group
group: African-American
   vars  n  mean    sd median trimmed   mad
X1    1 24 49.33 11.09   50.5   49.25 13.34
   min max range skew kurtosis   se
X1  30  71    41 0.08    -1.13 2.26
---------------------------------
group: Hispanic
   vars  n  mean    sd median trimmed   mad
X1    1 26 57.77 14.71   60.5   58.64 16.31
   min max range  skew kurtosis   se
X1  28  79    51 -0.55    -0.68 2.89
---------------------------------
group: Caucasian
   vars  n  mean    sd median trimmed   mad
X1    1 55 65.89 10.54     67   65.91 11.86
   min max range  skew kurtosis   se
X1  46  85    39 -0.04    -1.13 1.42
```

The output shows various basic statistics, such as the mean of the variable, its median, mean, and so forth. Of interest here is the reported swkewness and kurtosis. The values suggest a normally distributed dataset for the *punish* variable for each of the three ethnic groups. A similar analysis is performed below for the two other variables in the analysis, *repeats* and *dispos*. In each case, the skewness and kurtosis are well within the limits.

DV: *repeats*

```
 Descriptive statistics by group
group: African-American
   vars  n  mean   sd median trimmed  mad
X1    1 24 24.12 2.17     24   24.05 2.97
   min max range skew kurtosis   se
X1  21  28     7 0.26    -1.24 0.44
---------------------------------
group: Hispanic
   vars  n  mean   sd median trimmed  mad
X1    1 26 24.23 2.25     24   24.32 2.97
   min max range skew kurtosis   se
X1  20  28     8 -0.3       -1 0.44
---------------------------------
group: Caucasian
   vars  n  mean   sd median trimmed  mad
X1    1 55 25.53 1.99     26   25.58 1.48
   min max range  skew kurtosis   se
X1  21  29     8 -0.21    -0.62 0.27
```

DV: *dispos*

```
 Descriptive statistics by group
group: African-American
   vars  n mean   sd median trimmed  mad min
X1    1 24   23 2.27     23    22.9 2.97  20
   max range skew kurtosis   se
X1  28     8 0.26    -1.03 0.46
--------------------------------
group: Hispanic
   vars  n  mean   sd median trimmed  mad
X1    1 26 23.46 2.04     24   23.45 1.48
   min max range  skew kurtosis  se
X1  20  27     7 -0.15     -1.1 0.4
--------------------------------
group: Caucasian
   vars  n  mean   sd median trimmed  mad
X1    1 55 24.91 2.47     25   25.07 2.97
   min max range  skew kurtosis   se
X1  20  29     9 -0.44    -0.84 0.33
```

Visual Evaluation of Data Normality

For analysis we produce the appropriate *qq-plots* (Fig. 18) and *histograms* (Fig. 19). A cursory review of the plots seem to suggest that departure from normality may exist, at least for some of the groups. Observe the top-left plot in Fig. 18 (level of punishment for African Americans). The points representing the data align neatly along the diagonal line, which suggests that data is normally distributed. Other plots show an S-shaped distribution of the data point alongside the diagonal, which suggests departures from normality. The histograms in Fig. 19 seem to suggest similar possible departures from normality. In this situation further analysis is necessary to determine if the assumption of normality is met.

Shapiro-Wilk Test of Normality

Shapiro-Wilks is used to test normality of data for each subgroup. Table 17 combines summary statistics and the Shapiro-Wilk test for the design groups.

Table 17 shows that most of the cells (with the exception of the last line in the table), have a low significance level that may be indicative of the fact that even if the Shapiro-Wilk test suggests that the data is within the normality conditions, its distribution in each cell is not as close to the normal distribution as one would want it to be.

Multivariate Normality

The homogeneity of covariances can be analyzed using Box's test. With an observed significance of .2 the null hypothesis, stating that the covariance matrices of the dependent variables are equal across groups, is accepted.

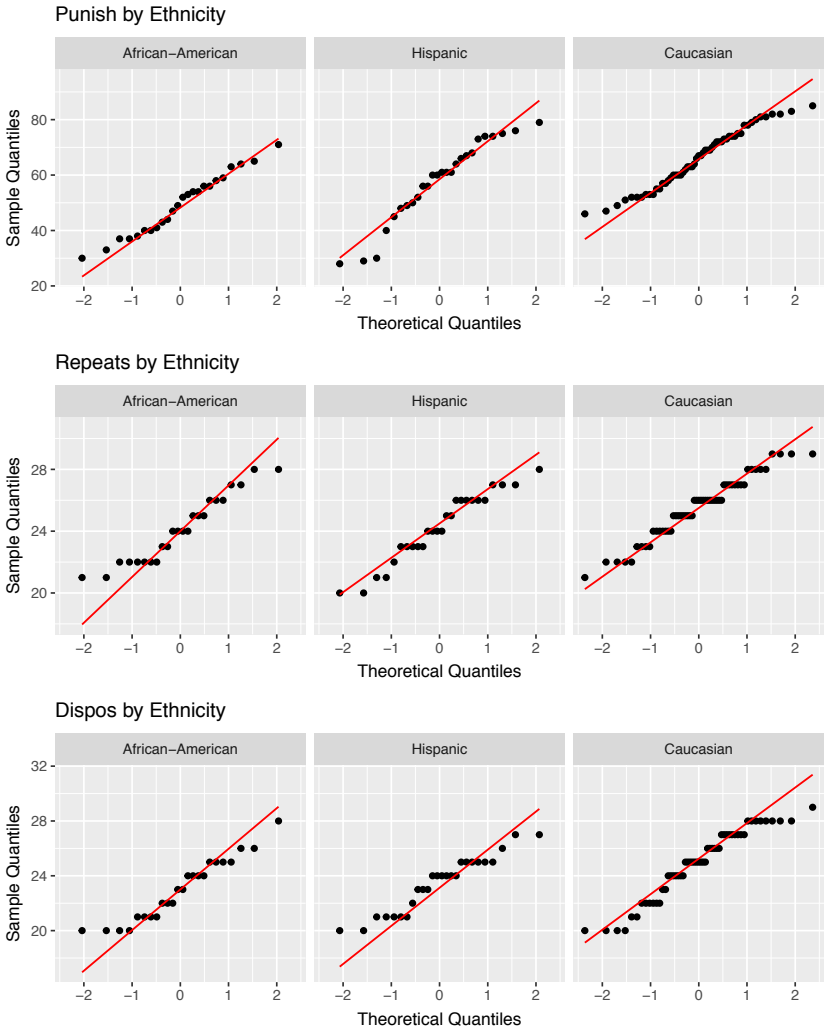

Figure 18: qq-plots

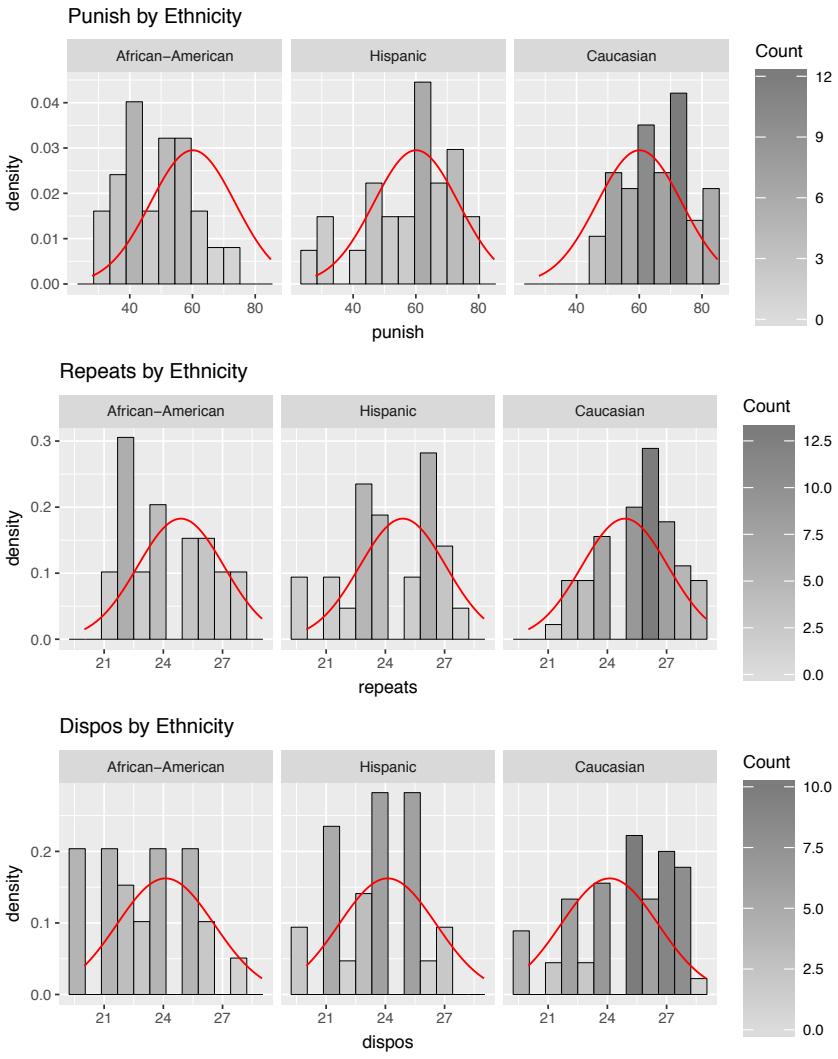

Figure 19: Histograms

Dependent Variable	Ethnicity	Statistic	df	Sig.
Level of punishment	African American	0.97	25	0.703
	Hispanic	0.94	26	0.110
	Caucasian	0.97	55	0.130
Potential to repeat	African American	0.93	25	0.101
	Hispanic	0.95	26	0.183
	Caucasian	0.96	55	0.087
Natural disposition	African American	0.95	25	0.132
	Hispanic	0.93	26	0.079
	Caucasian	0.94	55	0.005

Table 17: Shapiro-Wilk test of normality

```
    Box's M-test for Homogeneity of
    Covariance Matrices

data:  variables
Chi-Sq (approx.) = 17, df = 12, p-value
= 0.2
```

Levene's test of equality of error variances:

```
Level of Puhishment


Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2    1.14   0.32
      102


Potential to Repeat


Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2    0.55   0.58
      102


Natural Disposition


Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2    0.65   0.52
      102
```

Levene's tests of equality of variances shows that the error variance of all three dependent variables is equal across groups.

One last assumption to consider - *independence of observations* - seems reasonable because the treatments were administered individually.

The MANOVA Test

Now that the assumptions have been tested and the results show that the data set is appropriate for analysis, run the MANOVA test.

