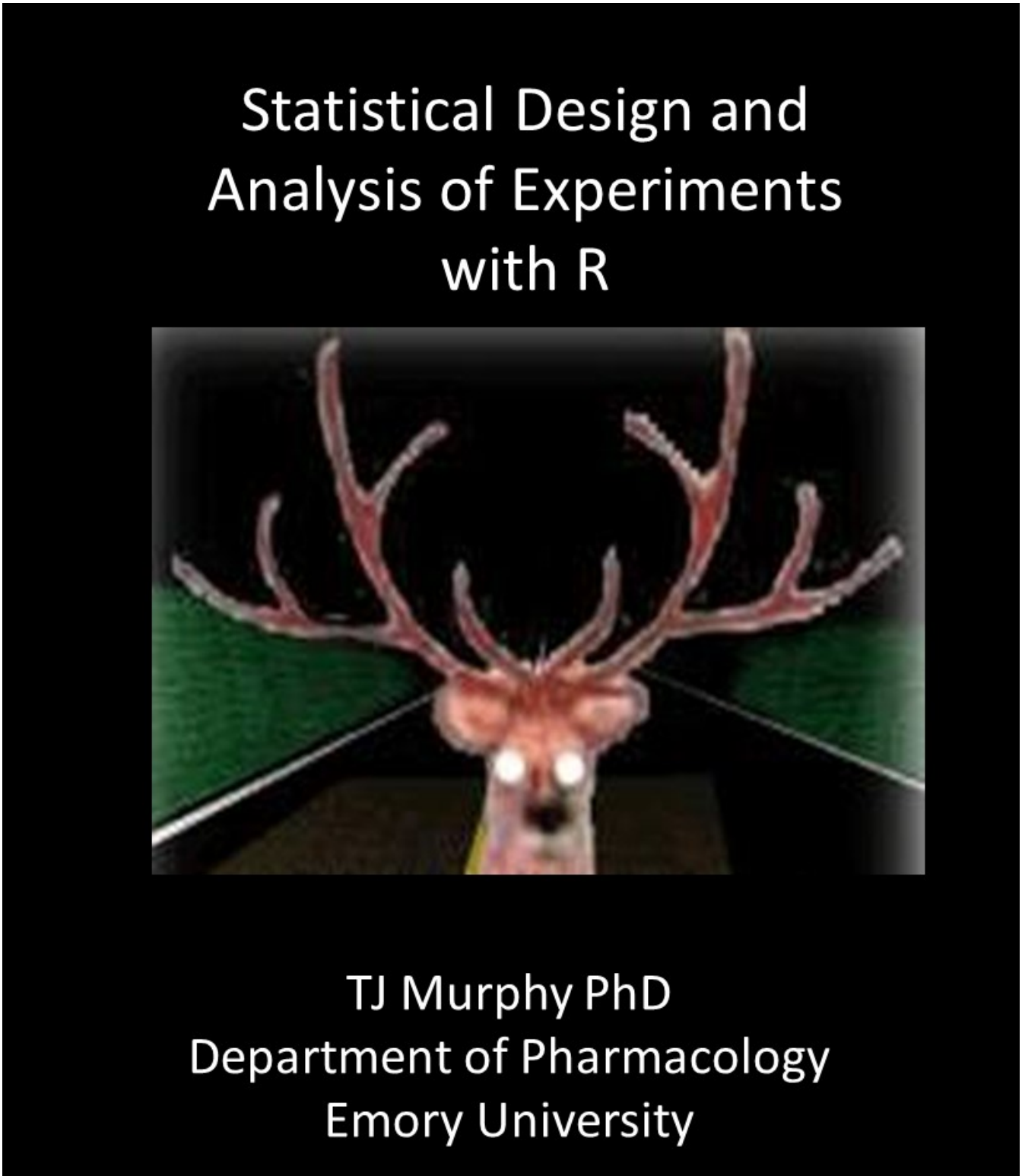
JABSTB: Statistical Design and Analysis of Experiments with R

TJ Murphy PhD, Department of Pharmacology, School of Medicine, Emory University, Atlanta, GA biostats538@gmail.com

2018-12-30

Contents

Preface


Statistical Design and
Analysis of Experiments
with R
TJ Murphy PhD
Department of Pharmacology
Emory University

This book is a resource for students enrolled in my biostats course.

The students are primarily in one of Emory's biomedical and biological sciences PhD programs. There are the ocassional Emory honors program undergrads, students from Emory's public health school, and usually a few Georgia Tech graduate students, also.

When you teach stats you have to choose software and maybe a textbook. I'd been wrestling with the decision to switch my course over to R for a few years, but never really found the right stats book that covers the subject the way I like to emphasize.

When I finally took the plunge I needed to prepare a bunch of handouts with coding examples for the various types of analyses I'd been teaching. Then I decided to write some intro handouts.

Before you know it, you have a "book". This is that. JABSTB.

Not included in this book are additional materials for the course (eg, take home and group exercises, slide decks, data sets, my extensive collection of stats cartoons, etc).

The scope of the book is to provide a) some background on statistical fundamentals that are most relevant to the biomedical researcher who is testing ideas by generating data (sampling, error, statistical hypotheses, experimental design) and then, b) provide examples for running and interpreting various statistical functions.

These are meant to be starting points. What's really nice about R is its adaptability. Copy and paste my examples into your R script or R markdown file, understand how they work by trying to break them, and then get to work modifying them for your own purposes.

Most chapters have code that require R packages beyond those that come in the base R installation. Those are listed at the head of each chapter.

Each chapter has a corresponding RMarkdown document. If you wish to grab those documents instead of using this material as HTML, go grab it on Github. Simply fork, clone or download them from the Github jabstb repo. You'll just need the Rmd files.

All of the other files are there for publishing this up as bookdown. By the way, I've optimized these Rmd's to make good HTML pages. I have not even checked to see if they work as PDFs.

This book is a living document. It is subject to a lot of on-the-fly revision. Stuff will be added and eliminated over time.

As I write these words, in Dec 2018, my main disclaimer is that it is definitely an MVP.

If you find errors, have any suggestions, or would otherwise like to contribute, I welcome your help. Please submit a pull request and/or contact me by email. I welcome your additions.

Chapter 1

About the author

A few years ago I adopted this course from Frank Gordon, a colleague who had retired. Like Frank, I'm a biomedical scientist who happens to have high level of interest in statistical methods.

I learned this material as a graduate student at Mizzou. There I took several stats courses as electives. The ones that impacted me the most were taught by the late Gary Krause, then a professor and statistician in Mizzou's agricultural college.

The light turned on for me during Gary's *Experimenal Design* course. That's when the fog of mathematical statistics cleared enough so I could finally "get" the pragmatic value of statistics for the researcher.

What became most clear is that experimental design is a statistical framework for conducting unbiased research. That concept permeates my course and this book.

I was working on my PhD in pharmacology within the medical school. But most of my classmates in Gary's courses were working on a PhD in one of the agriculture programs, usually in some area of agronomy or in animal science.

The problem my classmates shared, which was not one that really affected me, is having one growing or mating season by which to run a fully replicated experiment.

One shot.

That one shot changes everything.

Planning was a priority for them. They needed to map out their experimental design in advance. Once the experiment began, any new wrinkles or oversights would have to wait until the next growing season. They didn't have the luxury of running out to the field to plant another row of the crop, or to arrange additional breeding groups.

Planning was based upon statistical design principles, often in consultation with Gary. Statistics were *a priori* planning and *post-hoc* tests. At the end of the season the samples were harvested. After all the biochemistry was completed at their lab benches, the final statistical analysis was performed according to the planned approach.

In contrast, it is fair to say that most biomedical scientists fail to incorporate statistical design into their

plans. That failure opens up a whole can of worms that can generally be characterized as doing statistics in ways it was never meant to be done.

All too common is the biomedical researchers who takes a more "fly by the seat of their pants" approach to running experiments and collecting data. In this approach, bunches of near and partial replicates are munged together before looking at the results and making a decision about what statistical analysis would be most appropriate to confirm their inclined interpretation.

Unfortunately, that approach is riddled with biases, and sometimes other negative consequences that are even more challenging.

Experimental statistics was invented by the founders as a means of instilling some structure into the planning, discovery and inference process so that unbiased interpretations can be made.

The focus of this course is in teaching statistics as experimental design. The ideal learner will finish the course knowing how to map out the statistical plan for an experiment in advance and appreciate why this is so important to reduce bias.

That same learner will also know how to analyze, interpret, visualize, and write up the results for a wide array of experimental designs. Most of which she will forget immediately.

And since I emphasize pre-planning, this book is full of simulations. That's the really great advantage of using R to teach biostats, in my view.

I'm not a mathematician so I only offer enough theoretical and mathematical statistics to provide a glimpse of how how things work "under the hood". When I do, it is mostly for stuff I think is helpful to interpret statistical output, or illustrate why a test works in a specific way. I very much believe there is an important place for mathematical statistics, I just don't believe I'm the person who should be teaching it.

Scientists have a lot of overt biases and are the last to realize it. Data frequently has a lot of hidden biases we fail to see. That's why operating within a statistical design framework is so important.

For the biomedical PhD student hoping to graduate while still young, a statistical design framework also offers potential to keep things rolling downhill for you. Statistical thinking should help you avoid the time-sucking rabbit holes that are associated with sloppy, inconclusive or uninterpretable experiments and prolonged time to degrees.

Chapter 2

A Brief History of Experimental Design

Researchers in the pre-statistics days lacked the statistical framework that today's researchers take for granted. Our ancestor scientists were remarkably adept at the scientific method, in making observations, and in collecting data with great care. However, they struggled with designing experiments, in summarizing the data, and in drawing unbiased inference from it.

The statistical approach to experimental design we use today was first enumerated about a century ago, largely by Sir RA Fisher. His story is interesting in part because it is just so classically accidental.

At the outset of his career Fisher did not foresee authoring the foundational principles of experimental design and statistics practiced by most of us today. He took that trajectory by accident.

For about five years after graduating from Cambridge, Fisher worked as a census bureaucrat and part time math teacher.

He was smitten by Darwin's theory of evolution, which was the hot discovery of the day, of course. Fisher's side hustle was to work on mathematical problems related to evolutionary genetics. Today, we would probably recognize him as a hobbyist quantitative geneticist or perhaps even as one of the first bioinformaticians. That's certainly where his career ambitions seem laid. He never lost an interest in evolution and would go on to become, unfortunately, a prominent eugenicist. The take-away from that, alone, is that statistics is not a fool-proof antibias framework.

Still, one big contribution he made during this early stage was no small feat. He defined variance as the square of the standard deviation. He proposed that variance is useful as a descriptive statistic for the variability within a population. Further developed, it would soon become the foundation of the multigroup exprimental designs that called ANOVA, the analysis of variance, which are widely used today.

In 1919 Fisher was hired as a temporary statistician by Sir John Russell, the new director of the Rothamsted Experimental Research center in England.

After decades of underfunding Rothamsted had become a bit rundown. Russell, an agricultural chemist who today we would probably categorize as a biochemist, was hired to beef up postwar (WWI) agricultural research in the UK. Upon arrival he realized the station had a large repository of data. Fully expecting to create even more under his leadership. Russell believed bringing a mathematician on board could help him make sense of this data repository.

Thus, Russell hired Fisher to take a temporary position. Today, we would recognize Fisher in his Rothamsted role as a freelance data scientist charged with conjuring meaning from reams of the station's data, some of which represented serial experiments that had been running for decades.

Figure 2.1: RA Fisher in 1913, from the [Adelaide Digital Archive](https://digital.library.adelaide.edu.au/dspace/handle/2440

As he dug in Fisher saw a lot of flaws in the Rothamsted dataset. He had difficulty making sense of much of it. Mostly because the experiments were, in his view, so poorly designed the results were uninterpretable. If that sounds familiar then I've achieved my objective for mentioning it.

Here's when the paradigm shifted. Fisher began to think about the process by which experimental data **should** be collected. Almost immediately after digging into his Rothamsted work he invented concepts like confounding, randomization, replication, blocking, the latin square and other factorial designs. As I mentioned above, his invention of the **analysis of variance** extended his prior work on variance. The procedure of maximum likelihood estimation soon followed, as well.

It was a truly remarkable period. In 1925 Fisher published a small book, *Statistical Methods for Research Workers*. In 1934 he published its extension, *Design of Experiments*. In these works lay the foundations of how researchers today approach their experiments. His statistical procedures, developed with agricultural science in mind, would soon cross oceans…and then disciplines.

Today, experiments that we would recognize as statistically rigorous are those in which Fisher's early principles operate as procedures. We know today that randomization and pre-planned levels of replication are essential for doing unbiased research. The block ANOVA designs he mapped out then are among the most common experimental designs that we see in the biological and biomedical literature today.

There's much more to this history, including many additional players and plenty of controversy that remains unsettled to this day. I emphasize Fisher mostly because his experimental design and analysis procedures remain the standard for prospective experiments today.

Chapter 3

The Big Picture

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. ~ Sir Ronald Aylmer Fisher

```
library(tidyverse)
library(Hmisc)
```

Let's start by listing out some key characteristics that most biomedical experiments share in common.

They…

- tend to involve the use of relatively small sample sizes.
- are usually highly exploratory in nature.
- generate data that are either discrete counts or measurements of continuous scalars.
- are structured by a small group of fairly common experimental designs.
- are usually interpreted in a binary way; as having "worked", or not.
- test hypotheses (though too often these are unstated).
- aspire for rigor, replicability and reproducibility.
- aspire to be unbiased.

The stakes of our work can be pretty high. These include the higher ideals such as the validation of novel scientific paradigms, the steady advancement of knowledge, and opening the door to create impactful solutions, particularly in the realm of human diseases and suffering.

But no less motivating are the issues more related to the professional practice of science. These include ego, the completely natural impulse to seek out validation for an idea, publication and/or commercialization, time to degree, career viability, scientific reputations, and coveted research/investment funds.

The point is that the process of scientific discovery is driven both by ideals and by biases. This is nothing new.

The one big concept that I hope you embrace is that the statistical design and analysis of experiments serves as a working framework within which the biomedical researcher can conduct reasonably unbiased work. The statistical approaches covered in this course, it turns out, were invented long ago with all of these drivers in mind.

3.1 What are experimental statistics?

Experimental statistics are used

- to summarize data into simpler descriptive models.
- as procedures to draw inferences from samples.

- as procedures that guide the design of experiments.
- to serve as framework for conducting unbiased research.

Chances are you thought biostats was just one or two of those bullets, and probably not the latter two.

3.1.1 Descriptive modeling

Statistical models are ways of simplifying or summarizing data so that they can be more readily described and interpreted.

For example, if we have a sample in which blood glucose levels are measured in each of many subjects, clarity demands we explain those results in terms of summary statistics. Thus, we use parameters like the sample mean and standard deviation, or median and ranges or percentiles. The alternative is unthinkable today (but common long ago), which is to discuss each replicate individually.

To emphasize that sample parameters differ from population parameters, the standard in statistical notation is to use roman characters to indicate samples and greek characters to indicate the population. For example,

parameter	sample	population
mean	$\bar{y}$	μ
standard deviation	s	σ
variance	s^2	σ^2

Thus, the sample mean, $\bar{y}$ is an estimate of the population mean, μ.

Statistical tests also have a descriptive element in that they convey information about the experimental design. If you say, "I'm working up a two-tailed paired t-test," say no more. From that alone I know something about your hypothesis, how your replicates are handled, the number of predictor groups, and the type of data you're measuring.

Regression models also describe data. For example, here is the well-known Michaelis-Menten model that describes product formation as a function of substrate concentration.

$$[P] = \frac{[S][Vmax]}{[S] + Km}$$

That's a model we might fit to certain kinds of enzyme kinetic data, because we use it to estimate scientifically meaningful parameters, like V_{max} and K_m.

In fact, mathematical statistics is actually just modeling. Modeling is the process of simplifying data into something more coherent.

Take a simple example of two groups shown here. Each group has been fit to a simple model: that for the mean and standard deviation. Clearly, that model fits the control group much better than it fits the treatment group.

Why do I say that? The treatment group data are much more skewed. Most of the data values are greater than the mean of the group. Sure, a mean can be calculated for that group, but it serves as a fairly crappy summary. Perhaps some other model (or group of statistical parameters) would better convey how these data behave?

This is to point out that learning statistics is about learning to make judgments about which models are best for describing a given data set.

3.1.2 Statistical inference

There are two main types of inference researchers make.

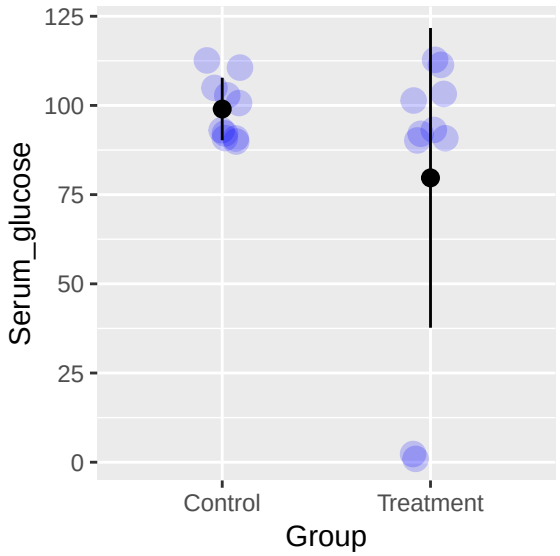


Figure 3.1: Is the mean for each these groups a good descriptive model?

One type is to infer whether an experiment "worked" or not…the so-called "significance test". This familiar process involves calculating a test statistic from the data (eg, t-test, F-tests, etc) and then applying a threshold rule to its value. If the test passes the rule, we conclude the experiment worked. I cover this type of inference in much more detail in the p-value chapter **??**, and we'll talk about it over and again throughout the course.

A second type of inference is to extrapolate from a sample some **estimate** for the values of the variables within the population that was sampled.

Both descriptive and statistical inference are subject to error. By random chance alone our sample could be way off the mark, even with perfectly calibrated instrumentation.

The real difficulty with inference is we can never know for certain whether we are right or wrong.

They are called random variables for a reason. It pays to have a very healthy respect for the role played by random chance in determining the values of our parameter estimates. If we were to completely redo a fully replicated experiment once more, we would almost certainly arrive at different numbers. In a well behaved system, they'd likely be in the same ballpark as those of the first experiment. But they would still differ.

To illustrate, copy and paste the code chunk below. It replicates a random triplicate sample six times, taking six means. Unlike in real life, the population parameters are known (because I coded them in): $\mu = 2$ and $\sigma = 0.4$. You can run that chunk tens of thousands of times and never get a "sample" with one mean that has a value of exactly 2, even though that's the true mean of the population that was sampled.

```
x <- replicate(6, rnorm(3, 2, 0.4))
apply(x, 2, mean)
```

```
## [1] 2.055037 1.770020 2.174044 1.855520 1.997851 2.437189
```

3.1.3 Experimental design

Experimental planning that involves dealing with statistical issues is referred here as experimental design.

This involves stating a testable statistical hypothesis and establishing a series of decision rules in advance of data collection. These rules range from subject selection and arrangement, predetermination of sample size using *a priori* power analysis, setting some data exclusion criteria, defining error tolerance, specifying how

the data will be transformed and analyzed, declaring a primary outcome, on up to what statistical analysis will be performed on the data.

Experimental design is very common in prospective clinical research. Unfortunately, very few basic biomedical scientists practice anything remotely like this. Most biomedical researchers begin experiments with only vague ideas about the statistical analysis, which is usually settled on after the fact. Much of the published work today is therefore retrospective, rather than prospective. Yet, most researchers tend to use statistics that are largely intended for prospective designs. That's a problem.

3.1.4 Statistics as an anti-bias framework

If you are ever asked (for example, in an examination) what purpose is served by a given statistical procedure, and you're not exactly sure, you would be wise to simply offer that it exists to prevent bias. That may not be the answer the grader was hunting for, but it is almost surely correct.

The main purpose of "doing" statistical design and analysis of experiments is to control for bias. Humans are intrinsically prone to bias and scientists are as human as anybody else. Holding or working on a PhD degree doesn't provide us a magic woo-woo cloak to protect us from our biases.

Therefore, whether we choose to admit it or not, bias infects everything we do as scientists. This happens in subtle and in not so subtle ways. We work hard on our brilliant ideas and, sometimes, desperately wishing to see them realized, we open the door to all manner of bias.

Here are some of the more important biases.

####Cognitive biases

From a statistical point of view biases can be classified into two major groupings. The first are Cognitive biases. These are how we think (or fail to think) about our experiments and our data.

These frequently cause us to make assumptions that we would not if we only knew better or were wired differently. If you ever find yourself declaring, "how could this not work!" you are in the throes of a pretty deep cognitive bias. In bench research, cognitive biases can prevent us from building adequate controls into experiments or lead us to draw the wrong interpretation of results, or prevent us from spotting confounding variables or recognizing telling glitches in the data as meaningful.

####Systematic biases

The second are systematic biases. Systematic biases are inherent to our experimental protocols, the equipment and materials we use, the timing and order by which tasks are done, the subjects we select and, yes (metaphorically), even whether the data are collected left-handed or right-handed, and how data is handled or transformed.

Systematic biases can yield the full gamut of unintended outcomes, ranging between nuisance artifacts to false negatives or false positives. For example, poorly calibrated equipment will bias data towards taking inaccurate values. Working forever on an observed phenomenon using only one strain of mouse or cell line may blind us from realizing it might be a phenomenon that only occurs in that strain of mouse or cell line.

####Scientific misconduct

More malicious biases exist, too. These include forbidden practices such as data fabrication and falsification. This is obviously a problem of integrity. Very few scientists working today are immune from the high stakes issues that pose threats to our sense of integrity.

In the big picture, particularly for the biomedical PhD student, I like to call bias the event horizon of rabbit holes. A rabbit hole is that place in a scientific career where it is easy to get lost for a long, long time. You want to avoid them.

The application of statistical principles to experimental design provides some structure to avoid making many of the mistakes that are associated with these biases. Following a well-considered, statistically designed

protocol enforces some integrity onto the process of experimentation. Most scientists find a statistical framework quite livable.

If you give it some thought, the only thing worse than a negative result from a statistically rigorous experiment is a negative result from a statistically weak experiment. With the former at least you know you've given it your best shot. That is hard to conclude when the latter occurs.

Chapter 4

Statistical Sampling

An experiment is no more reliable than is its sample. -TJ Murphy

A statistically valid sample is comprised of independent replicates of the experimental unit, which are generated using some random process.

To unpack this let's think about each of the following terms:

- What are experimental units?
- What do we mean by independent replicates?
- What is a random process?
- When is statistical validity even important?

4.1 Experimental units

The experimental unit is the source of the measurement. An experimental unit can generate one or many measurement values.

I prefer the concept of an experimental unit to the concept of subject, though they often mean the same thing. I've found the word subject carries more ambiguity, especially for people first learning sampling and sample size concepts.

In some experimental designs (eg, unpaired or completely randomized) each experimental unit generates a single measurement value. Here there is a one-to-one correspondence exists between the number of experimental units and the number of measurement values within a data set.

In other designs (eg, paired or matched or repeated/related measure), a single experimental unit can generate more than one measurement values for the same variable. Such data sets have more values than experimental units.

Here are some guidelines for deciding what is the experimental unit in an experiment, with full recognition that sometimes there are gray areas. Ultimately the researcher has to use scientific judgment to recognize or define the experimental unit.

4.1.1 A simple test to define the experimental unit

When defining an experimental unit I recommend using a simple test:

Are these measurements intrinsically-linked?

If two or more measurement values are intrinsically-linked then they would comprise paired or matched or related measures from a single experimental unit.

So how could you judge whether two or more measurements are intrinsically-linked? For the most part, this happens when the source of those measurements doesn't differ.

Here are a few examples:

A before and after design. A mouse is scored on how well it performs a behavioral test at baseline, before a treatment. After that same mouse receives a treatment it is run through the behavioral test once more to get a second score. Those two scores are intrinsically-linked because they were taken from the same mouse. All that differs between the scores is the absence or presence of the treatment, the effect of which the researcher is trying to measure.

We would also say those two scores are matched, paired or related/repeated measures. A single mouse from which two scores are derived is an independent replicate of the experimental unit.

Twinning. Take for example a study involving human identical twins. In these studies identical twin pairs are modeled as a single experimental unit due to their high level of instrinsic relatedness. There are two human subjects but they are modeled statistically as a single experimental unit. The two measurements would be analyzed using a statistical method configured for paired or matched or repeated/related measures.

One of the pair receives a control condition while the other receives a treatment condition. A measurement is taken from each person. There are two measurements in total, and two people, but only a single experimental unit. Given that the twins are so identical we could reasonably conclude these two measurements are intrinsically-linked. We can model the pair as one.

The two measurements would be analyzed using a statistical method configured for paired or matched or repeated/related measures.

Unpaired or completely randomized In contrast, imagine a study using the same control and treatment conditions using unrelated humans (or some other outbred animal species) as subjects. Each subject is assigned either a treatment or a control, and only a single measurement is taken from them. Since the subjects are each very different from each other, we could not conclude that measurements taken from them are intrinsically-linked. Each person stands alone as an experimental unit. The data would be analyzed using an unpaired, unmatched or completely randomized test.