```
           Df Wilks approx F num Df den Df
ethnicity   2 0.731     5.66      6    200
Residuals 102
           Pr(>F)
ethnicity 1.9e-05 ***
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

To better understand what the results tell us let's look at the effect size (η^2). The threshold values for evaluating the effect size (partial η^2) are .01 for small effect size, .08 for moderate effect size, and .14 for large effect size. Any value below .01 indicates no effect.

```
           eta^2
ethnicity 0.1452
```

The write-up may look like below:

A significance value $p < .001$, together with F(6,200) = 5.66, $p < .05$ and a large effect size, suggested by a partial $\eta^2 = .145(> .14)$ for the multivariate tests, indicate a significant multivariate effect of ethnicity on the three dependent variables. Therefore, the null hypothesis (H_0) can be rejected.

Univariate Statistics

While the MANOVA multivariate statistics looks the dependent variables together, it may also be of interest for researchers to look how ethnicity affects the behavior of each dependent variable individually in multivariate context.

```
 Response 1 :
             Df Sum Sq Mean Sq F value
ethnicity     2   4768    2384    17.1
Residuals   102  14243     140
            Pr(>F)
ethnicity    4e-07 ***
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

 Response 2 :
             Df Sum Sq Mean Sq F value
ethnicity     2     48    23.8    5.42
Residuals   102    449     4.4
            Pr(>F)
ethnicity   0.0058 **
Residuals
```

```
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

 Response 3 :
            Df Sum Sq Mean Sq F value
ethnicity    2     76    37.8    6.97
Residuals  102    553     5.4
           Pr(>F)
ethnicity  0.0014 **
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1


Level of Punishment


          Partial eta^2
ethnicity        0.2508
Residuals            NA


Potential to Repeat


          Partial eta^2
ethnicity       0.09601
Residuals            NA


Natural Disposition


          Partial eta^2
ethnicity        0.1203
Residuals            NA
```

The univariate statistics analysis show that ethnicity has univariate effects on all three dependent variables as the significance p values for all three dependent variables are < .01. For effect size, ethnicity shows a large effect size on the Level of punishment (punish) (partial η^2 *= .251), and moderate effect size on the other two dependent variables, Potential to repeat (repeat) and Natural disposition (dispos), which have values of partial* $\eta^2 = .096$ *and .12 respectively.*

Multiple Linear Regression Analysis

Linear Regression is used to study/describe the predictive relationship between *one criterion (DV) variable* and *one or more predictor (IV) variable(s)*. The procedure can be used to either explore the predictive relationships between a set of variables or to test a causal model. It can be applied when:

- The dependent variable is continuous;
- The independent variable(s) is continuous, discrete or categorical.

It is a highly flexible procedure, especially helpful in non-experimental research in the social sciences field where researchers often deal with naturally occurring variables[102]. The procedure allows researchers to determine if a given set of variables (predictors) is useful in predicting a criterion variable.

[102] Variables measured as they naturally occur and are not manipulated in any way by the researchers.

The intent of the linear regression analysis is to find the best fitting line of a set of data points. Remember, the simple regression equation is of the form:

$$Y = a + bX$$

It is the equation representing a straight line in the two dimension plane. In this equation, Y is the *criterion* variable (the response), X is the *predictor* variable, a is the intercept[103], and b is the slope of the line.

[103] The value of y when $X = 0$, or the point where the line intersects the Y (vertical) axis.

Multiple linear regression expands on the *simple linear regression* model and attempts to account for how *multiple predictors* contribute to the value of the *criterion* variable. In this case the question has the form:

$$Y = a + b_1X + b_2X + ... + b_nX$$

The question is still one of straight line[104]. The b_i coefficients repre-

[104] Using algebra, we can rewrite the equation as:

$$Y = a + (b_1 + b_2 + ... + b_n)X$$

which is similar to the previous equation ($b = b_1 + b_2 + ... + b_n$)

sent the contributions of each individual predictor.

Multiple regression allows researchers to:

- Explore the relationship between the *criterion variable* and the *predictor variables* as a group and determine whether the relationships is statistically significant;
- Determine how much of the variance of the *criterion variable* can be accounted for by the *predictors*, individually and as a group;
- Determine the relative importance of the *predictor variables*.

Of note is the fact that regression analysis cannot provide strong evidence of a causal relationship between the variables but rather show that the findings are or are not *consistent* with the causal model under study[105]. In case of modeling causal relations, if none of the regression coefficients were found to be significant, the model failed to pass a test[106].

When designing the study researchers may not have sufficient knowledge to determine how the *predictors* affect the *criterion* variable. The model they build is their best guess based on their prior knowledge[107] of the problem they are investigating. For this reason, to find the model that best represents the phenomena, regression analysis offers multiple pathways for conducting the analysis. This example looks at the *enter*, *forward*, and *backward* regression models.

The *enter* model includes all predictors while the stepwise regression *forward* and *backwards* models attempt to reach a more parsimonious (see Occam's Razor for details) model by reducing the number of predictor variables while attempting to explain as much as possible of the criterion variable. The differences in the quality of the predicted scores can be studied by looking at the three models side by side. A comparative analysis of the forward and backward models will offer the opportunity to validate the results of the analysis when the two methods reach the same solution (regression equation) or if they are very close.

The *enter* model, as its name suggests, will include all predictors in the regression equation from the beginning. The *forward* model will enter the variables in the model one by one, starting with the predictor that has the highest order partial correlation with the criterion variable, and then calculating the regression statistics. The process continues by adding the remaining predictor variables, one by one, in descending order of their partial correlations with the dependent variable. This continues as long as the addition of a new predictor variable to the model brings *significant*[108] changes in the model's predictive power[109]. The backwards model starts with the full model[110] and continues by

[105] The correlational nature of the data leads to multiple possible ways of interpreting the results.

[106] Alternatively, if significance was observed, the causal model passed an attempt at disconfirmation.

[107] Based on existing literature, prior experience, other experiments they have conducted, etc.

[108] The goodness of fit of an estimated statistical model is measured by either the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). The lower the value or AIC or BIC is, the better the model is. When the AIC/BIC of two successive models are compared and found significantly different, the model with the better (lower) values is retaind.

[109] The predictive power of the model is reflected in the value of R^2.

[110] The same model as the *enter* model

removing predictor variables one by one, in the reverse order of their partial correlation with the criterion variable. The process continues until the updated model does not improve on the previous one.

Prior to the analysis we should familiarize ourselves with the data. Here are a few things to do to better understand what we are working with.

- Check for errors[111];
- Outliers;
- Unusual distributions;
- Clustering or systematic patterns[112];
- Anything else unexpected.

[111] E.g., data entry, missing values, etc.

[112] If the data groups together in certain ways, forming areas where data is concentrated.

Assumptions

In regression analysis assumptions are intertwined with the analysis itself. That is, some of the assumptions can only be verified once the model has been defined and after the analysis has been run and the results computed. Regression analysis assumptions are:

- *Linear relationship between the predictor (IV) variables and the criterion (DV) variable*
  - Detected using plots of the residuals[113] against each continuous predictor and predictive values;
  - The assumption is violated when large or systematic deviations of the fit line are observed around the 0-line[114]. A violation is indicative of a non-linear relationship.
- *Correct model specification*
  - Detected by performing the R Squared Change test to determine if adding a variable (given that the theoretical model supports it) to the regression model significantly increases R Square.
  - If the R Square Change test is significant, the variable should be included in the regression model. Otherwise, if the test is not significant, do not include the variable.
- *No measurement error in the predictor (IV) variables*
  - Detected by examining the reliability coefficients for the predictor (IV) variables;
  - Inadequate reliability if the coefficients are less than .70.
- *Homoscedasticity, residuals have constant variance*

[113] The *residual* is the difference between the observed value of the dependent variable and the value predicted by the regression model. The mean of residuals is zero, as is their sum. Residuals help us understand how well the regression line (equation) approximates the data from which it was generated.

[114] Representing the mean of the residuals.

  - Detected by plotting the residuals against each continuous predictor and predicted values;
  - Relationship between variability of the residuals and either the predictor (IV) variables or the predicted values indicate heteroscedasticity.