Intrinsically-linked measurements are very common in bench work. In fact they are too often overlooked for what they are and mistakenly analyzed as unmatched. Experiments involving batches of biological material, cultured cells and/or littermates of inbred animal strains routinely involve intrinsically-linked measurements. As a general rule, these should always be designed and analyzed using matched/paired/related measures procedures.

Cell cultures Cell cultures are remarkably homogeneous. The typical continuous cell line is a monoculture passaged across many doubling generations.

Imagine a test conducted on a 6 well multi-well cell culture plate. Each well receives a different level of some treatment condition, such as a dosing or time-course study.

All of the wells were laid down at the same time from a common batch of cells. Each well is very highly related to all of the other wells. The intrinsic differences between wells would be relatively minor and mostly due to technical variation. There's no real inherent biological variation from well-to-well other than that attributable to the level of treatment the well receives.

As a result, all of the measurements taken from a plate of wells are intrinsically-linked to each other. The experimental unit is the plate. They should be designed and analyzed using matched/paired/related measure statistical procedures.

Furthermore, any other plates laid down at the same time from the same source of cells are virtually identical clones of each other. If we were to expose the wells in all of those plates to various treatments followed by taking some measurement, then it is pretty easy to argue that all of those measurements taken on that passage of cells are intrinsically-linked. None of the wells are independent of any of the other wells, irrespective of the plate. Together, all of the plates represent a single experimental unit.

Inbred mice In many regards, the high level of relatedness within inbred mouse strains doesn't differ from human identical twins, or from cultured cells, for that matter.

A given strain of these animals are inbred to genetic homogeneity across several generations. For all intents and purposes all mice derived from a given strain are immortalized clones of each other. Two mice from the same litter are identical twins. Indeed, two mice from different litters from the same strain are identical twins.

Due to their clonal identity all measurements taken from any of these highly related subjects are intrinsically-linked.

Just as for cell culture, protocols must be contrived to break up the homogeneity. A common approach is to treat the litter as the experimental unit and take measures from littermates as intrinsically-linked.

Split tissue Imagine two slices of an organ (or two drops of blood) taken from a single animal. Although the two slices (or drops of blood) are obviously different from each other, any measurements derived from each are intrinsically-linked. The experimental unit would be the animal from which that biological material is derived.

Batches Finally, imagine a batch of a purified protein or other biochemical material. The batch was isolated from a single source and prepared through a single process. The material in the batch is highly homogeneous, irrespective of whether it is stored away in aliquots. Any measurement taken from that batch are highly related to any other measurement. They are intrinsically-linked. The batch would be the experimental unit.

###Blocking

We have to contrive protocols to break up experimental units that have high inherent homogeneity. The statistical jargon used for this is blocking, such that blocks are essentially grouping factors that are not scientifically interesting.

Going back to culture plates. Let's say we prepared three plates on Friday. An assay performed on one plate on Monday would represent one experimental unit of intrinsically-linked measures. An assay repeated on Tuesday on a second plate would represent a second experimental unit. Wednesday's assay on the third plate is also its own experimental unit.

Here the blocking factor is the day of the week. Assuming we created fresh batches of reagents each day, there would be some day-to-day variation that wouldn't exist if we assayed all threee plates at once on a single day. But we're not particularly interested in that daily variation, either.

More conservatively, cell line passage number can be used as a blocking factor to delineate experimental units. Each passage number would represent an experimental unit and the overall replicated experiment would be said to be blocked on passage number.

Defining the experimental unit and any blocking factors requires scientific judgement. That can be difficult to do when dealing with highly homogenous material. What should be avoided is creating a design that limits random chance too severely. To measure on Monday all three plates that were laid down on Friday will probably yield tighter results than if they were blocked over the course of the week.

This has to be thought through carefully by the researcher in each and every case. Reasonable people can disagree what whether one approach is superior to some other. Therefore, what is important is to make defensible decisions. To do that, you need to think through this problem carefully. When in doubt, I suggest leaning towards giving random chance a fair shot at explaining the result you're observing.

For example, you can make the case that measurements from two cell culture plates that were laid down on the same day but are collected on different days are not intrinsically-linked. That's a harder case to make if they are collected on the same day.

You will almost certainly have to make the case that measurements taken from two mice on different days or if they are from different litters are not intrinsically-linked.

Before going there, we need to chat about what we mean by independent replication.

4.2 Independent Replicates

That we should strive for biological observations that are repeatable seems self evident.

An experiment is comprised of independent replicates of treatment conditions on experimental units. The total number of independent replicates comprises an experiment's sample size.

A primary goal in designing an experiment is to assess independent replicates that are not biased to the biological response of a more narrowly defined group of experimental units.

A replicate is therefore independent when a repeat is on an experimental unit that differs materially from a previous experimental unit. A material difference could involve a true biological replicate. Measurements taken from two unrelated human subjects have a material difference. In bench biological work with fairly homogenous systems (eg, cell lines and inbred animals) a material difference will usually need to be some separation among replicates in time and space in applying the experimental treatments.

4.2.1 A simple test for independence

How willing am I to certify this is a truly repeatable phenomenon when replicated in this way?

A new scientific discovery would be some kind of repeatable phenomenon.

4.2.2 Some replication examples

If we are performing an experiment using pairs of human twins, each pair that is studied stands as an independent replicate. Because the pair is the experimental unit, a study involving 5 pairs will have five, rather than ten, independent replicates.

If we conduct an experiment using unrelated human volunteers, or someother out bred animals, each person or animal from whom a measurement is recorded is considered an independent replicate. Their biological uniqueness defines their independence.

We wander into gray areas pretty quickly when thinking about the independence of experimental units in studies involving cultured cells, batches of biological material, and inbred mice. Working with these systems it is difficult to achieve the gold standard of true biological independence. The focus instead should be on repeatability….“Working with new batches of reagents and different days do I get the same response?”

Imagine a 6 well plate of cultured cells. No well differs biologically from any other. If each well received a repeat of the same treatment at the same time we shouldn't consider any measurements from that plate independent from others. Otherwise, the sample would be biased to that plate of cells measured at that particular time with a given set of reagents under those particular conditions. It is too biased to that moment. What if we screwed up the reagents and don't know it?

Rather than being independent, it is best to consider the 6 measurements drawn from the plate as technical replicates or pseudo replicates. The data from the 6 wells should be averaged or totaled somehow to improve the estimate of what happened on that plate that day.

A better approach with cultured cells is to use passage numbers to delineate independence. Thus, a 6 well plates from any one passage are independent experimental units relative to all other passages.

Obviously, given the homogeneity of cells in culture, it's unlikely there is much biological variation even by these criteria. But to achieve true biological independence would require re-establishing the cell line each time an independent replicate was needed. That's rarely feasible.

Inbred mice pose much the same problem. Scientific judgment is needed to decide when 2 mice from the same strain are independent of each other.

One mark of delineation is the litter. Each litter would be independent of other litters. Outcomes of two (or more) littermates could be considered matched or related-measures and thus one experimental unit.

4.3 Random process

You can probably sense intuitively how randomization can guard against a number of biases, both systematic and cognitive. Systematic artifacts become randomly distributed amongst the sample replicates, whereas you are less tempted to treated a replicate as preferred if you don't know what is its treatment level.

Mathematical statistics offers another important reason for randomization. In classical statistics the effect size of some treatment is assumed to be fixed. Our estimate of that real value is the problem. Thus, when we measure a value for some replicate, that value is comprised of a combination of these fixed effects and unexplained effects.

The variation we observe in our outcome variables, the reason it is a random variable, arises from these unexplained effects. These can be particularly prominent in biological systems. Randomization procedures assures those random effects are truly random. Otherwise we might mistake them for the fixed effects that are of more interest us! This concept will be discussed more formally in the section on general linear models.

Suffice to say for pragmatic purposes that random sampling is crucial for limiting intentional and unintentional researcher biases.

Either the experimental units should be selected at random, or the experimental units should be assigned treatments at random, and/or the outcome data should be evaluated at random (eg, blind). Sometimes, doing a combination of these would be even better.

Usually, the researcher supervises this randomization using some kind of random number generator. R's `sample()` function gets that job done for most situations.

Let's design an experiment that involves two treatments and a total of 12 independent experimental units. Thus, 6 experimental units will each receive either of the two treatments. Let's say that my experimental units each have an ID, in this case, a unique letter from the alphabet. Using `sample(1:12)` we randomly assign a numeric value to each ID. This numeric value will be the order by which the experimental unit, relative to the other experimental units, is subjected to the experimental treatment. ID's that are assigned even random numbers get one of the two treatments, and odd numbered ID's get the other treatment.

What we've done here is randomize both the order of replication and the assignment of treatment. That's a well-shuffled deck. You can see how this approach can be readily adapted to different numbers of treatment levels and sample sizes.

```
set.seed(1234)
ID <- letters[1:12]
order <- sample(1:12, replace=F)
plan <- data.frame(ID, order)
plan
```

```
##    ID order
## 1   a     2
## 2   b     7
## 3   c    11
## 4   d     6
## 5   e    10
## 6   f     5
## 7   g     1
## 8   h    12
## 9   i     3
## 10  j     8
## 11  k     4
## 12  l     9
```

4.4 Statistically valid samples

For any statistical test to be valid, each replicate within a sample must satisfy the following two criteria:

- The replicate should be generated by some random process.
- The replicate must be independent of all other replicates.

Why? Statistical tests are one of the last stages of a hypothesis testing process. All of these tests operate, formally, on the premise that at least these two conditions are true.

When these conditions have not been met the researcher is collecting data without testing a hypothesis. To run a statsitical test is to pretend a hypothesis has been tested, when it has not.

4.4.1 Select random subjects

Let's say we want to do an experiment on graduate students and need to generate a representative sample. There are 5 million people in the US who are in graduate school at an given time. Let's imagine they each have a unique ID number, ranging from 1 to 5,000,000. We can use R's `sample()` function to randomly select three individuals with numbers corresponding to that range.

Sampling with replacement involves throwing a selection back into a population, where it can potentially be selected again. In that way, the probability of any selection stays the same throughout the random sampling process.

Here, the `replace = FALSE` argument is there to ensure I don't select the same individual twice.

```
sample(x=1:5000000, size=3, replace = FALSE)
```

```
## [1] 1413668 4617167 1461579
```

All that needs to be done is to notify the three people corresponding to those IDs and schedule a convenient time for them to visit so we can do our experiment.

You can imagine several variations to randomly select graduate students for measurements. You just need a way to find graduate students, then devise a way(s) to ensure the sampling is as representative as possible. Selecting subjects from a real population is pretty straight forward, a bit like picking 8 lotto balls from a spinning container.

A lot of times in experimental work the number of subjects available to the researcher is fixed and smaller. The size of the population to be sampled can be much closer to the number of replicates needed for the experiment rather than a sample from a large pool.

In these cases we have to come up with other ways to randomize.

4.4.2 Randomize to sequence

For example, let's say we want to compare condition A to condition B. We have 6 subjects to work with, each of which will serve as an independent replicate. We want a balanced design so will have 3 replicates for each of the 2 conditions.

Let's imagine we can only perform an experiment on one subject, one day at a time. In that case, it makes sense to randomize treatment to sequence.

We can randomly generate a sequence of 6 even and odd numbers, and assign them to the daily sequence (MTWTFM) based on which random number is first on its list. We can make a rule that subjects assigned even numbers will receive condition A, whereas condition B is meted out to subjects associated with odd numbers.

```
sample(x=11:16, size=6, replace = FALSE)
```

```
## [1] 16 12 15 11 13 14
```

4.4.3 Randomize to location

Let's imagine 3 treatments (negative control, positive control, experimental), that we will code 1,1,2,2,3,3. These will be applied in duplicate to cells on 6-well cell culture plate. We'll code the plate wells with letters, a, b, c, d, e, f from top left to bottom right (ie, a and b are wells in the top row).

Now we'll generate a random sequence of those six letters.

```
sample(letters[1:6], replace=F)
```

```
## [1] "b" "a" "e" "d" "f" "c"
```

Next, we'll map the sequence 1,1,2,2,3,3 to those letters. Thus, negative control goes to the wells corresponding to the first two letters in that sequence, positive control to the 3rd and 4th letters, and so forth.

4.4.4 Randomize to block

In statistical lingo, a block is a subgroup within a sample. A blocked subject shares some feature(s) in common with other members of its block compared to other subjects in the overall sample. But usually, we're not interested in block as a variable, per se.

Here are some common blocks at the bench are

- One purified enzyme preparation vs a second preparation of the same enzyme, nominally purified the same way. The two enzyme preps represent two different blocks.
- A bunch of cell culture dishes plated on Friday from passage number 15 vs ones plated on Tuesday from passage number 16. The two passages represent 2 different blocks.
- A litter of mouse pups born in January vs a litter born in February. The two different litters represent two different blocks.
- An experiment run with freshly prepared reagents on Monday vs one run on Tuesday, with a new set of freshly prepared reagents. Each experimental day represents a block.

Frequently, each block is taken as an independent replicate.

4.5 Independence of replicates

In biomedical research the standard is for biological independence; when we speak of "biological replicates" we mean that each independent replicate represents a distinct biological entities.

That standard is difficult to meet when working with many common biological model systems, particularly cell lines and inbred animals.

The definition of statistical independence is grounded in the mathematics of probability: Two events are statistically independent when they convey no information about the other, or

$$p(A \cap B) = p(A)p(B)$$

.

Here the mathematics is not particularly helpful. Imagine two test tubes on the bench, each receives an aliquot of biological material from a common prep (eg, a purified protein). One tube then receives treatment A and the other treatment B. As best we know, the two tubes aren't capable of influencing each other. But we can reasonably assume their responses to the treatments will at least be correlated, given the common source of biological material. Should each tube be treated as if it were statistically independent?

Replicate independence that meets statistical validity therefore has to take on a more pragmatic and nuanced definition. My preference is to define a **replicate as the independent experimental unit receiving treatment.** I like this because it allows for defining the experimental unit differently depending upon the experimental design.

Chapter 5

Framing statistical hypotheses

"There is no more to science than its method, and there is no more to its method than Popper has said."- Hermann Bondi

Hypothesis-driven research tests predictions about the nature of the world.

Testing hypotheses statistically provides a pragmatic framework for making decisions about the validity of those predictions.

When planning an experiment the primary goal should be to bring hyper-focused clarity to the hypothesis. This is the time to distill your thinking down to the exact question you want answered. What are you studying? What is not known? What is your prediction? How will you measure it? What are your variables? Are they discrete or continuous? How will you test it? How will you decide whether what you predicted happened or not? Will this actually answer the question you're asking?

The statistics taught in this course are for assessing the validity of experimental outcomes in a somewhat odd way: Formally, we test the null hypothesis.

The expectation is to generate observations of such extreme magnitude that we can reject the null, the hypothesis that nothing happened. At first blush that might come off as absurd. Like a Seinfeld episode, where nothing is what is most important.

Hopefully this won't seem so odd after I describe what this accomplishes and explain why it is done this way.

5.1 The decision process

Everybody knows something about the p-value. When it's low enough, the experiment "worked". Before diving into the nitty gritty of p-values, let's jump into a wider angle format to flesh out how they are used.

The framework can be broken down into 5 key steps:

i. We begin with a null hypothesis–yes, the boring one about nothing.
ii. Experiments generate data.
iii. The data are transformed into test statistics.
iv. P-values are calculated from the experiment's test statistic value.
v. Based upon *a priori* thresholds, a decision is made to reject a null hypothesis, or not, depending upon the extremeness of the result.

```
knitr::include_graphics("images/hypothesis.jpg")
```

Low p-values are associated with extreme values of a test statistic. Extreme values of test statistics happen when the effect sizes of the results are high. Rejecting a null on the basis of a p-value means our test statistic

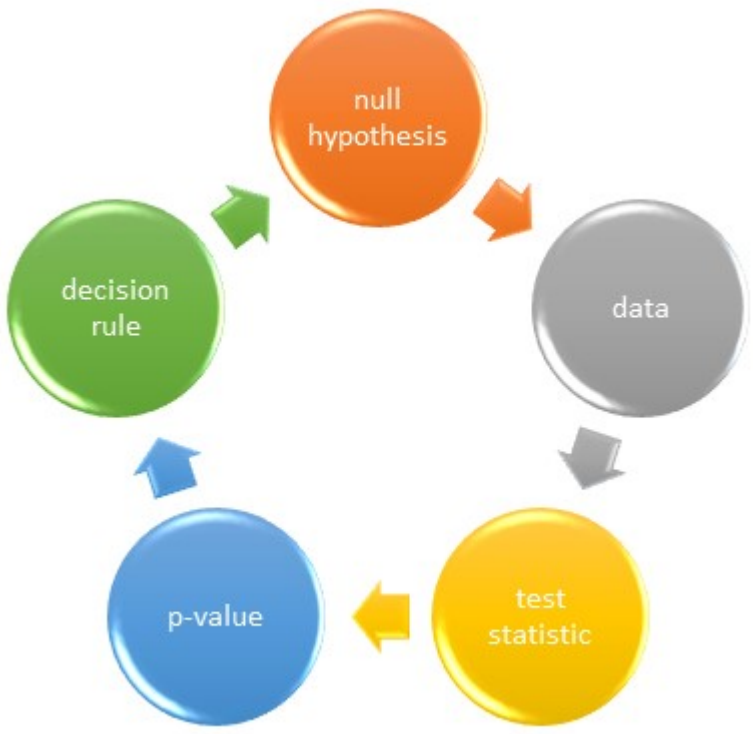


Figure 5.1: Statistical hypotheses test the null in a multistep process

value is too extreme to belong in the distribution of null test statistic values.

Thus, a low p-value means the effect size is improbably high if it were, in fact, not truly effective.

If you learn nothing more in this course, learn that the statistics discussed here are tests of the null hypothesis. Learn that every p-value you see in R output is coupled to a test statistic value. These p-values represent the probability your evidence belongs in the null test statistic distribution.

5.2 Popper and falsification

Using data to falsify an hypothesis, even if that hypothesis is the null, is a decision framework that plays well with philosopher Karl Popper's assertion that scientific theories are probative and that unscientific theories are not.

To Popper, the grandest scientific theories are those that can be falsified. In Popperian logic, the truth of nature is unknowable and unproveable….even though it is testable. Thus, the scientific method advocated by Popper doesn't allow for proving an hypothesis, but at the same time it doesn't forbid us from rejecting hypotheses that are inconsistent with observations.

Thus enters the null hypothesis, which predicts, of course, that nothing happens. The null is an incredibly handy device because if we make observations that are extremely inconsistent with the null, meaning we have observed that something happens, we are obligated to reject the null. Thus, the null is falsifiable when we have positive results!

Imagine an experiment to test whether a drug lowers blood glucose in people who have diabetes. When the glucose-lowering effect size for the drug in the sample is large enough, we can reject the hypothesis that the drug didn't have any effect.

In other words, we will accept an observation as evidence for a positive result by formally concluding that same evidence is inconsistent with a negative result.

Some argue that this logic forces the researcher to test the "wrong" hypothesis and to also accept an alternate hypothesis that itself may not be true. For example, although blood glucose may be lower in the drug treatment arm of the sample, that may have occured by random chance. An unknown confounder variable could be responsible for the observation that drug-treatment is associated with lower blood glucose.

In that case we would make an error by rejecting the null when it is actually true.

Of course, rejecting the null is provisional. All gained knowledge is provisional. Drawing a conclusion from one experiment doesn't preclude testing the experiment some other way. If the problem is important enough (and "real"), it will be tested from multiple angles. It will need to survive the preponderance of the evidence.

I'm convinced the alternative approach, which is to seek out evidence that affirms an hypothesis, is not better. This lies at the heart of what Popper stood against. There is an inherent confirmation bias in seeking out affirmation of ideas. In the proper light, any evidence can be made to look attractive. Furthermore, what if, in seeking affirmation, nothing happens? Negative results are very difficult to interpret because the absence of evidence cannot be interpreted as the evidence of absence.

So I'd hope the researcher who gives this some thought will find null falsification more pragmatic, if not ingenious. It allows us to move forward on the basis of positive evidence (granted, which may be wrong and we don't know it), while at the same time practicing a more sound, more unbiased scientific methodology (hypothesis falsification rather than affirmation).

Meanwhile, this statistical framework does allow for the possibility of designing experimental conditions to minimize false positive (type1) and false negative (type2) errors. We can operationalize our tolerance for those kinds of mistakes in meaningful ways such that we are less likely to become victims of bad luck.

Finally, the decision to reject a null hypothesis can stand alone. It need not be the same as a decision to accept the alternate. Rejecting the null only asserts that the experimental evidence is inconsistent with the null. In no way does that "prove"" the alternative hypothesis.

For most, some of these concerns should become even less of a problem when the null and alternate hypotheses are explicitly framed in terms of population parameters and their **mutually exclusive and collectively exhaustive** outcomes. This approach doesn't leave much room for ambiguity about what is being declared at the decision step.

For example, the null hypothesis for the diabetes case is very explicit: $null, H_0 : \mu_{placebo} = \mu_{drug}$, Here μ, since it is greek notation, represents the mean blood glucose in concentration units in the populations corresponding to the two sampled groups.

Now that we have a bona fide null hypothesis, we can state the alternate hypothesis as everything the null can't be: $alternate, H_1 : \mu_{placebo} \neq \mu_{drug}$

In other words, the inference operates on the basis of straightforward mathematical principles. Two parameters that are compared either meet our prescribed expectations, or they do not. In this case, if we reject the hypothesis that the means of the two groups are equal, then they can only be not equal.

Are they truly not equal? We can never know for sure, but we are operating within a framework of known error tolerances.

5.3 Statistical hypothesis rubric

Researchers have to grapple with two types of hypotheses. One type is the grand, paradigm-driving assertion of some key insight, which is designed to express the big picture in forward thinking terms. It is also designed to wow study sections and seminar audiences.

The other type is the null hypothesis, which is designed to be tested statistically. The null predicts nothing will happen. The null is as boring as it gets. You'd never propose the null in a specific aims page, but you

should get in the habit of thinking in terms of testing the null with your statistics.

Only the null hypothesis has any statistical utility, whereas the grand hypothesis has no statistical utility. This is a conceptual hurdle that most students struggle with. The grand hypothesis is for marketing, the null hypothesis is for mattering.

For that reason I've created a rubric for forming a statistically testable hypothesis. The rubric begins with a conceptual overview of a problem, and it ends with how the results will be interpreted. At some point during the semester you'll have a major assignment that asks you to go through this rubric for a problem of your own choosing. That assignment is a major test for whether you "get" statistical design of experiments.

Step 1: Lay out the big picture of the problem in a way that leads to a "What is not yet known" assertion.

Type 2 diabetes is associated with high blood glucose levels and obesity, which each have long term effects associated with high morbidity. Exenatide is GLP-1 receptor agonist that can control blood glucose levels. When delivered as an osmotic minipump exenatide lowers blood glucose. A standard of care for type2 diabetics is to put them on a weight loss program while giving them drugs that manage blood glucose. It is not known if continuous administration via osmotic minipump can lead to greater weight loss while on this standard of care.

Step 2: Transform the "What is not known" statement into a bold and simple scientific prediction, as if "what is not known" were answered:

Long-term administration of exenatide via osmotic minipump to type-2 diabetics will cause weight loss.

Step 3: Now frame the experimental plan in terms of the independent and dependent variables, written as an if/then statement. In narrative format, if you manipulate what predictor variables, then what outcome do you expect to observe?

If an exenatide osmotic minipump is implanted into type-2 diabetics, then their weight loss will differ compared to placebo.

Step 4: Define the dependent and the independent variables of the experiment. What type of variables are these? What are the experimental units? Are the measurements intrinsically-linked, or not?

The dependent variable will be weight loss, calculated as the weight difference between pre-study to post-study for each human subject. Each subject is the experimental unit. The independent variable is treatment. Treatment is a discrete, factoral variable that will be at two levels, placebo and exenatide. Although pre- and post-study weights will be measured for each subject and are themselves intrinsically-linked, they are used to derive the dependent variable (weight loss), which are not instrinsically-linked.

Step 5: Write the null and alternate hypothesis on the basis of the statistical parameters to be tested. Note here that greek notation is used to symbolize that the hypothesis is about the sampled population parameters, rather than the sample.

Where μ represents the mean weight loss of the populations corresponding to the sampled groups, the null and alternate hypotheses are

$$H_0 : \mu_{exenatide} = \mu_{placebo}$$

and

$$H_1 : \mu_{exenatide} \neq \mu_{placebo}$$

Step 6: What statistical test will be used to test the null hypothesis? What are the decision rules?

A two-sided, unpaired t-test for comparing group means. The sample size will be based upon a power of 90%, which means that the tolerance level for type2 error will be 10%. The decision threshold for type1 error will be 5%. Thus, the null hypothesis will be rejected at a p-value of less than 0.05.

5.3.0.1 Two-sided vs one-sided hypothesis

The above is an example for a two-sided hypothesis. In a two-sided hypothesis $\neq$ is mutually exclusive and collectively exhaustive of $=$. By rejecting the null that two things are equal, we implicitly (and provisionally) accept the alternative hypothesis that they are not equal.