- *Residuals are independent, reflecting a clustering or serial dependency problem*
  - Clustering is detected by using plots of residuals against the grouping/cluster variable using Box Plots;
    * The clustering assumption is violated if there is variability in the median value of the residuals in each group;
  - Serial dependency is detected by looking at the measure of autocorrelation using the Durbin-Watson test;
    * The serial dependency assumption is violated if the Durbin-Watson test results in values less then or greater than 2.

- *Normal distribution of residuals*[115]
  - Detected by using normal probability plots and histogram plots of residuals;
  - Non-normal distribution of residuals indicate a violation;
  - Residuals that fall far from the straight line indicate a violation.

[115] In case of linear regression, a plot of residuals should look as a random cloud of points alongside the regression line. If a pattern is observed in how the residuals are arranged (e.g., they seem to align along a curve), it may be an indication that a non-linear model may be a better fit.

The Study

A study was conducted in a Midwestern state in the US to gauge consumers' interest in purchasing and consuming the different kinds of nuts[116] available on the market. The primary focus of the study was to attempt to predict future market behavior based on consumption of, familiarity with, and interest in the product. The data was collected using a survey-type instrument. With the exception of demographic data, the questions used 5-point Likert scales.

[116] Chestnuts, pecans, and black walnuts.

For the purpose of this analysis, three variables, defined from the raw dataset, were chosen:

- Interest in buying raw nuts from farmers markets or grocery stores;
- Interest in buying prepared/semi prepared products that contain nuts at farmers markets or from grocery stores;
- Interest in consuming, in restaurants, prepared food that contains nuts;

The questionnaire was freely distributed during a local festival. A number of 232 responses were collected.

Preliminary Steps

For all quantitative studies is always helpful to familiarize yourself with the data, how it looks, etc. So, let's look at the first few lines of the dataset[117].

[117] Due to the large number of variables, only the firs 10 are shown.

Table 18: First few rows and variables of the chestnut data set.

D1.1	D2.1	D3.1	F1.1	F1.2	F1.3	F1.4	F1.5	F1.6	F2.1
3	4	4	3	4	4	5	4	3	3
2	3	2	3	5	5	2	2	4	4
3	5	5	1	1	1	1	1	1	3
2	2	1	2	3	4	3	3	2	2
1	1	1	2	5	5	4	5	5	2
1	1	1	2	2	2	1	1	1	3

Missing Cases

Due to the large sample available, for the purpose of this example, the cases with missing variables were deleted list wise.

Variable Recoding

Table 19: Variables included in the analysis

Name	Type	Value Range	Description
BuyRaw	DV	3-15	Interest in buying raw nuts
Price	IV	3-15	How much the product's price influences purchase/consumption decision
Quality	IV	3-15	How much the product's quality influences purchase/consumption decision
Taste	IV	3-15	How much the product's taste influences purchase/consumption decision
LocGrown	IV	3-15	How much the fact that the product is locally grown influences purchase/consumption decision
PrepEase	IV	3-15	How much product's ease of preparation influences purchase/consumption decision
Nutrition	IV	3-15	How much product's nutrition factor influences purchase/consumption decision

The survey was designed to enable a detailed study of the market, for which reason the questions used to collect the data were divided between the three categories of products: chestnuts, pecans, and black walnuts. To capture an overall view of the market the three categories are merged into a single variable. This example analysis will look at the following variables (Table 19[118]):

The new variables described in Table 19 are computed, using an addition-based model, based on the participants' responses to survey

[118] The value of a variable for all the variables listed in Table 19 is determined by summing the scores of three questions together. Therefore, if the Likert scale used is from 1 to 5, the minimum value is 3 and the maximum is 15.

questions (Table 20). Table 21 shows a few lines showing the newly created variables.

Variable	Component Questions
BuyRaw	D1.1 + D2.1 + D3.1
Price	F1.1 + F2.1 + F3.1
Quality	F1.2 + F2.2 + F3.2
Taste	F1.3 + F2.3 + F3.3
LocGrown	F1.4 + F2.4 + F3.4
PrepEase	F1.5 + F2.5 + F3.5
Nutrition	F1.6 + F2.6 + F3.6

Table 20: Variable recoding

BuyRaw	Price	Quality	Taste	LocGrown	PrepEase	Nutrition
11	9	12	12	15	10	9
7	11	13	13	6	6	12
13	9	11	11	11	11	11
5	6	9	13	9	9	6
3	6	15	14	13	15	15
3	6	6	6	3	3	5

Table 21: Recoded variables

Summary Statistics

As a first attempt to discover any possible abnormalities in the data it is always helpful to run summary statistics on the variables in the dataset.

A cursory analysis of the data in Table 22 shows that there are no unexpected problems with the data. For example, for all variables the number of included cases is the same (184), meaning that there are no missing cases, the *min (3)* and *max (15)* represents the correct range determined by the way the variables were computed (Tables 20 and 21).

Before moving forward we need to define the full regression model as the decision to buy as a function of price, quality, taste, growth place, ease of preparation, and nutrition qualities. It can be represented as a linear relationship of the study's predictors:

$$BuyRaw' = b_0 + b_1 \cdot Price + b_2 \cdot Quality + b_3 \cdot Taste$$
$$+b_4 \cdot LocGrown + b_5 \cdot PrepEase + b_6 \cdot Nutrition$$

Outliers

Let's look at residuals plot first (Figure 20). In this representation an outlier would be a point situated outside the ± 3 interval[119] represented on the vertical (y) axis. According to this test, only one of the cases is

[119] Colored in red or highlighted.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
BuyRaw	1	184	9.109	3.120	9.0	9.122	2.965	3	15	12	-0.1058	-0.5286	0.2300
Price	2	184	9.277	2.591	9.0	9.324	2.965	3	15	12	-0.1676	0.0790	0.1910
Quality	3	184	11.207	2.929	12.0	11.507	2.965	3	15	12	-0.8162	0.3884	0.2159
Taste	4	184	11.636	2.932	12.0	11.986	2.965	3	15	12	-0.8755	0.4133	0.2161
LocGrown	5	184	9.924	3.245	10.0	10.040	2.965	3	15	12	-0.3064	-0.5388	0.2392
PrepEase	6	184	10.114	3.034	10.0	10.250	2.965	3	15	12	-0.4631	-0.2139	0.2237
Nutrition	7	184	10.663	3.478	11.5	10.946	3.707	3	15	12	-0.4941	-0.6215	0.2564

Table 22: Summary statistics for the study variables

situated close to the -3 limit and could be considered an outlier. Because there are other cases that are relatively close to the limits and because the sample size is sufficiently large, further analysis should be considered to determine if the data point is a real outlier or not. For this purpose further testing will use Mahalanobis distances[120].

The histogram in Figure 21 suggests that there is at least one outlier in the dataset, indicated by the highlighted bar at the far right of the histogram. While the analysis so far tells us that outliers may exist in the data set, it is not able to help us determine what impact these outliers may have[121]. Computing Cooks distance[122] will provide the means for a closer analysis of potential outliers. Figure 22 presents a graphical representation with possible outliers highlighted.

Based on Cook's distance, the dataset may have three outliers (observations 5, 26, and 110). Considering what we've learned so far about the outliers in the dataset, we have two options:

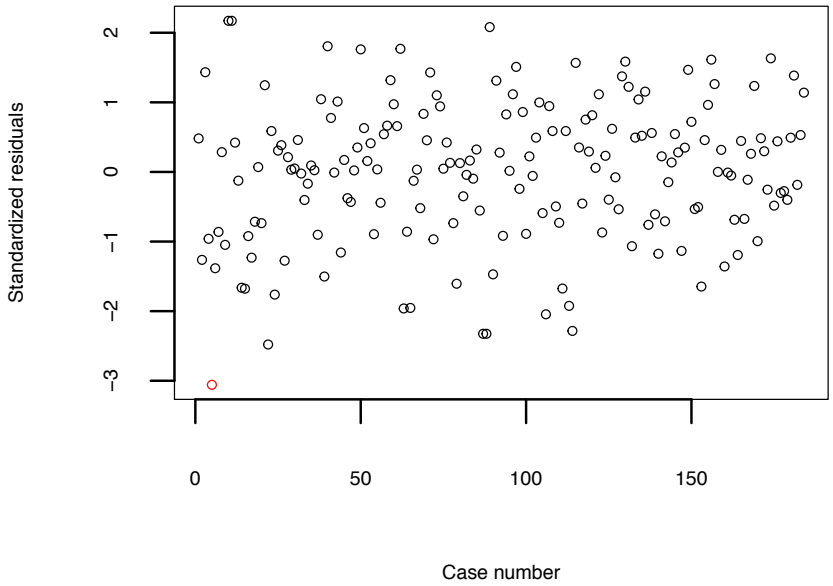

Figure 20: Case-wise plot of standardized residuals

1. Remove the outliers from the dataset and re-run the outlier analysis

[120] Mahalanobis distance (MD) is the distance between two points in multivariate space. It measures the distance relative to a base or central point considered as an overall mean for multivariate data (*centroid*). The *centroid* is a point in multivariate space where the means of all variables intersect.

[121] Cases that may look like outliers at first sight may not be so after a more in-depth analysis.