Notice how this hypothesis doesn't predict the direction of an effect. It only predicts there will be a difference between the two groups.

If you're willing to predict the direction of an effect, you would choose to make a one-sided hypothesis.

One-sided hypotheses can happen in either of two ways. In one case we can predict one mean will be greater ($>$)than another mean. In the other case, we can predict one mean will be less than ($<$) another mean. The mutually exclusive and collectively exhaustive alternatives to these one sided hypotheses are therefore $\geq$ and $\leq$, respectively. In other words, if one mean is not greater than another mean, then the only alternative possibilities are that it is less than or equal to it.

The decision to test a one- or two-sided hypothesis should be based upon scientific reasoning.

In the example above, I'm unwilling to test a one-sided hypothesis that exenatide will cause a greater weight loss than placebo, even though that is the expectation (and hope!). Were I willing to test the direction of the effect, the one-sided hypothesis test would be written like this:

$$H_0 : \mu_{exenatide} < \mu_{placebo}$$

and

$$H_1 : \mu_{exenatide} \geq \mu_{placebo}$$

If the data show that mean weight loss is greater in the exenatide group, as expected, that null hypothesis can be rejected.

But what if, unexpectedly, weight loss is greater in the placebo group? It would generate a high p-value. According to the pre-planned hypothesis, the null could not be rejected.

Worse, given they are already enrolled in a standard of care weight loss program, to know the drug actually impairs weight loss would be an important finding. But in choosing the incorrect one-sided hypothesis, there is nothing to do with the result. It is a negative result. I can't flip the tail to the other direction to get a significant result that I wasn't planning upon. That would be extremely biased!

In practice, some researchers caught in this conundrum create a whole new can of worms by simply changing the pre-planned hypothesis after the fact. It's done flippantly but is actually a fairly serious violation of scientific integrity. Changing the hypothesis so that it is consistent with the results is not what anybody would consider sound scientific method.

5.3.0.2 Stick to two-sided hypotheses

Unlike the case above, when being wrong about the direction of an effect is not a big deal, then one-sided tests are not a bad option.

The example above serves to illustrate how a two-sided hypothesis would have been a better choice than a one-sided hypothesis. There are a few other reasons why it is probably better to get in the habit of always testing two-sided nulls: the two-sided test is more conservative because the p-value threshold is a bit lower. Furthermore, multiple tests and confidence intervals easier perform and to interpret, respectively.

Chapter 6

Error

```
library(tidyverse)
library(treemapify)
library(pwr)
```

In a jury trial under the American system of justice the defendant stands accused of a crime by a prosecutor. Both sides present evidence before a jury. The jury's duty is to weigh the evidence then vote in favor of or against a conviction.

The jury doesn't know the truth.

A jury is at risk of making two types of mistakes: An innocent person might be convicted, or a guilty person might be acquitted. They can also make two correct calls: Convict a guilty person or acquit someone who is innocent.

Without ever knowing for sure what is actually true, they are instructed by the judge to record their decision on the basis of a threshold rule. In a trial the rule is vote to convict only when you believe "it is beyond a reasonable doubt" the accused is guilty.

In science the researcher is like a jury. The experiment is like a trial. At the end, the researcher has the same problem that jurors face. There is a need to conclude whether the experiment worked or not. And there's no way to know with absolute certainty. Mistaken judgments are possible.

Whereas the jury works within the "beyond a reasonable doubt" framework, researchers operate within a framework that establishes tolerance limits for error.

Every hypothesis tested risks two types of error. A type 1 error is committed when the researcher rejects the null when in fact there is no effect. This is also known as a false positive. A type 2 error is not rejecting the null when it should be rejected, which is known as a false negative.

Or the researcher might not make an error at all. The sensitivity of an experiment is conclude correctly there is no effect, and power (also known as specificity) is concluding correctly there is an effect. Sensitivity and power are the complements of type 1 and type 2 error, respectively

6.1 Setting type 1 and type 2 error thresholds

In the planning stages of an experiment the researcher establishes tolerance for these errors. A balance has to be struck between aversion for each error type, the ability to make the right call, and the costs involved for being either wrong or right.

6.1.1 Setting alpha-the type 1 error

In the biological sciences the standard for type 1 error is 5%, meaning in any given experiment (no matter the number of comparisons to be made), the chance of generating a false positive should be limited to 5%.

The acceptable type 1 error limit is labeled alpha, or α. In several R statistical functions, it is controlled by adjusting its complement, the confidence level.

Why is α 5% and not some other value? Credit for that is owed largely to R.A. Fisher who offered that a 1 in 20 chance of making such a mistake seemed reasonable. That number seems to have stuck, at least in the biological sciences.

The researcher is always free to establish, and defend, some other level of α. In the field of psychology, for example, α is historically 10%.

There is nothing to stop a researcher from selecting a threshold below or above 5%. She just needs to be prepared to defend the choice.

6.1.1.1 The decision rule

The α is stated before an experiment begins, but operationalized during the final statistical analysis on the basis of p-values generated from statistical tests. **The null hypothesis is rejected when a p-value is less than this preset α.**

6.1.1.2 Experimentwise error

An experiment that just compares two groups (eg, placebo vs drug) generates only one hypothesis. An experiment comparing k groups (eg, placebo vs drug1, vs drug2…drugk-1) generates $m = \frac{k(k-1)}{2}$ hypotheses.

For experiments that generate multiple hypotheses it is important to maintain the overall α for the experiment at 5%. If not checked, the experiment-wise error would inflate with each hypothesis tested.

Several methods have been devised to maintain experiment-wise α for multiple comparisons. The most conservative of these is the Bonferroni correction $\alpha_m = \frac{\alpha}{m}$. Thus, if $m = 10$ hypotheses are tested, the adjusted threshold for each, α_m, is 0.5%, or a p-value of 0.005. If 1000 hypotheses are tested, such as in a mini-gene screen, the p-value threshold for each would be 0.00005.

6.1.2 Power: Setting beta-the type 2 error

In the biological sciences the tolerance for type 2 error, otherwise symbolized as β, is generally in the neighborhood of 20%.

It's a bit easier to discuss β through its complement, $1 - \beta$ or power. Thus, experiments run at 80% power, which are generally regarded as well-designed, run at 20% risk of type 2 error.

Operationally, an experiment is designed to hit a specific level of power via planning of the sample size. "Power calculations" return sample size by integrating intended power, α, and an estimated effect size.

Students tend to fret over effect size estimates. They are nothing more than a best guess of what to expect. A crude estimate. The researcher should use values representing a minimum for a scientific meaningful effect size. The effect size is estimated on the basis of scientific judgment and preliminary data or published information.

If the effect size estimate turns out to be accurate, an experiment run at that sample size should be close to the intended power.

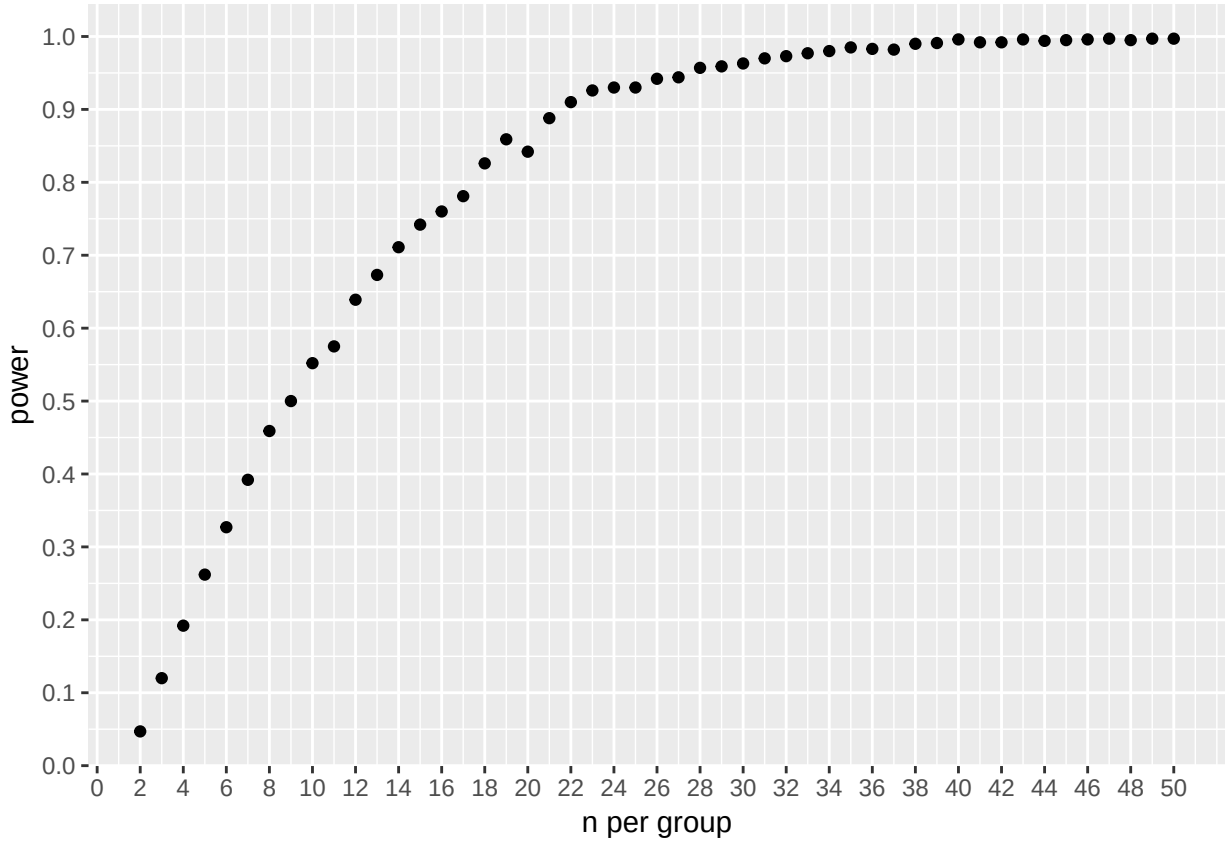
In a perfect world, we might consider powering up every experiment to 99%, completely minimizing the risk of β. As you'll see in the simulation below, the incremental gain in power beyond ~80% diminishes with sample size. In other words, perfect power and very low β comes at a high cost. The choice of what power to run an experiment should strike the right balance between the risk of missing out on a real effect against the cost burden of additional resources and time.

R's `pwr` package has a handful of functions to run power calculations for given statistical tests. These, unfortunately, do not cover all of the statistical tests, particularly for the most common experimental designs (eg, ANOVA).

In this course, we'll emphasize performing power calculations using custom Monte Carlo functions, which can be custom adapted for any type of experiment involving a statistical test.

Here's a custom Monte Carlo-based power function for a t-test. To illustrate the diminishing returns argument, the function calculates power comparing samples drawn from $N(0,1)$ to samples drawn from $N(1,1)$. The graph is generated by passing a range of sample sizes into the function. Note how the gain in power plateaus.

```
t.pwr <- function(n){
  #Intitializers. Means and SD's of populations compared.
  m1=1; sd1=1; m2= 0; sd2=1
  # the monte carlo
  ssims=1000
  p.values <- c()
  i <- 1
  repeat{
    x=rnorm(n, m1, sd1);
    y=rnorm(n, m2, sd2);
    p <- t.test(x, y,
                paired=F,
                alternative="two.sided",
                var.equal=F,
                conf.level=0.95)$p.value
    p.values[i] <- p
    if (i==ssims) break
    i = i+1
    pwr <- length(which(p.values<0.05))/ssims
  }
  return(pwr)
}
#Run t.pwr over a range of sample sizes and plot results
frame <- data.frame(n=2:50)
data <- bind_cols(frame,
                  power=apply(frame, 1, t.pwr))
#plot
ggplot(data, aes(n, power))+
  geom_point() +
  scale_y_continuous(breaks=c(seq(0, 1, 0.1)))+
  scale_x_continuous(breaks=c(seq(0,50,2)))+
  labs(x="n per group")
```




```
## Validation by comparisonC to pwr package results
pwr.t.test(d=1,
           sig.level=0.05,
           power=0.8,
           type="two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 16.71472
##               d = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

6.2 Striking the right balance

The script below provides a way to visualize how the relationship between correct (green) and incorrect (red) decisions varies with error thresholds.

The idea is to run experiments under conditions by which green is the dominant color.

Unfortunately, most published biomedical research appears to be severely underpowered findings.

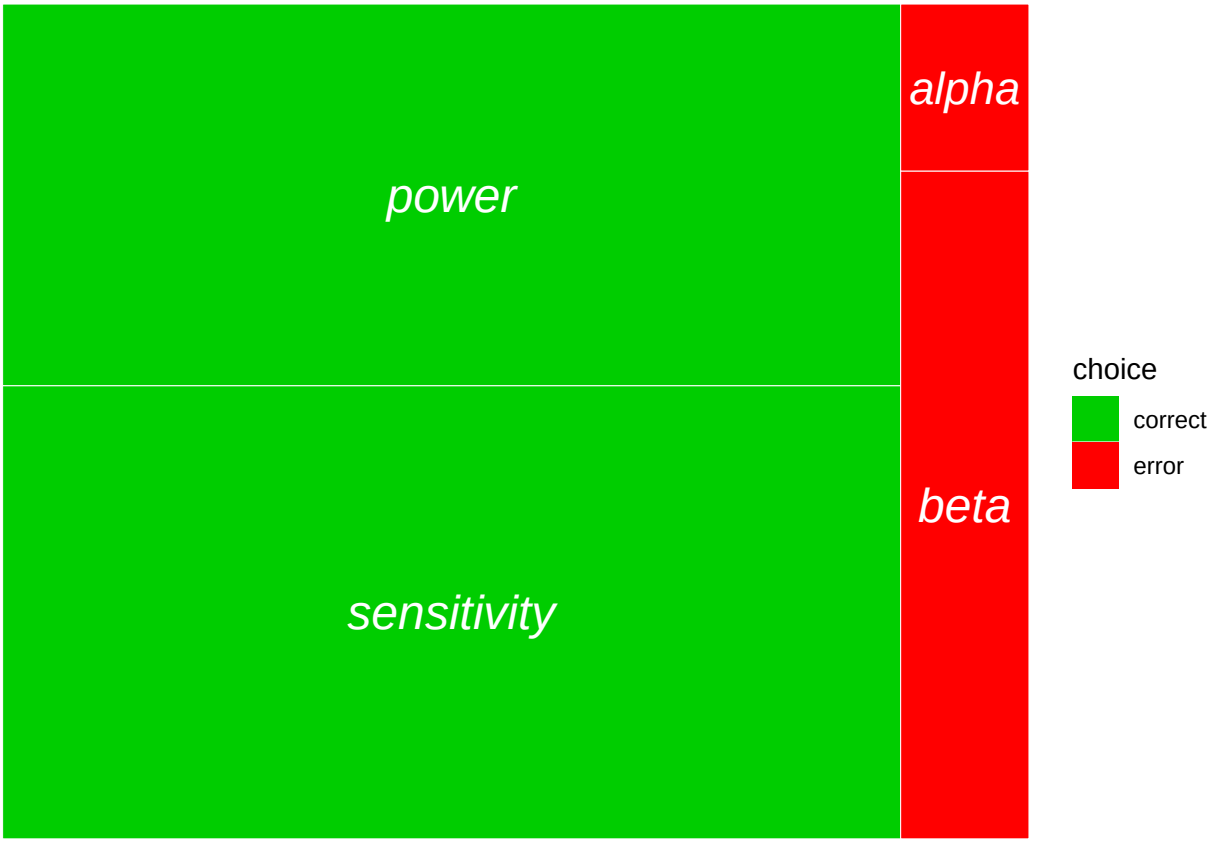
```
alpha <- 0.05
beta <- 0.20

panel <- data.frame(alpha,
                    sensitivity=1-alpha,
                    power=1-beta,
                    beta)
panel <- gather(panel, key="threshold",
         value="percent")

panel <- bind_cols(panel,
                   truth=c("no effect", "no effect", "effective", "effective"),
                   decision=c("effective", "no effect", "effective", "no effect"),
                   choice=c("error", "correct", "correct", "error"))
panel
```

```
##     threshold percent     truth  decision  choice
## 1       alpha    0.05 no effect effective   error
## 2 sensitivity    0.95 no effect no effect correct
## 3       power    0.80 effective effective correct
## 4        beta    0.20 effective no effect   error
```

```
ggplot(panel, aes(area=percent, fill=choice, label=threshold))+
  geom_treemap(color="white")+
  geom_treemap_text(
    fontface = "italic",
    colour = "white",
    place = "centre",
    grow = F
  )+
  scale_fill_manual(values = alpha(c("green3", "red"), .3))
```



6.3 False discovery rate

The false discover rate, or FDR is another way to estimate experimental error.

$$FDR = \frac{false\ positives}{false\ positives + false\ negatives}$$

FDR varies given α, β and the probability of the effect. The probability of the effect bears some comment. Think of it as a prior probability, or the likelihood that an effect being studied is "real". It takes some scientific judgment to estimate these probability values.

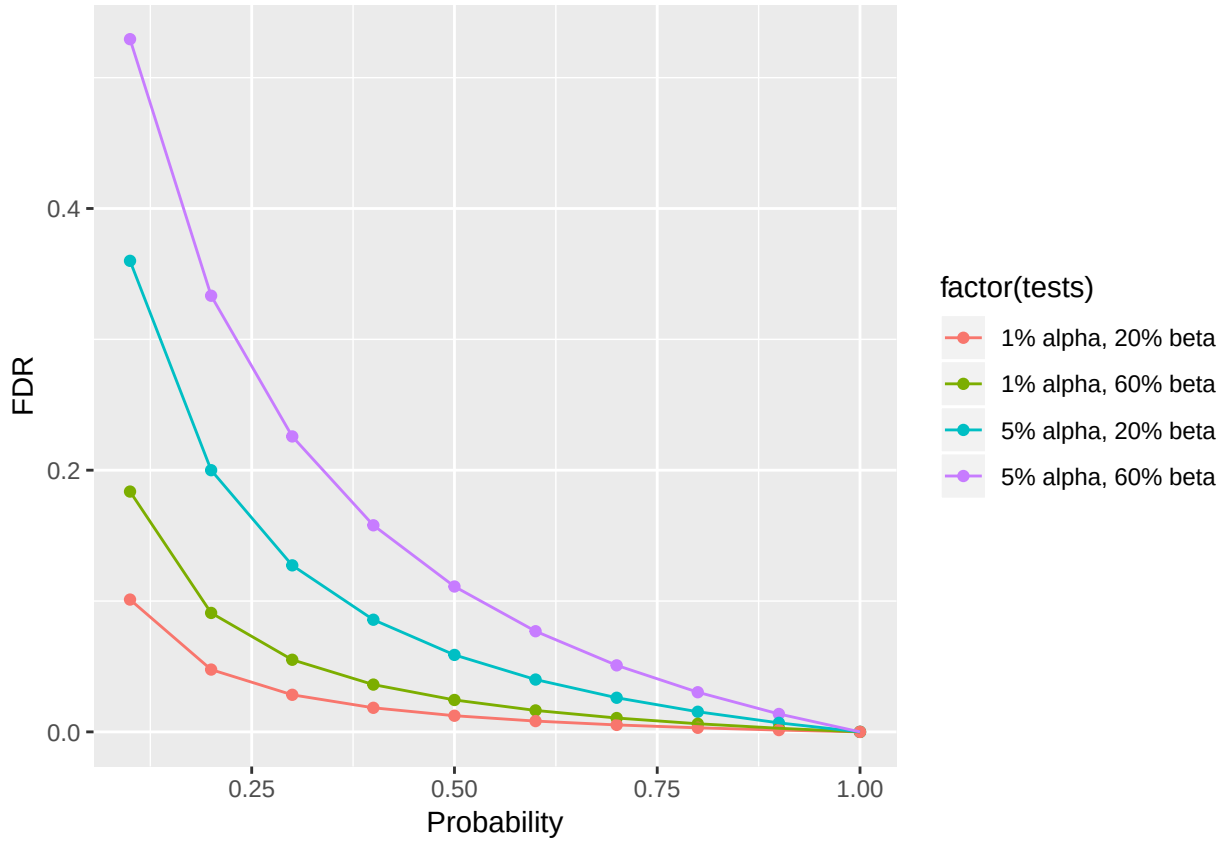
The graph below illustrates how FDR inflates, particularly when running experiments for low probability effects when tested at low power, even at a standard α.

These relationships clearly show that the lower the likelihood of some effect that you would like to test in an experiment, the higher the stringency by which it should be tested.

```
px <- seq(0.1, 1.0, 0.1) #a range of prior probabilities
tests <- 10000
fdr_gen <- function(beta, alpha){
  real_effect <- px*tests
  true_pos <- real_effect*(1-beta)
  false_neg <- real_effect*beta
  no_effect <- tests*(1-px)
  true_neg <- tests*(1-alpha)
  false_pos <- no_effect*alpha
  FDR <- false_pos/(true_pos + false_pos)
```

```
  return(FDR)
}
upss <- fdr_gen(0.6, 0.05)#under-powered, standard specificity
wpss <- fdr_gen(0.2, 0.05)#well-powered, standard specificity
uphs <- fdr_gen(0.6, 0.01)#under-powered, high specificity
wphs <- fdr_gen(0.2, 0.01)#well-powered, high specificity
fdrates <- data.frame(px,upss, wpss, uphs, wphs)
colnames(fdrates) <- c("Probability",
                       "5% alpha, 60% beta",
                       "5% alpha, 20% beta",
                       "1% alpha, 60% beta",
                       "1% alpha, 20% beta")

#convert to long format
fdrates <- gather(fdrates, tests, FDR, -Probability)
ggplot(fdrates, aes(Probability,FDR, group=tests))+
  geom_point(aes(color=factor(tests)))+
  geom_line(aes(color=factor(tests)))
```



Chapter 7

P Values

```
library(tidyverse)
```

You'll see soon enough that when you run a statistical test function in R, it generates list objects that are chock full of useful information.

Invariably, the researcher's eyes will go right to the p-value. This is understandable, since most researchers have been trained to associate their own success with a p-value falling below some pre-set α. Who could really blame them for peeking at the p-values first?

The p-value is an instrument by which an important decision will be made. As such, it is worth understanding how that instrument works.

Hypothesis-driven experiments are designed to test the null hypothesis. That null will be rejected if the effect size is large enough. Extreme effect sizes correspond to extreme values of test statistics.

A p-value is the probability that a given test statistic value could be as large as it is, or even more extreme, if the null hypothesis were actually true.

In other words, the p-value is an error probability. It is also a random variable. Which means that it is always possible for an experiment to generate an extreme test statistic by simple random chance. The p-value asserts the probability that this is the case.

One of the reasons I like using R for experimental statistics is that R makes it easy to simulate p-values. Because of that you build an intuitive sense for how they operate. R makes it easy to understand p-values.

7.1 How p-values are calculated

You can think of test statistics as a transformation of sample data. There are many test statistics. The one to use for a given data set depends on the experimental design. Each test statistic has a probability distribution. P-values are derived from the probability distributions of these test statistics and serve as a way to standardize the decision making process irrespective of the experimental design and test statistic.

Probably the simplest test statistic to understand is the z-score. The z-score is a transformation of data from whatever scale it is on, to a standard normal scale. It's usually appropriate for continuous data.

$$z_i = \frac{y_i - \mu}{\sigma}$$

Let's say we have single blood glucose value of 122 mg/dl. What is its p-value? Is the z-score corresponding to that glucose value too extreme to belong in the null distribution of z-scores?

First, the blood glucose values is transformed into a z-score. We'll say the mean and standard deviation of blood glucose in the sampled population is 100 and 10 mg/dl, respectively. The z-score for a value of 122 is therefore:

```
z <- (122-100)/10; z
```

```
## [1] 2.2
```

z-score units are in standard deviations. Thus, a z-score value of 2.2 indicates it is 2.2 standard deviation units greater than the standard normal mean (which is zero, of course).

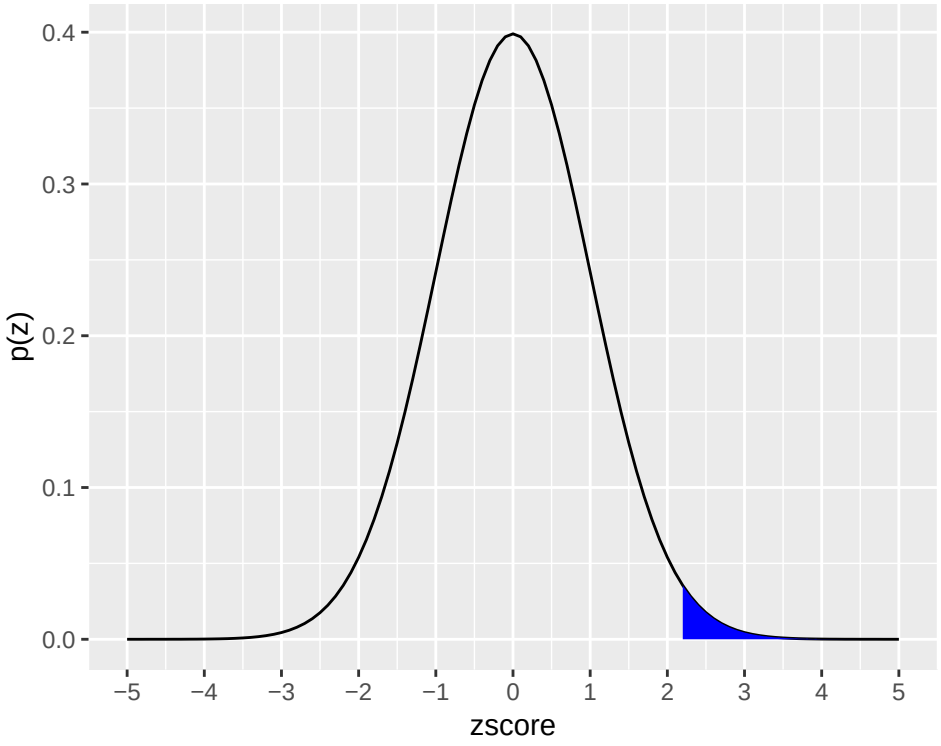
Next, we'll pass that z-score value of 2.2 into the standard normal density function, `pnorm`. We cause the function to produce a p-value for that z-score by using a `lower.tail=FALSE` argument:

```
pnorm(2.2, mean=0, sd=1, lower.tail=FALSE)
```

```
## [1] 0.01390345
```

In the z probability distribution below, the blue shaded region illustrates what this p-value looks like. The p-value covers the probabilities for z-score values of 2.2 and higher. The p-value is thus the area under the curve for the z probability distribution for that value of z and for more extreme values.