[122] Cook's distance is a measure used to estimate the influence of a data point in regression analysis. It measures how much deleting an observation influences the results. Data points with a large value of the computed Cook's distance are potential outliers and should be subjected to further examination.

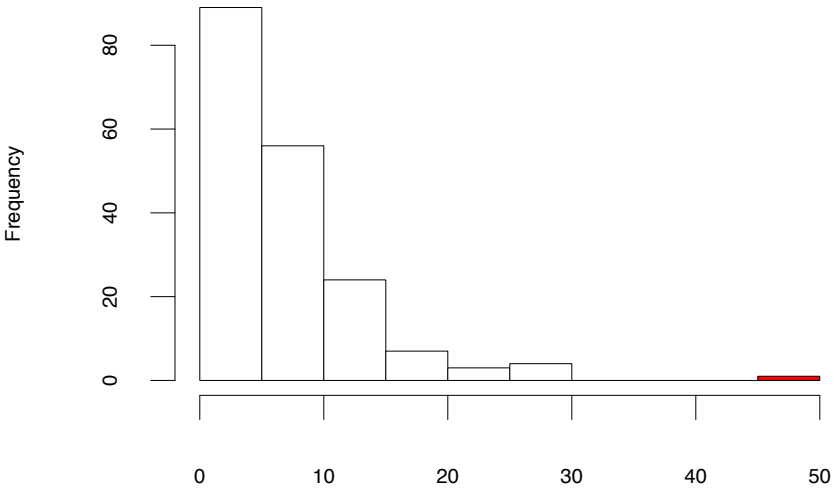

Figure 21: Histogram of Mahalanobis distances

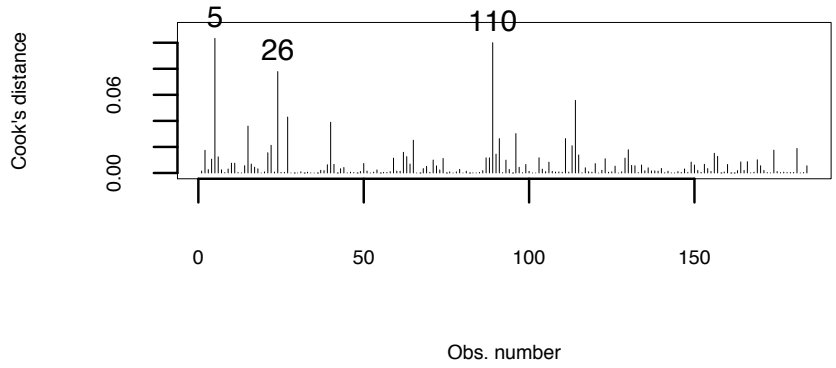

Figure 22: Cook's distance

to determine if the new dataset has more;

2. Investigate these outliers further and learn more about their influence and determine what potential leverage they may have on the results of the analysis.

If we have sufficient data points for the analysis[123], the entire observation or record including the outliers can be removed from the data set. Alternatively, if the removal of outliers reduces sample size too much or if the analysis begins with an undersized sample, the outliers should be analyzed further to determine if they should indeed be removed or if they can be retained thus improving the relevance of the findings to the population[124].

With this study being used as an example, we will pursue the second option and look further into what leverage outliers may have. For that

[123] If the remaining number of records after removing the outliers is larger than the sample size needed for analysis. The sample size is one of the deciding factors in how well the results can be generalized to the population the model attempts to represent.

[124] Removing observations from the dataset may have unwanted effects. For example, when studying opinions, by removing a record showing an extreme, we may unintentionally remove an opinion that may be relevant to the results and affect the outcomes. That is, outliers may be influential observations that have a reason to be kept in the analysis. In addition to numerical analysis, the theory, literature, and previous studies should be used as well to guide the decision.

purpose we can use a scatter plot representation of deleted vs. raw residuals (Figure 23). If in the graphical representation all the points, and especially those highlighted so far as outliers, align along the diagonal of the plot area, represented as a line, it can be concluded that none of the points representing potential outliers has a significant leverage and that it may be reasonable to keep them in the dataset.

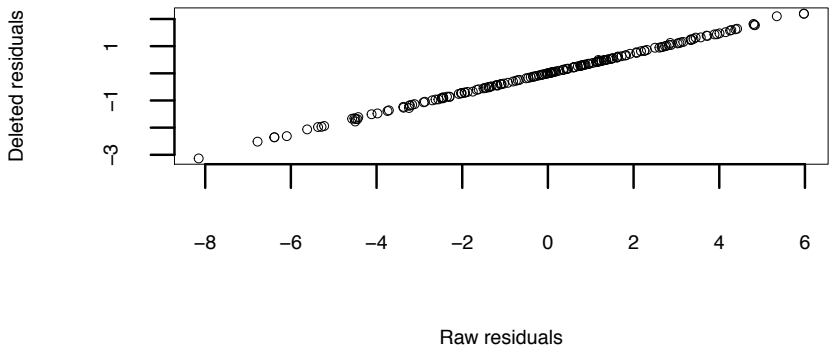

Figure 23: Deleted vs. raw residuals plot

While prior analysis may indicate the existence of potential outliers, leverage analysis suggests their influence is sufficiently weak to warrant including the observations in the analysis.

Residuals Analysis

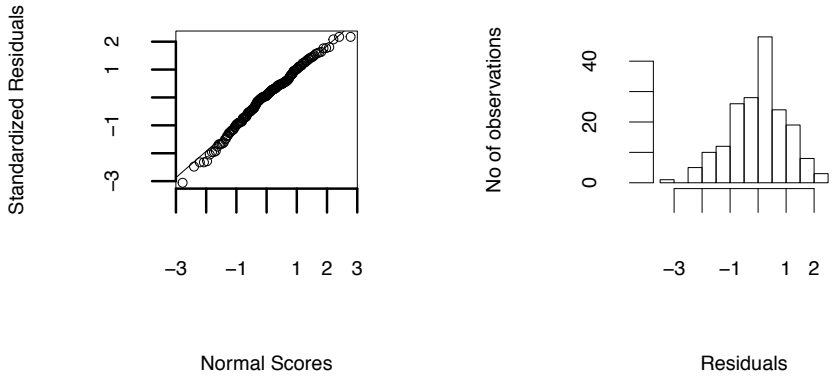

Figure 24: Normal probability plots

Multiple regression analysis assumes linear relationships between the variables included in the equation and the normal distribution of residuals. When these assumptions are violated, the final conclusion drawn may not be accurate. Normal probability plots are used to test these assumptions (Figure 24 left). The grouping of the points close to the straight line[125] supports the assumption that the residuals are normally

[125] The representation is a quantile-quantile plot and is interpreted similarly.

distributed. The same assumption is also supported by the histogram representation of the residuals (Figure 24 right).

Linearity

A simple residuals vs fitted plot can be used to test the linearity assumption (Figure 25). If linearity is observed, the data points plotted on the graph should resemble a homogeneous cloud around the center line, which should be approximately horizontal at zero. Alternatively, a pattern is indicative of a problem with the linear model. In this analysis, the distribution of the points and the center line (red) suggest that a linear relationship between the criterion and predictor variables can be assumed.

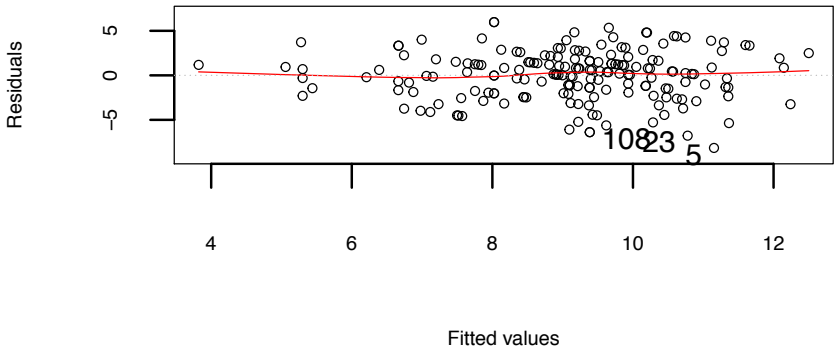

Figure 25: Residuals vs. fitted linearity plot

Enter or Full Model

Up to this point the preliminary analysis of the assumptions has been conducted using the full regression model, which includes all predictor variables. As a reminder, the regression equation for the enter model is[126]:

$$BuyRaw' = b_0 + b_1 \cdot Price + b_2 \cdot Quality + b_3 \cdot Taste + b_4 \cdot LocGrown + b_5 \cdot PrepEase + b_6 \cdot Nutrition$$

In regression analysis the correlations should be high between the criterion and predictor variables and low between predictor variables. There are multiple ways to look at the correlations between the predictor variables and a wide variety of graphical representations to help the analysis. For example, a visualization of the variable correlations is shown in Figure 26.

[126] *BuyRaw'* is the estimated value of the criterion variable.