```
ggplot(data.frame(zscore = c(-5, 5)), aes(zscore)) +
  stat_function(fun = dnorm) +
  stat_function(fun = dnorm,
                xlim= c(2.2, 5),
                geom = "area", fill="blue")+
  ylab("p(z)")+
  scale_x_continuous(breaks=seq(-5,5,1))
```



7.2 How p-values should be interpreted

The question that's ringing in your ears right now is, "Is a z-score value of 2.2 so extreme we can reject that it belongs to the null distribution of z-scores?"

The answer to that question depends upon what threshold you deem is too extreme. Remember, a threshold is our tolerance for error; in this case, for type 1 error.

If the threshold for an acceptable risk of type 1 error is 5% ($p < 0.05$), then let's see how those look on the z-distribution.

First, let's calculate z-scores corresponding the area outside 95% of the z-scores. Since extreme z-scores can lay on both the right and the left sides of the z-distribution, which is symmetrical. Therefore we split the 5% in half and use the quantile function `qnorm` to calculate z-scores for each:

```
qnorm(0.025, lower.tail = F)
```

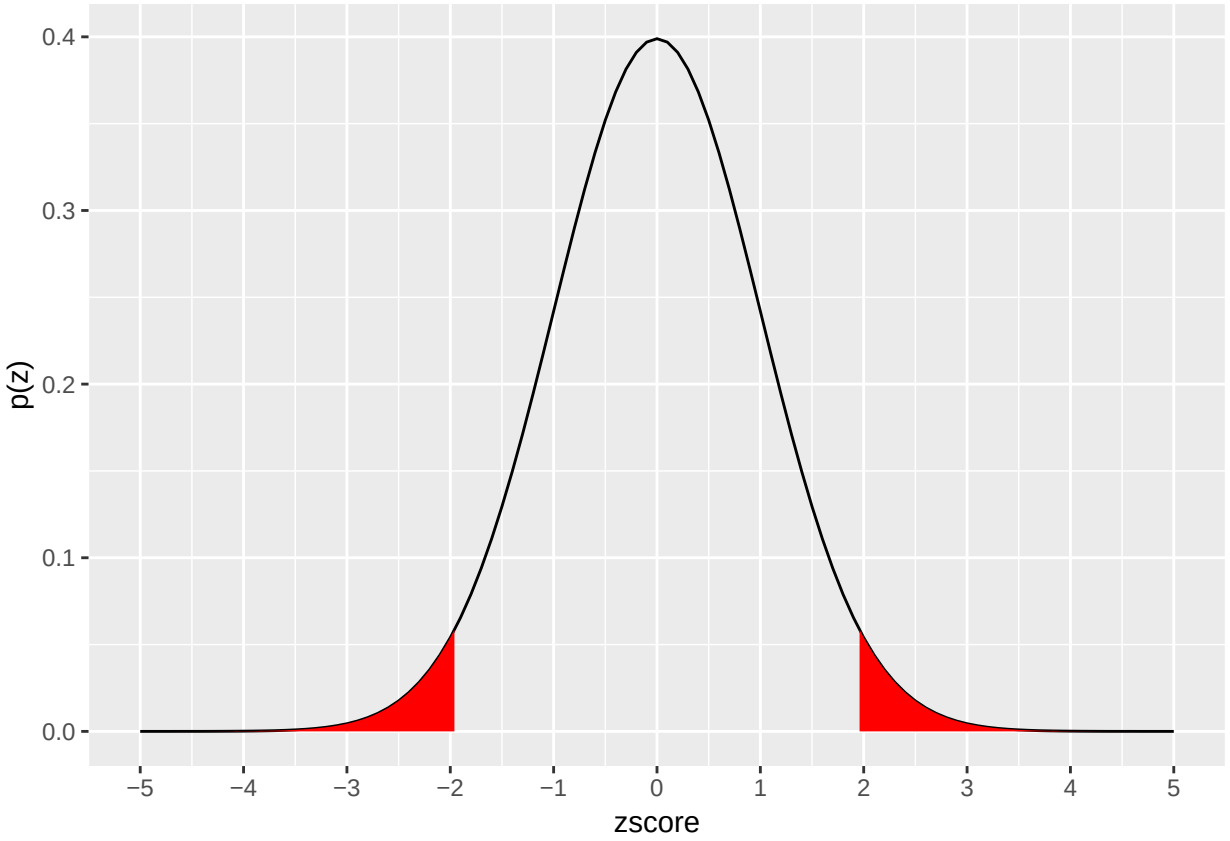
```
## [1] 1.959964
```

```
qnorm(0.025, lower.tail = T)
```

```
## [1] -1.959964
```

Thus, the 95% confidence limits for the z-scores are ~ +/- 1.96, almost 2 standard deviations from the mean. We plug those values as limits into our plot:

```
ggplot(data.frame(zscore = c(-5, 5)), aes(zscore)) +
  stat_function(fun = dnorm) +
  stat_function(fun = dnorm,
                xlim= c(1.96, 5),
                geom = "area", fill="red")+
  stat_function(fun = dnorm,
                xlim= c(-1.96, -5),
                geom = "area", fill="red")+
  ylab("p(z)")+
  scale_x_continuous(breaks=seq(-5,5,1))
```



Any z-score values corresponding to the red-shaded areas would be deemed too extreme to belong to the null. The limit on the right side is 1.96. Therefore, yes, a z-score of 2.2 ($p = 0.0139$) is too extreme to belong to the standard null distribution.

7.3 Interpretation

Every time we do an experiment we operate on the assumption that our data represent the null. This is analogous to considering a defendant innocent until proven guilty. So we think of test statistic values we calculate from our data, unless proven otherwise, as belonging to the null distribution of test statistic values.

The interpretation of $p = 0.0139$ is the probability that z-score (and its corresponding glucose value of 122 mg/dl) are that large by chance is 0.0139. There's about a 1.4% chance we are making an error by rejecting the null that it belongs to the $N(100, 10)$.

7.4 Criticisms of p-values

There are several criticisms of p-values, many of which are legitimate. I'll address a few key ones here.

1. *They are too confusing, nobody understands them.* I get that. I confess that p-values are a struggle to teach in a way that's simple and memorable. Especially for students who only consider statistics episodically, perhaps a few times a year.

This year I'm teaching this with a bit more emphasis upon Popper and the merits of null hypothesis falsification as the cornerstone of the scientific method and how p-values fit into that tradition.

Here it is: All statistical tests (the ones I teach in this course) are tests of the null hypothesis. When the test result is extreme, we reject the null. The p-value is the probability we're rejecting the null in error.

Despite the merit of this particular criticism, p-values are not going away. They are an important inferential tool used by most biological scientists, even if poorly understood and implemented.

Like any tool in the lab, it is incumbant upon the researcher to learn how it works. I think a great way to get a better intuitive understanding for p-values is to play around with the various test statistic probability and quantile distributions in R (`pnorm, qnorm, pt, qt, pf, pf, pchisq, qchisq, psignrank, qsignrank` etc). Use them to run various scenarios, plot them out…get a sense for how the tools work by using them.

2. *p-Values poorly protect from false discovery* This is undoubtedly true. Since David Colquhoun goes over this in blistering detail I won't repeat his thorough analysis here. The researcher MUST operate with skepticism about p-values.

Since Colquhoun's argument is entirely based in simulation it also inspires an approach for dealing with this problem. Through simulation *a priori*, a researcher can design and run experiments *in silico* that strikes the right balance between the threshold levels she can control (eg, α and β) and feasibility in a way that best minimizes the risk of false discovery. Before ever lifting a finger in the lab.

This criticism explains why I am such a strong advocate of Monte Carlo simulation in experimental design. With software like R, there really is no excuse anymore for the researcher being a victim of this problem.

3. *p-Values aren't the probability I'm interested in* Researchers who raise this criticism generally are interested in something the p-value was never designed to deliver: the probability that their experiment worked.

A p-value doesn't provide that information because it is an error probability. Specifically, it is the probability of making a type 1 error.

For these researchers, embracing Bayesian statistics is probably a better option. I don't teach Bayesian statistics in this course for a couple of reasons, but mostly because I don't understand it well enough to teach it, and I don't see how it offers a superior approach for experimental research.

4. *People use p-values as evidence for the magnitude of an effect.* Sure, but they are wrong. This is more a criticism of the people who use p-values, and not the p-value. But the criticism raises the point that it is a mistake to rely solely on a p-value to interpret the outcome of an experiment.

A p-value$<$0.05 only means that there is less than 1 out of 20 chance of having detected an extreme effect when the null is true. A low p-value doesn't provide evidence that the treatment effect is real. As a result, a p-value can't provide any information about the magnitude of the treatment effect.

Neither is a low p-value synonymous with scientific significance. A simple example of this comes from 2 way ANOVA F test analysis. When the test suggests a postive result for an interaction effect, that finding supercedes the main effects. Thus, should any main effects also have low p-values they are not scientifically meaningful.

Researchers should therefore analyze p-values in conjunction with other parameters, such as effect sizes and the confidence intervals.

Chapter 8

Data Classification

```
library(datapasta)
library(tidyverse)
```

The starting point in any statistical design is to understand the types of data that are involved. Ask yourself whether the variables are discrete or continuous. Then ask if they measured, ordered or sorted?

If you don't understand those two questions, just read on.

Because the answers will point you in the proper analytical direction.

This is one of the most important things to learn in this course. If you don't get the concept that not all data types are equivalent, you won't get statistics.

In this section data classification will be discussed. In all likelihood this material will sound simplistic or even obvious to you, but I cannot emphasize enough the importance of data classification in mastering a statistical framework.

If for no other reason, understanding how data are classified is crucial in selecting the most appropriate statistical analysis. If you were to approach me to ask, "here's my stuff, what statistical test should I do?" I would ask, "tell me more about your data." And we would probably spend a lot of time with you answering my questions until I was sure I understood your data classification.

Therefore, a major learning objective for you is, given a data set, to know which variables are dependent and which are independent, and whether the variables involved are continuous (measured) or discrete (ordered or sorted).

8.1 Dependent and independent variables

For the experimental researcher there are two basic types of variables.

An independent variable is the predictor or explanatory variable imposed by the researcher upon a system. Independent variables have values, the levels of which are determined by the researcher. For example, in a blood glucose drug study, the independent variable "Treatment" would come in two levels, "Placebo" and "Drug". In R, we'd call treatment a factor variable with two levels.

Conventionally, the independent variable is plotted on the abcissa, or x-axis, scale of some graph.

A dependent variable is the response or outcome variable collected in an experiment. The values that dependent variables take on are determined by, or dependent upon, the level of the independent variables. For example, the dependent variable in the blood glucose drug study would be a measurement called "blood_glucose".

Most of the time the dependent variable is plotted on the ordinate, or y-axis, scale.

In statistical notation the dependent variable is usually depicted by the uppercase symbol Y. The values that variable can assume are symbolically represented as lowercase symbol y_i, where i is the sample size, ranging from 1 to n independent replicates.

Similarly, the indepedent variable is usually depicted by uppercase X (or some other letter) and its values are lowercase x_i.

I'm going to use that convention but with a twist. Independent variables denoted using X will represent continuous scaled variables, whereas independent variables denoted using A or B, or C, will represent discrete, factoral variables. These will take on values denoted by lowercases, eg, a_i, b_i, c_i) .

To illustrate dependent and independent variables think about a linear relationship between two continuous variables, X and Y . This relationship can be expressed using the model $Y = \beta_0 + \beta_1 X$. X would be a variable the researcher manipulates, such as time or the concentration of a substance. Y would be a variable that the researcher measures, such as absorption or binding or fluorescence. The parameters β_0 and β_1 are constants that modify the relationship between the two variables, which I'm sure you recognize as representing the y-intercept and slope, respectively, of the regression line between the two variables.

Thus, Y takes on different values as the researcher manipulates the levels of X. Which explains why Y depends on X.

For example, here's how the data for a protein standard curve experiment would be depicted. In the R script below the variable X represents known concentrations of an immunoglobulin protein standard in $\mu g/ml$. The researcher builds this dilution series from a known stock, thus it is the independent variable. The variable Y represents A_{595}, light absorption in a spectrophotometer for each of the values of the standard protein. The A_{595} values depend upon the immmunoglobulin concentration. Estimates for β_0 and β_1 are derived from running a linear regression on the data with the `lm(Y~X)` script. Thus, for every one unit increment in the value of X, there is a 0.02497 increment in the value of Y. Again, Y depends upon X.