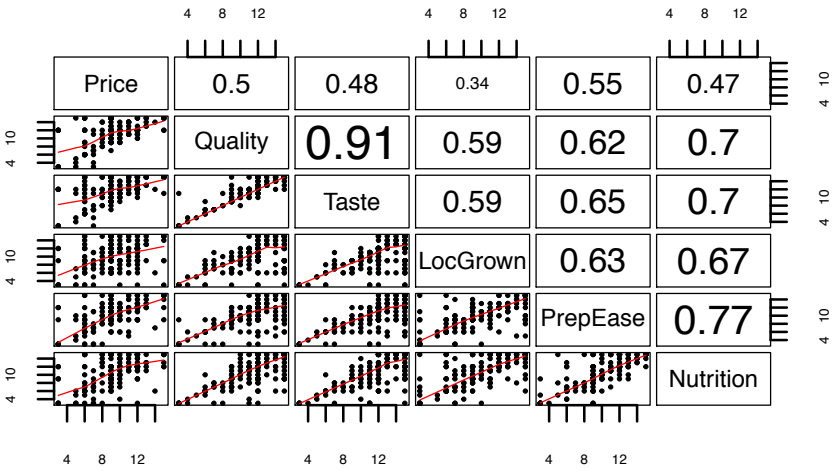

Figure 26: Predictors correlation matrix

Figure 26 shows the predictors on the diagonal, the value of the correlation coefficient between each pair of two predictors above the diagonal, with larger font for higher correlations, and a scatter plot of each pair below the diagonal. The numerical values of the correlation coefficients are fairly easy to interpret. The closer the value is to 1, the higher the correlation is. The closer the value is to 0, the lower the correlation is. Therefore, we are looking to have low values across the board. In this case, *quality* is highly correlated to *taste*, as indicated by the value of correlations coefficient of 0.91.

Looking at the scatter plots below the diagonal, we want to see the data points distributed across the plot area rather than forming a pattern. Looking at the same two variables as discussed above, the scatterplot representation shows the points arranged close to and alongside the red diagonal line, forming a diagonal pattern. Other scatter plots may look more distributed, but a closer inspection shows that in some of them the many of the points are concentrated in a pattern along the red diagonal line.

A table representation of the correlations coefficients between the variables in the study (Table 23) may be more helpful, especially when it is associated with a table showing the significance levels (Table 24) of these correlations.

The data in Table 23 suggests that the higher correlations between the criterion variable BuyRaw and any of the predictor variables is with Quality, at 0.47.

Looking at the correlations between the predictor variables (Table 23) the the associated p-values[127] (Table 24) it appears that this data set has a problem. The correlations between predictors are relatively high,

[127] Significance levels of correlations

	BuyRaw	Price	Quality	Taste	LocGrown	PrepEase	Nutrition
BuyRaw	1.00	0.18	0.47	0.41	0.33	0.25	0.38
Price	0.18	1.00	0.50	0.48	0.34	0.55	0.47
Quality	0.47	0.50	1.00	0.91	0.59	0.62	0.70
Taste	0.41	0.48	0.91	1.00	0.59	0.65	0.70
LocGrown	0.33	0.34	0.59	0.59	1.00	0.63	0.67
PrepEase	0.25	0.55	0.62	0.65	0.63	1.00	0.77
Nutrition	0.38	0.47	0.70	0.70	0.67	0.77	1.00

Table 23: Variable correlations

	BuyRaw	Price	Quality	Taste	LocGrown	PrepEase	Nutrition
BuyRaw	NA	0.015	0	0	0	0.001	0
Price	0.015	NA	0	0	0	0.000	0
Quality	0.000	0.000	NA	0	0	0.000	0
Taste	0.000	0.000	0	NA	0	0.000	0
LocGrown	0.000	0.000	0	0	NA	0.000	0
PrepEase	0.001	0.000	0	0	0	NA	0
Nutrition	0.000	0.000	0	0	0	0.000	NA

Table 24: Significance levels of variable correlations

for which reason their contribution is overlapping.

It has become evident at this point that the dataset has problems. Nevertheless, the analysis can still be conducted to inform next steps. With this knowledge, let's look at the results of the regression analysis.

```
Call:
lm(formula = BuyRaw ~ Price + Quality + Taste + LocGrown + PrepEase +
    Nutrition, data = my.chestnut)

Residuals:
   Min     1Q Median     3Q    Max
-8.152 -1.784  0.176  1.620  5.976

Coefficients:
            Estimate Standardized Std. Error
(Intercept)   3.9396       0.0000     0.9405
Price        -0.0607      -0.0504     0.0980
Quality       0.4736       0.4446     0.1754
Taste        -0.0404      -0.0379     0.1737
LocGrown      0.0909       0.0945     0.0893
PrepEase     -0.1725      -0.1677     0.1179
Nutrition     0.1629       0.1816     0.1081
            t value Pr(>|t|)
(Intercept)    4.19  4.4e-05 ***
Price         -0.62   0.5363
Quality        2.70   0.0076 **
Taste         -0.23   0.8165
LocGrown       1.02   0.3102
PrepEase      -1.46   0.1453
Nutrition      1.51   0.1336
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 2.77 on 177 degrees of freedom
Multiple R-squared:  0.24,  Adjusted R-squared:  0.214
F-statistic: 9.29 on 6 and 177 DF,  p-value: 7.43e-09
```

The explanatory power of this regression model is indicated the by the value of the Adjusted $R^2 = 0.214$[128] which tell us that the current model accounts for 21.4% of the variance of the criterion variable. The regression coefficients[129] can be found in the column labeled Estimate. Using these coefficients, the regression equation is:

$$BuyRaw' = 3.9396 - 0.0607 \cdot Price + 0.4736 \cdot Quality - 0.0404 \cdot Taste$$
$$+0.0909 \cdot LocGrown - 0.1725 \cdot PrepEase + 0.1629 \cdot Nutrition$$

Standardized β (beta) coefficients (the *Standardized* column in the output) are an indication of the relative importance of the respective predictor variable in predicting the value of the criterion variable. To help us understand which of the regression coefficients may be relevant, *t-Tests* are run behind the scenes to determine the significance[130] of each b coefficient in the regression equation. In this example the last column of the output[131] shows that only the *Quality* predictor variable is significant at a level of significance of 0.05.

The last line of the regression analysis output shows the results of the ANOVA test if the model has a significant explanatory power[132]. In this case the model with a *p-value* < 0.05 the model is significant. A full ANOVA table for the regression model may help understand it a bit further.

```
Analysis of Variance Table

Response: BuyRaw
            Df Sum Sq Mean Sq F value  Pr(>F)
Price        1     57      57    7.49  0.0068
Quality      1    335     335   43.73 4.3e-10
Taste        1      0       0    0.05  0.8312
LocGrown     1     12      12    1.52  0.2191
PrepEase     1      5       5    0.71  0.4008
Nutrition    1     17      17    2.27  0.1336
Residuals  177   1355       8

Price     **
Quality   ***
Taste
LocGrown
PrepEase
Nutrition
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

An analysis of the full ANOVA table[133] suggests Price as being a significant predictor besides Quality. Looking at the last column of the table, the *p-values* for each of the predictor variables are different from those listed in the regression analysis output. The differences are indicative of overlap between the predictors which is an indication that the variables are not perfectly uncorrelated. The largest differences are

[128] The value is computed together with the other parameters and made available in the output. In this case (R output) it is in the notes at the bottom, the line titled "Multiple R-squared".

[129] The values that represent the b coefficients in the regression equation.

[130] If the coefficient is significantly different than 0.

[131] The column titled Pr(>|t|) lists the significance (*p-value*) of the coefficients. The asterisks indicate the level of significance based on legends shown in the*Signif. codes* line below the table.

[132] The ANVOA test is run to determine if the null hypothesis holding that all coefficients are 0 holds or not. If this null hypothesis can be safely rejected ($p < 0.05$), then the regression model defined in the analysis has a significant explanatory power.

[133] As a reminder, *total variance* is the sum of two variances: the variance due to the predictors (model) and the variance due to error. In the ANOVA output it is represented by the sum of squares (column "Sum Sq"). The value for each variable indicates the variance accounted for by the respective predictor. The "Residuals" line at the bottom lists the unexplained (error) variance.

observed for the two variables in question[134]. Considering that the standardized coefficient of the Quality predictor is significantly larger than the one of the Price predictor, Quality is probably more relevant than price in this analysis.

[134] Common sense tells us that usually quality is correlated with price in the sense that higher quality usually demands higher prices while higher prices many not necessarily mean higher quality.

The *variance inflation factor (VIF)*[135] can be used to study predictor multicollinearity[136] among predictors. The higher the value of the *VIF* is, the more significant the collinearity. The *VIF* values for the current model are listed below.

[135] When multicollinearity exists, the coefficients of the regression equation are inflated. *VIF* is a measure of how much the variance is inflated due to multicollinearity.

[136] Multicollinearity happens when two or more predictors are correlated with each other. It occurs in many studies, especially when the researchers only observe a process and do not have control over the variables. Significant multicollinearity can also be an indication of potential issues with the instrument used to collect the data.

```
   Price   Quality     Taste  LocGrown
   1.542     6.310     6.201     2.007
PrepEase Nutrition
   3.059     3.379
```

There are multiple ways to interpret the results in this output One way would be to analyze the raw *VIF* values and flag those with values greater then 10. The output indicates that all model predictors respect the collinearity assumption.

Another way is to compute square root of the *VIF* values and flag those with values greater then 2 as having collinearity issues (output of running this computation is included below). The output below flags *Quality* and *Taste* as being collinear.

```
   Price   Quality     Taste  LocGrown
   FALSE      TRUE      TRUE     FALSE
PrepEase Nutrition
   FALSE     FALSE
```

Given that there is no agreement in the analysis, alternative models[137], discussed in the next sections, could be helpful in bringing some clarity to these issues.

[137] For example, *forward* and *backward* stepwise regression models.

With a 21.4% explanatory power this model is not capable of explaining much of the criterion variable variability using a linear combination of all predictor variable even though the model is significant overall.

As the model output suggests any prediction made using this model will be relatively weak. Further analysis suggests that the model's predictors share a lot of explanatory power with overlapping and collinearity being the main issues. Additional models can be useful in understanding if different, more parsimonious[138] models, can explain a larger percentage of the variance of the criterion variable.

[138] With fewer predictors.

Backward Model

As an example, let's run a backward stepwise regression to see if a more parsimonious model exists and how that model might look like. In

backward stepwise regression the analysis starts with the most complex model, which includes all predictors. It then starts to build new models by removing predictors in reverse order of their partial correlations with the criterion variable. This process continues until a new model brings no significant improvement in the model's explanatory power[139]. The predictors are being considered one by one to determine their effect when removed from the regression equation. The predictor who's deletion brings the smallest reduction in R^2 is the candidate for removal in the next iteration..

[139] The goodness of fit of an estimated statistical model is measured by either the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). The lower the value or AIC or BIC is, the better the model is. Therefore, stepwise regression will compute a value for the AIC or BIC for each successive regression model and compare it with the one computed for the preceding model. If the value is significantly lower, meaning a significantly better fit of the model, the model is retained. Otherwise, the analysis is stopped and the last retained model is considered to be the best fit.

```
Start:  AIC=381.4
BuyRaw ~ Price + Quality + Taste + LocGrown + PrepEase + Nutrition

            Df Sum of Sq  RSS AIC
- Taste      1       0.4 1355 379
- Price      1       2.9 1358 380
- LocGrown   1       7.9 1363 380
<none>                   1355 381
- PrepEase   1      16.4 1371 382
- Nutrition  1      17.4 1372 382
- Quality    1      55.8 1411 387

Step:  AIC=379.4
BuyRaw ~ Price + Quality + LocGrown + PrepEase + Nutrition

            Df Sum of Sq  RSS AIC
- Price      1       2.9 1358 378
- LocGrown   1       7.8 1363 378
<none>                   1355 379
- Nutrition  1      17.3 1373 380
- PrepEase   1      17.9 1373 380
- Quality    1     135.8 1491 395

Step:  AIC=377.8
BuyRaw ~ Quality + LocGrown + PrepEase + Nutrition

            Df Sum of Sq  RSS AIC
- LocGrown   1       8.8 1367 377
<none>                   1358 378
- Nutrition  1      17.2 1376 378
- PrepEase   1      25.3 1384 379
- Quality    1     134.2 1492 393

Step:  AIC=377
BuyRaw ~ Quality + PrepEase + Nutrition

            Df Sum of Sq  RSS AIC
<none>                   1367 377
- PrepEase   1      20.5 1388 378
- Nutrition  1      26.2 1393 379
- Quality    1     154.2 1521 395



Call:
lm(formula = BuyRaw ~ Quality + PrepEase + Nutrition, data = my.chestnut)

Residuals:
   Min     1Q Median     3Q    Max
-7.779 -1.819  0.221  1.842  6.002

Coefficients:
            Estimate Std. Error t value
(Intercept)   3.8264     0.8458    4.52
Quality       0.4470     0.0992    4.51
PrepEase     -0.1764     0.1074   -1.64
Nutrition     0.1929     0.1038    1.86
            Pr(>|t|)
(Intercept)  1.1e-05 ***
Quality      1.2e-05 ***
```

```
PrepEase       0.102
Nutrition      0.065 .
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

Residual standard error: 2.76 on 180 degrees of freedom
Multiple R-squared:  0.233, Adjusted R-squared:  0.22
F-statistic: 18.2 on 3 and 180 DF,  p-value: 2.32e-10
```

The output above shows a trace of the successive models considered in the analysis. It starts with the full model[140] and continues to remove the predictor with the smallest contribution to the variance until a more parsimonious model does not improve its predictive power significantly. The criteria used is AIC, listed at the top of each iteration. In this case, by using *backward stepwise regression* we were able to slightly improve the predictive power of the model by using only three of the predictors: *Quality*, *PrepEase*, and *Nutrition.* Model summary shows that the predictive power increased to 22% (from 21.4% of the full model).

[140] The predictors are sorted in ascending order of their contribution to the total variance, represented by the "Sum of Sq" column.

Summary

The study exemplified here was designed as an exploratory study. For that reason any outcome was expected despite the fact that the researchers believed, based on existing literature, that the six predictor variables would be fairly reliable in explaining consumers' level of interest in buying raw products.

The multiple linear regression analysis has produced weak models, with at most 22% explanatory power. During the analysis it became evident that predictor collinearity and overlapping seems to be a significant issue with this analysis. Stepwise regression was able to offer an improved model, although the 0.8% increase in explanatory power is too small to be meaningful in real life.

The results of the analysis seem to suggest that the study's design was flawed and that the current data does not have sufficient potential to explain market behavior and cannot be used as predictor for future behavior. Nevertheless, it has provided the foundation for further studies using a revised data collection instrument.

Multiple Logistic Regression

Logistic regression analysis responds to the needs in many domains to predict a categorical binary response[141] based on two or more predictors. For example, for a *response (criterion)* variable with two possible values (e.g., Yes or No), logistic regression offers the possibility to attach probability values to the responses given a set of predictors. That is, logistic regression helps understand how multiple predictor variables, together, predict the response or criterion variable membership in one or the other of the two categories of the dependent variable.

[141] A binary response takes values 0 or 1. For example, between the right or wrong answer, between survival and death, or between to buy and not to buy.

The dichotomous nature of the response variable prevents the calculation of a numerical value, as it is the case with regular regression tests. Instead, it uses the binomial probability theory, with only two values to predict, and the maximum likelihood method to generate a best fitting equation that is used to classify the data to the appropriate category based on the regression coefficients.

The basic formula for Logistic Regression is similar to the one used in Linear Regression:

$$logit(P) = a + b \cdot X$$

For multiple predictors, the formula changes to:

$$logit(p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_n \cdot X_n$$

In the equation above, p is the probability the outcome or characteristic of interest is attained, X_i represent the predictors, and β_i represent the relative contribution of these factors.

The dependent variable in Logistic Regression is a *logit*, the natural logarithm of the odds:

$$logit(p) = log(odds) = ln(\frac{P}{1-P})$$

where P is the probability of predicting a 1 (attain the outcome of interest).

In the end, what we are interested in is the probability (P) that the desired outcome occurs. For that, we need to do a little bit of simple algebra[142]:

[142] While they may look scary at first sight, the computations are simple and will be explained in context later in this chapter.

$$
\begin{aligned}
ln(\tfrac{P}{1-P} &= a + bX) \\
\tfrac{P}{1-P} &= e^{a+bX} \\
P &= \tfrac{e^{a+bX}}{1+e^{a+bX}}
\end{aligned}
$$

The Study

To understand a bit better where and how logistic regression is useful, let's look at an example study designed to investigate the effects of self-explanation[143] on learners' performance on causal reasoning tasks. Specifically, this study was designed as a completely randomized[144], two group (control and treatment), between-subjects[145] experiment that used self-explanation to elicit causal mechanism explanations when reasoning about causally linked events. The target field was medicine, a domain that relies heavily on the understanding and use of extensive and complex causal processes. The overarching research question was:

[143] *Self explanation* is an explanation a learner generates on his or her own, as opposed to explanations provided by an external source (e.g., book, instructor, peer).

[144] The participants were randomly assigned to one or the other of the two experimental groups.

[145] Because each participant is member in only one of the two experiment groups: control or treatment.

In the medical field, when learners are reasoning causally, does using self-explanation to elicit an explanation of the causal mechanism(s) improve, on average, learners' performance on tasks involving such reasoning processes?

Based on the existing literature, the study used prompts to train learners to self-explain before answering a question. That is, to think about and attempt to formalize in writing the principle(s) involved in solving practice problems presented to them. For the purpose of this experiment, the participants were randomly assigned to one of the two groups, control or treatment. The participants in both groups were asked to answer the same multiple choice questions in the training stage, with the difference that those participants in the treatment group were asked, in addition to answering the question and before choosing an answer, to formally explain the causal mechanism behind the problem posed in the question. The participants in the control group were only prompted to choose an answer, without being prompted to explain the mechanism first. It was hypothesized that the participants that had a chance to practice self-explanation (those in the treatment group, who were asked to explain before responding) would perform better, on average, than the control group on a subsequent similar problem, for which the prompt was removed and all participants, in both groups, performed the same task.

Learners' performance was measured using a single multiple-choice question with only one correct answer for which learners first selected an answer they believed to be correct and then explained the mechanism that supported their choice. An opportunity to change the answer was offered to the participants after they submitted their explanation of the phenomena with a request to explain why the new answer is better than the previous one. The instrument also assessed the participants' prior knowledge related to the topic used in testing, both as self-assessment and, more objectively, through a set of multiple choice questions. In addition, *age group, gender, income group, undergraduate major, and intended medical speciality* were collected as demographic variables.

About 350 first- and second-year medical students were invited to participate in the study and processes were set in motion to convince enough students to participate to at least meet the minimum sample size of about 100[146]. In the end, the recruitment efforts generated a sample of 117 valid responses.

[146] A basic sample size of 88 was computed based on recommendations from Keppel (1991, p. 74) of at least 44 participants per group for medium 0.6 effect size, a power of 0.8, and an alpha level of 0.05. Recommendations from other authors ranged from 40 to 60 per group. Therefore, a choice was made to consider 50 participants per group an acceptable value, which makes 100 participants the minimum sample size for the experiment.

This example will cover only a subsection of the full study that was answered using logistic regression and will use a curated data set. Data manipulation and transformation procedures used to generate the data set used in this example are not covered. This example will focus on the following research question:

Does the practice of self-explanation as causal mechanism elicitation technique affects, on average, learners' performance on causal reasoning tasks?

The variables included in the model are:

Categorical Performance Score (criterion/response, nominal scale) - calculated assigning a value of 0 to a wrong answer choice or a value of 1 to the correct answer choice.

Experiment Group (predictor, nominal scale) - determined by the group (control or treatment) to which the participant was randomly assigned to.

Year of Study (covariate) - introduced as covariate to control for potential differences in performance due to where the student is situated on the progression timeline in medical school (first or second year medical students). This attempts to account for additional knowledge, experience, and other skills that may help their performance on causal reasoning tasks.

Assumptions

As with all other statistical analysis tests, Logistic Regression has some requirements to be met:

- The *response (criterion)* variable has to be *dichotomous* (has only two values). For this example, the response variable is dichotomous by design. Therefore, this assumption is verified.
- The *groups (categories)* are *mutually exclusive*, meaning that one case can only be in one of the groups. The random assignment of participants to one of the treatment groups, either control or treatment, and only one group, verifies this assumption.

Analysis

The data file has been prepared beforehand to include, from the more than 60 variables in the raw data set, only variables that may be relevant to this analysis. So, first, let's familiarize with the data. Table 25 shows the first few data rows.

ids	case	year	group	pk	score
7	1	1	1	6	0
9	1	1	1	12	1
10	1	1	2	8	0
11	1	1	1	11	0
12	0	1	2	6	0
14	1	1	2	14	0

Table 25: Logistic regression data

The variables of interest are *group*, *score*, and *year*. A first step in the analysis is to convert the variables into factors[147]. For this purpose two new variables, *groupf* and *yearf* will be added to the data set, representing *group* and *year* as factors. Once this step has been performed, it is time to define the model. The research question being investigated here asks if the treatment (practice of self-explanation) affects, on average, performance on causal reasoning tasks. Therefore, the model will look at how performance (DV) represented by the variable *score* is related to the predictor *group*. The variable *year* is introduced to account for the potential effect of the year in medical school[148]. A summary shows the count of records (frequencies) for each category for each of the predictors converted to factors.

[147] In R, logistic regression analysis requires the predictors to be defined as factors.

[148] First or second year students.