```
#Protein assay data, X units ug/ml, Y units A595.
X <- c(0, 1.25, 2.5, 5, 10, 15, 20, 25)
Y <- c(0.000, 0.029, 0.060, 0.129, 0.250, 0.371, 0.491, 0.630)
#derive the slope and intercept by linear regression
lm(Y~X)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##  -0.0008033    0.0249705
```

8.1.1 When there is no independent variable

This is a course for experimental biologists. In other types of research, particularly in the public health, behavioral and social science fields, studies are often not strictly experimental. Researchers in these fields generally work with data sets lacking true, experimentally-manipulated independent variables as defined above. Yet these researchers are still very interested learning whether certain phenomena cause other phenomena.

The problem of drawing causal inference from studies in which all of the variables are observed is beyond the scope of this course. Pearl offers an excellent primer on considerations that must be applied to extract causality from observational data here.

8.2 Discrete or continuous variables

At their most fundamental level, the dependent and independent variables of experiments can each be subclassified further into two categories. They are either discrete or continuous.

Discrete variables can only take on discrete values, while continuous variables can take on values over a continuous range. If that's not clear just yet, it should become more clear by reading below.

Variables can be subclassified further as either **measured**, **ordered**, or **sorted**. This subdivision fulfills a few purposes.

First, it's alliterative so hopefully easier to remember. It reminds me of Waffle House hashbrowns, which can be either scattered, smothered or covered, and that is just something you'll never forget once you've visited a Waffle House.

Second, it covers all types of data and statistical testing, and thus forms the basis for drawing a pragmatic statistical modeling heuristic.

```
knitr::include_graphics("images/testing_heuristic.jpg")
```

Third, the "measured, ordered, sorted" scheme classifies variables on the basis of their information density, where measured >> ordered >> sorted.

Different authors/softwares give these three types of variables different names, which creates some confusion. In SPSS, for example, when setting up variables you can choose to classify it as scalar, ordinal, or nominal, which correspond to measured, ordered and sorted. Another fairly common descriptive set for the three types is interval, ordinal, and categorical. These correspond to measured, ordered, and sorted, too.

Though they are named differently, for the most part everybody seems to agree that all variables can be reduced to 3 subtypes, even if they can't agree on what to name them.

8.2.1 Measured variables

Because everything is measured in some sense of the word "measured"" is probably not an ideal choice to describe what is meant when refering to a continuous variable. My rationale for the choice is that it almost always requires some kind of measuring instrument to grab this type of data.

For the present purposes let's considered the terms measured variables and continuous variables as synonymous. Measured variables are fairly easy to spot. Any derivative of one of the seven base SI units will be a measured variable.

```
knitr::include_graphics("images/si_units.jpg")
```

Take mass as an example. The masses of physical objects can be measured on a continuous scale of sizes ranging from super-galaxian to subatomic. Variables that are in units of mass take on a smooth continuum of values over this entire range because mass scales are infinitesimily divisible.

Here's a thought experiment for what infinitesimily divisible means. Take an object that weighs a kilogram, cut it in half and measure what's left. You have two objects that are each one half a kilogram. Now repeat that process again and again. After each split something always remains whose mass can be measured. Even though it gets smaller and smaller. Even when you arrive at the point where only a single atom remains it can be smashed into yet even smaller pieces in a supercollider, yielding trails of subatomic particles….most of which have observable masses.

But here's what's important about continuous variables: That continuity between gradations means that continuous variables can carry more information than other types of variables. That's what I meant by information density, in the comment above.

On a scale of micrograms, an object weighing one kilogram would have one billion subdivisions. If you have an instrument that can accurately weigh the mass of kilogram-sized objects to the microgram level,

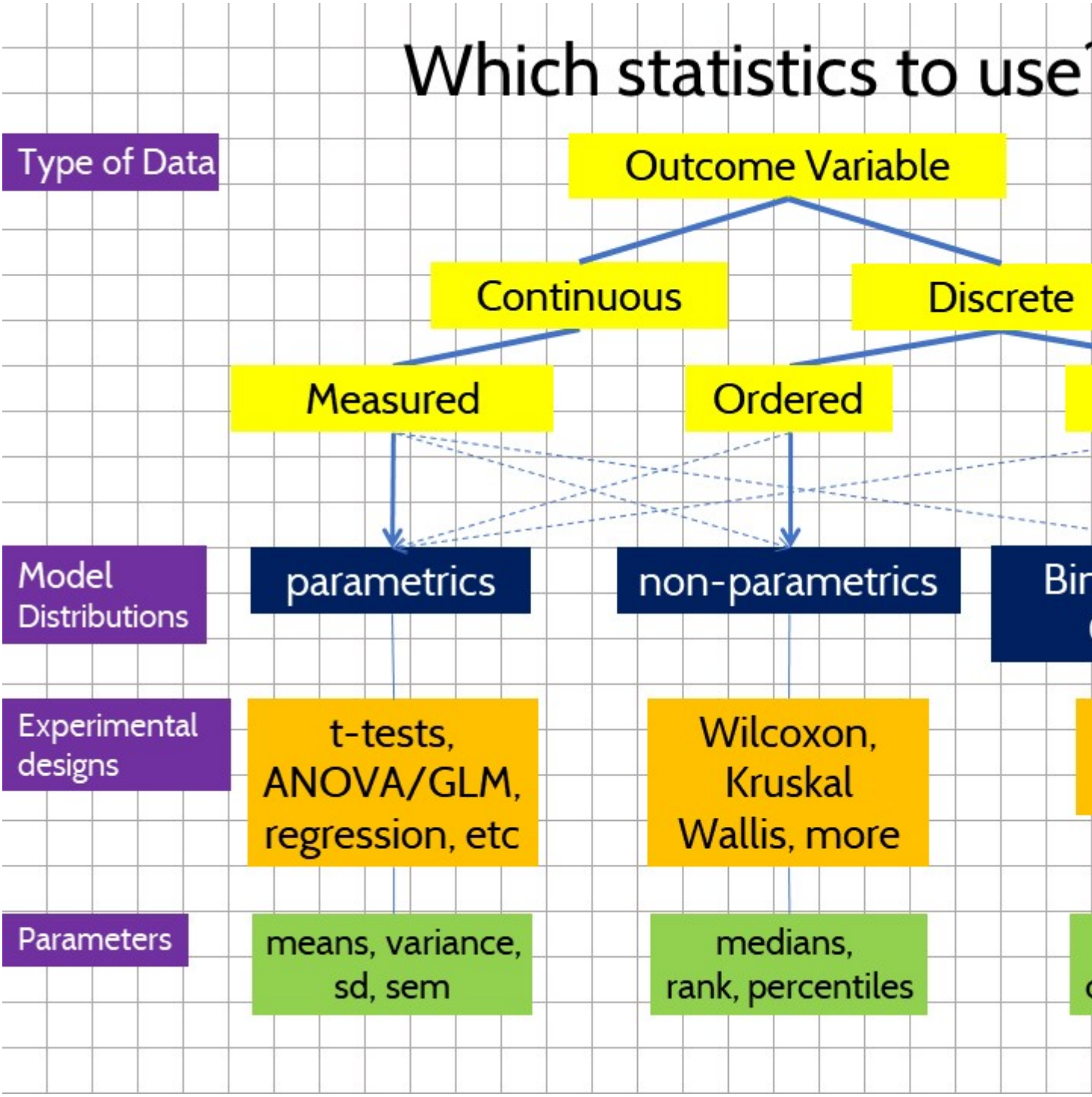


Figure 8.1: The type of data dictates how it should be modeled.

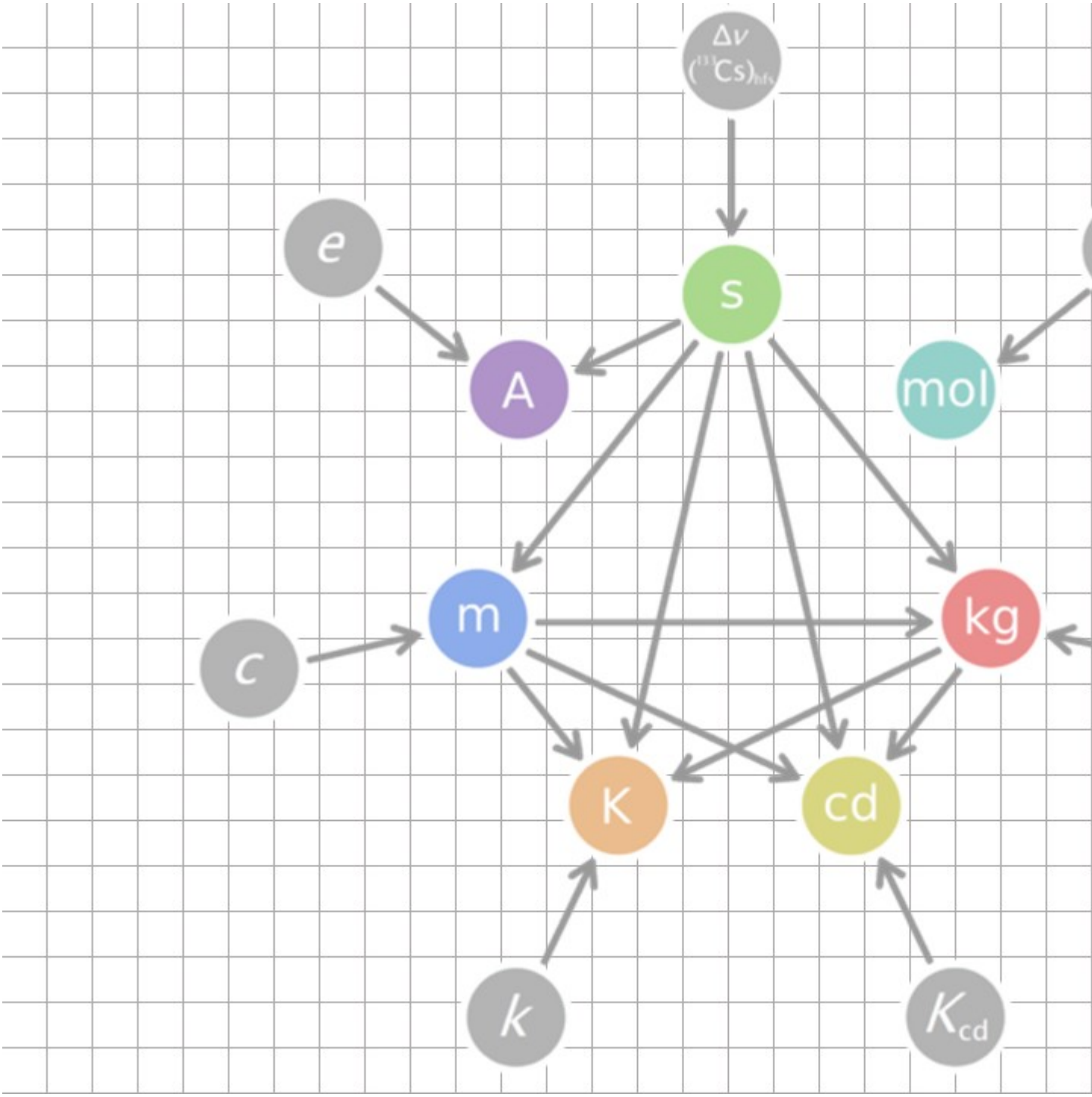


Figure 8.2: The seven SI units