```
Group (1=Control Group, 2=Treatment Group)

 1  2
61 56
```

Variable	Values	Description
Treatment Group	0	Control group (recoded as 1 in original dataset
	1	Treatment group (recoded as 2 in original dataset
Year of Study	0	First year medical students (recoded as 1 in original dataset
	1	Second year medical students (recoded as 2 in original dataset

Table 26: Logistic regression predictor (factor) recoding

```
Year of Study (1=First Year Students, 2=Second Year Students)

 1  2
44 73
```

For the analysis, R[149] converts the predictors (factors) to values of 0 and 1. For the current analysis and data set, the recoding (or dummy coding[150] as it is sometimes known) is performed as shows in Table 26.

The logistic regression equation and the model we start this analysis with is:

$$logit(p) = \beta_0 + \beta_1 \cdot group + \beta_2 \cdot year + \beta_3 \cdot group \times year$$

In this model *group* and *year* are the main effects[151] while the term $group \times year$ represents the interaction effect[152]. Let's look at how a summary of this model looks like.

[149] Other statistical analysis packages, such as SPSS, perform a similar conversion.

[150] Is the process of recoding a categorical variable with 2 or more levels into a binary variable (categorical variable with only 2 levels), with values 0 and 1, variable known as *dummy variable*.

[151] The effect of each variable taken individually on the response (DV) variable.

[152] The combined, simultaneous, effect of the two variables taken together.

```
Call:
glm(formula = score ~ groupf + yearf + groupf * yearf, family = binomial,
    data = logReg)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.394  -0.992  -0.781   0.975   1.634

Coefficients:
              Estimate Std. Error z value
(Intercept)    -0.9445     0.4454   -2.12
groupf2        -0.0852     0.6854   -0.12
yearf2          0.4925     0.5615    0.88
groupf2:yearf2  1.0336     0.8376    1.23
              Pr(>|z|)
(Intercept)      0.034 *
groupf2          0.901
yearf2           0.380
groupf2:yearf2   0.217
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.10  on 116  degrees of freedom
Residual deviance: 148.74  on 113  degrees of freedom
AIC: 156.7

Number of Fisher Scoring iterations: 4
```

Looking at the output above, it can be noted that the name of the variables in the left column is accompanied by a value. That is because R[153] converts the predictors in a process known as *dummy* or *treatment* coding. It creates a set of dichotomous[154] variables (see Table 26 as well) where each level of the predictor is contrasted to a predefined reference level chosen to be as one of the values of the respective predictor variable. In this analysis the variables have only two levels, therefore the process consists of choosing a reference level as one of the two values. In this case the value 1 (representing the control group) is selected as the reference level. The value 2[155] in the output indicates that the treatment group (represented by the value 2 in the original dataset) is contrasted to the reference level, the control group. If the predictor has more than two levels, one of the levels will be chosen as reference level and two or more dichotomous variables will be generated for the remaining levels, that contrast each of them with the reference level. Each of these levels will be entered as a separate factor in the output.

[153] Other statistical analysis software application perform the same conversion.

[154] Dichotomous variables have only two values

[155] In the output R uses the values of original variables 1 and 2 and not the internal values of 0 and 1 it uses for analysis. This is because while the analysis uses numbers, the output can use strings to provide more information, if the data was collected and entered with string labels or values.

The analysis of the full model shows that there may be no significant main effects or interaction effects. Nevertheless, further analysis can be conducted to learn if there may be a model that, using only a subset of the variables, may show significance. For this purpose a stepwise logistic regression can be conducted.

Similar to multiple linear regression, the stepwise analysis can be conducted either *forward* or *backward.* The *forward* approach starts with a blank model and enters each term one at a time, computes the model, and compares it against the previous one. The process will continue as long as the difference in predictive power between the more complex model[156] and its predecessor is significant. Once it finds an insignificant gain in predictive power, the process stops. The *backward* approach looks at things in reverse. It starts with the full model and starts removing variables in decreasing order of their contribution to total variance.

[156] Has more variables than the previous one.

For this specific case, considering that we started with the full model[157] we'll use the *backward* approach. The output of this model suggests that a more parsimonious model exists. It includes only the main effects and shows significance for *year* and the model's *constant*[158]. An ANOVA analysis conducted between the competing models shows which factor(s) were eliminated.

[157] Known as *enter* model, in which case all terms of the model are entered at the beginning.

[158] The (Intercept) line of the output.

```
Start:  AIC=156.7
score ~ groupf + yearf + groupf * yearf

               Df Deviance AIC
- groupf:yearf  1      150 156
<none>                  149 157

Step:  AIC=156.3
score ~ groupf + yearf
```

```
        Df Deviance AIC
<none>           150 156
- groupf  1      153 157
- yearf   1      156 160


Call:
glm(formula = score ~ groupf + yearf, family = binomial, data = logReg)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.319  -1.061  -0.705   1.042   1.739

Coefficients:
            Estimate Std. Error z value
(Intercept)   -1.264      0.393   -3.22
groupf2        0.608      0.389    1.56
yearf2         0.984      0.416    2.37
            Pr(>|z|)
(Intercept)   0.0013 **
groupf2       0.1181
yearf2        0.0180 *
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.10  on 116  degrees of freedom
Residual deviance: 150.28  on 114  degrees of freedom
AIC: 156.3

Number of Fisher Scoring iterations: 4
```

So let's look at what the ANOVA output tells us.

```
            Step Df Deviance Resid. Df
1                NA       NA       113
2 - groupf:yearf  1    1.539       114
  Resid. Dev   AIC
1      148.7 156.7
2      150.3 156.3
```

The ANOVA analysis shows that when the interaction factor was eliminated, the model's AIC has improved slightly, effectively making this more parsimonious model a better predictor for the response variable than the full model. To interpret the results we need to look again at the generic equation of the logistic regression:

$$logit(p) = \beta_0 + \beta_1 \cdot group + \beta_2 \cdot year + \beta_3 \cdot group \times year$$

From the logistic regression output, the β coefficients in the logistic regression equation are found in the Estimate column. Therefore, with values, the equation becomes:

$$logit(p) = -1.264 + 0.6084 \cdot group + 0.9836 \cdot year$$

The *odds ratio* computed for each of the parameters are indicators of the odds of performing better (answering correctly) versus the odds of

performing worse (answering incorrectly) is increased or decreased by a factor indicated by the value of the odds ratio. The direction is provided by the sign of the raw estimated β coefficient. If the coefficient is negative, the odds are decreased by the computed value while if the coefficient is positive, the odds are increased by the value of the *odds ratio*.

The odds ratio for *group*:

$$group\,odds\,ratio = e^{0.6084} = 1.84$$

Therefore, holding the year of study constant, being in the treatment group[159] increases the odds of preforming better rather than worse by a factor of 1.84. That is, being in the treatment group[160] increases by 84% the the odds of a better performance score than by being in the control group.

[159] Indicated by the number 2 at the end of the variable name in the logistic regression output.

[160] Participants in the treatment group were prompted to use self-explanation to help improve score.

The odds ratio for *year*:

$$year\,odds\,ratio = e^{0.9836} = 2.67$$

This suggests that, holding the treatment constant, being a second year[161] medical student increases the odds of performing better rather than worse by a factor of 2.67, meaning second year medical students see a 167% increase in odds for a better performance score.

[161] Indicated by the number 2 at the end of the variable name in the logistic regression output.

This concludes the logistic regression analysis test. Nevertheless, *every statistical test is run in the context of a study and once the results are know, they should be interpreted in that context.* Next section, while not relevant to the application of the Logistic Regression analysis, is intended to offer insights in how the results of the analysis may be interpreted in the context of the study.

Additional Analysis - TL/DR

The results of the analysis so far are mixed, showing that the treatment itself, while still included in the equation, does not show a significant main effect in the overall sample. Let's look at some of the elements that may have impacted the results, additional information about the study's design, and how these affect data analysis.

First, given the population of students at the medical school was relatively small and considering the expected percentage of respondents, second year medical students offered an insufficient participant pool. Therefore, based on the timeline of the study and the curricula at the

medical school, which ensured that the participants had sufficient knowledge of relevant domains, a decision was made to include first year medical students as well.

Second, the literature and prior pilot studies suggested that prior knowledge in domains relevant to the practice and test questions matters. Therefore, the study included both a subjective measure of prior knowledge, as a self-evaluation assessment reported by the participants, and a more objective, though brief, evaluation of the participants' prior knowledge using multiple choice questions. Including first year medical students offered a chance to better understand the effects of prior knowledge as it is expected this prior knowledge to be less extensive than that of second year medical students.

With this new knowledge, the model that has been presented thus far can be extended to account for the effects of prior knowledge while controlling for the year of study. By introducing prior knowledge in the regression equation, the model selected included an interaction between prior knowledge and treatment group which suggests that the treatment works differently for different levels of prior knowledge.

The existence of an interaction term is relevant because the interpretation can no longer be conducted for each individual predictor while holding the others constant as explained for an interpretation of a logistic regression showing only main effects. In this case, the interpretation covers multiple simple regression equations, for each level of the predictors that are part of the interaction term. For example, consider the following regression equation[162]:

[162] pk = prior knowledge

$$logit(p) = \beta_0 + \beta_1 \cdot group + \beta_2 \cdot pk + \beta_3 \cdot group \times pk + \beta_4 \cdot year$$

The interaction term is represented by $group \times pk$. For this equation, analysis can be conducted for the two levels of the group (0 = control group, 1 = treatment group), by entering the values 0 or 1 into the equation. This will produce the following two regression equations, which only include main effects and can be interpreted as described before.

For $group = 0$ (control group):

$$logit(p) = \beta_0 + \beta_2 \cdot pk + \beta_4 \cdot year$$

The interpretation will now discuss the odds ratio of prior knowledge to affect the response variable for the participants in the control group only.

For $group = 1$ (treatment group):

$$logit(p) = \beta_0 + \beta_1 + \beta_2 \cdot pk + \beta_3 \cdot pk + \beta_4 \cdot year$$

Which can be further reduced to:

$$logit(p) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \cdot pk + \beta_4 \cdot year$$

The resulting equation is interpreted only in the context of the treatment group and can look a the odds ratio for the levels of the prior knowledge predictor to influence the response variable.

References

Andrews, Frank M., Laura Klem, Terrence N. Davidson, O'MalleyPatrick M., and Willard L Rodgers. 1981. *A Guide for Selecting Statistical Techinques for Analyzing Social Science Data.* 2nd ed. Survey Research Center, Instituted for Research, The University of Michigan, An Arbor, MI.

Arnheim, Valentin, Sander Greenland, and McShane Blake. 2019. "Retire Statistical Significance." *Nature* 567: 305–7. https://doi.org/10.1038/d41586-019-00857-9.

Baron, R. M., and D. A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51 (6): 1173–82.

Bulmer, M.G. 1979. *Principles of Statistics.* Dover.

Cochran, W.G. 1977. *Sampling Techniques.* 3rd ed. New York: John Wiley & Sons.

Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

———. 1992. "A Power Primer." *Psychological Bulletin* 112: 155–59.

Everitt, Brian, and Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R.* Springer, New York, NY.

Filzmoser, P., R.G. Garret, and C. Reimann. 2005. "Multivariate Outlier Detection in Exploration Geochemistry." *Computers and Geosciences*, no. 31: 579–87.

Frey, Bruce B. 2016. *There's a Stat for That! What to Do and When to Do It.* Sage Publications, Inc.

Hatcher, E.J., L.; Stepanski. 1994. *A Step-by-Step Approach to Using the Sas System for Univariate and Multivariate Statistics.* SAS.

Hatcher, L. 1994. *A Step-by-Step Approach to Using Sas for Factor Analysis and Structural Equation Modeling.* SAS Institute Inc.

Holmbeck, Grayson N. 1997. "Toward Terminological, Conceptual, and Statistical Clarity in the Study of Mediators and Moderators: Examples from the Child-Clinical and Pediatric Psychology Literatures." *Journal of Consulting and Clinical Psychology* 65 (4): 599.

Jones, Mark, Val Gebski, Mark Onslow, and Ann Packman. 2002. "Statistical Power in Stuttering Researcha Tutorial." *Journal of Speech, Language, and Hearing Research* 45 (2): 243–55.

Kenny, David A. 1987. *Statistics for the Social and Behavioral Sciences*. Little, Brown; Company.

Keppel, G. 1991. *Design and Analysis. A Researcher's Handbook.* Englewood Cliffs, NJ: Prentice Hall.

Kruskal, William Henry. 1988. "Miracles and Statistics. The Casual Assumption of Independence." *Journal of the American Statistical Association*, no. 83: 929–40.

Lewicki, Pawel, and Thomas Hill. 2005. *Statistics: Methods and Applications*. StatSoft Inc.

Lodico, Marguerite G., Dean T. Spaulding, and Katherine H. Voegtle. 2006. *Methods in Educational Research*. Jossey-Bass. A Wiley Imprint, San Fancisco, CA.

McDonald, J.H. 2014. *Handbook of Biological Statistics*. 3rd ed. Baltimore, Maryland: Sparky House Publishing.

Menard, S. 2010. *Logistic Regression. From Introductory to Advanced Concepts and Applications.* Thousand Oaks, CA: SAGE Publications, Inc.

Osborne, Jason, and Elane Waters. 2002. "Four Assumptions of Multiple Regression That Researchers Should Always Test." *Practical Assessment, Research & Evaluation* 8 (2).

Pan, M Ling. 2016. *Preparing Literature Reviews: Qualitative and Quantitative Approaches*. Routledge.

Pedhazur, E.J. 1997. *Multiple Regression in Behavioral Research. Explanation and Prediction.* Third Edition. Wadsworth/Thomson Learning Inc.

Rodgers, J.L., and W.A. Nicewander. 1988. "Thirteen Ways to Look at the Correlation Coefficient." *The American Statistician* 42 (1): 59–66. http://www.jstor.org/stable/2685263.

Salmon, Wesley C. 1998. *Causality and Explanation*. New York: Oxford University Press.

Sawilowsky, Shlomo S. 2009. "New Effect Size Rules of Thumb." *Journal of Modern Applied Statistical Methods* 8 (2): 597–99. https://doi.org/10.22237/jmasm/1257035100.

Schommer, M. 1990. "Effects of Beliefs About the Nature of Knowledge on Comprehension." *Journal of Educational Psychology*.

Schraw, G., L.D. Bendixen, and M.E. Dunkle. 2002. "Personal Epistemology. The Psychology of Beliefs About Knowledge and Knowing." In, edited by B. K. Hofer and P. R. Pintrich, 261–175. Lawrence Erlbaum Associates.

Seltman, Howard J. 2018. *Experimental Design and Analysis*. Downloaded from http://www.stat.cmu.edu/ hseltman/309/Book/Book.pdf.

Sloman, Steven. 2005. *Causal Models. How People Think Abou the World and Its Alternatives.* New York: Oxford University Press.

Stevens, J P. 2002. *Applied Multivariate Statistics for the Social Sciences. Fourth Edition. Mahwah, NJ: Laurence Erlbaum and Associates*. Mahwah, NJ: Laurence Erlbaum; Associates.

STHDA. 2018. "Statistical Tools for High-Throughput Data Analysis." Website. http://www.sthda.com/english/.

Venables, D.M., W.N.; Smith. 2018. "An Introduction to R." CRAN (The Comprehensive R Archive Network). 2018.

Verzani, John. 2014. *Using R for Introductory Statistics*. Second. CRC Press.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The Asa's Statement on P-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Wiley, Larry A., Joshua F.; Pace. 2015. *Beginning R. An Introduction to Statisticsl Programming*. Second Edition. Apress.

Yamane, Taro. 1967. *Statistics, an Introductory Analysis*. 2nd ed. New York: Harper; Row.