جامعة نيويورك أبوظبي

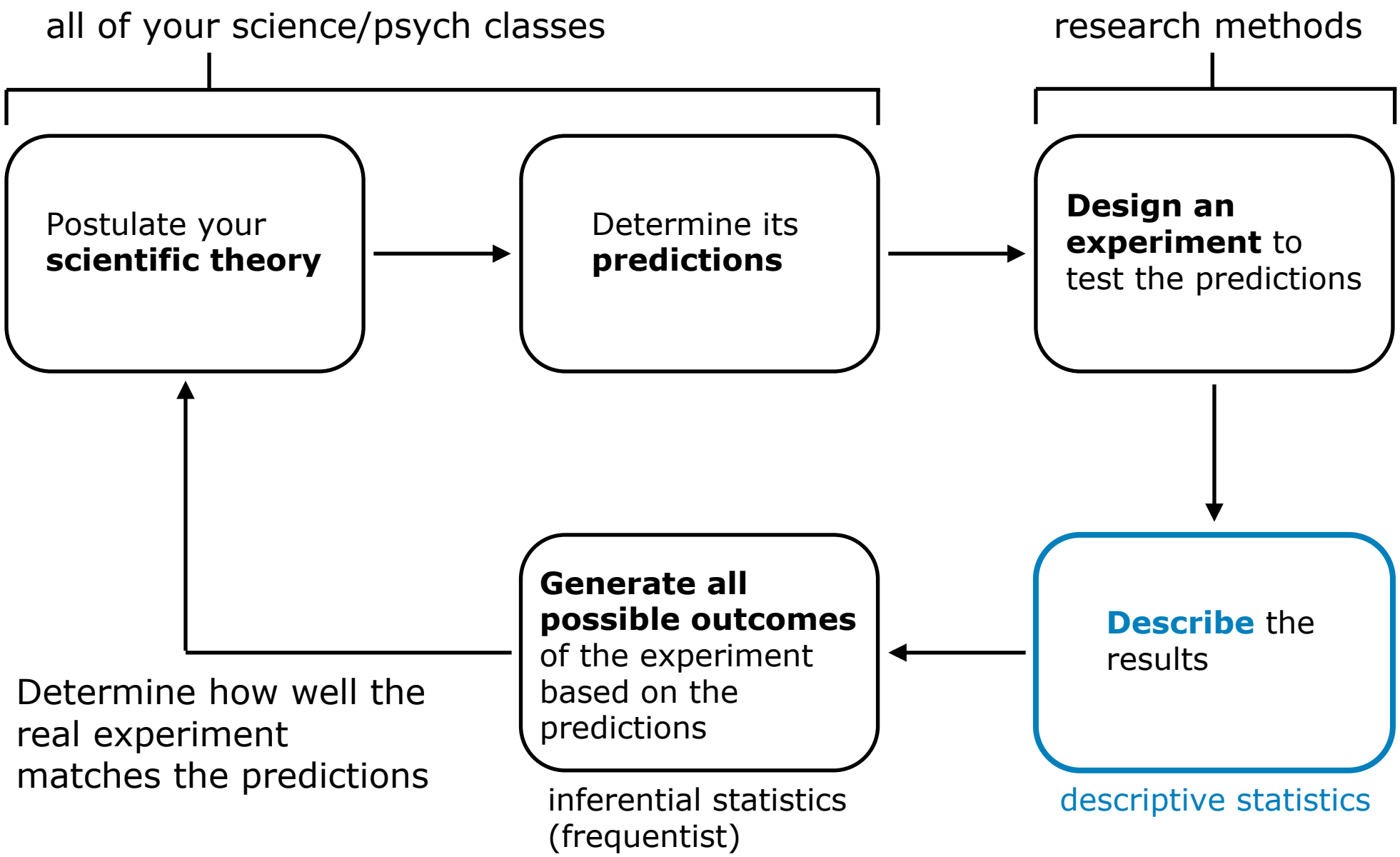
NYU | ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

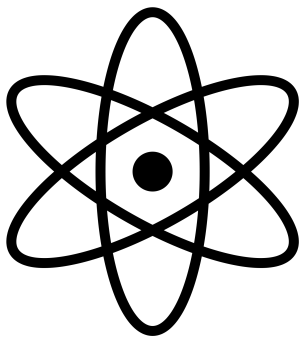
Class 6: Describing our results - putting it in practice

Prof. Jon Sprouse
Psychology

Putting it all together: describing our data



Disclaimer



The practices of the various subfields of psychology will sometimes differ from each other in the details of descriptive analyses.

Similarly, the practices associated with different data types can also differ in the details: survey data, reaction time data, electrophysiological data, hemodynamic data, etc.

What we can do in this class is discuss a general framework for describing (and exploring) your data. It will be up to you to learn the finer details of your chosen subfield and data type when you do your own studies.

General steps for descriptive data analysis

1. Data wrangling

Getting your data into a format that you can use for your data analysis.

2. Plot the distributions

Always check to see if your experiment worked the way that you expected!

3. Decide how to deal with outliers

These are data points that lie outside the range that you expected.

4. Plot the means of the conditions

This is the plot that will tell you, and your readers, what you found.

5. Plot a measure of variability as error bars

Never plot means without error bars showing variability!

Data Wrangling

(Just a summary of important steps - a deep dive goes way beyond the scope of this course!)

Data wrangling

Data wrangling is a term that has arisen in the last 10 years or so in data science. It refers to all of the steps you need to do get your data into a format that you can use for your data analysis, such as:

1. Importing your data into R: use read_csv()
2. Formatting the data — typically into long format or wide format.
3. Organizing the factors/levels (and creating any new ones that you need).

You can do all of this in base R. But, the dominant tool for this is the R package **tidyverse**: https://www.tidyverse.org/. It is a set of functions designed by data scientists specifically to make data wrangling even easier.

Some free, online books that teach data science with R and tidyverse:
https://bookdown.org/f_lennert/introduction-to-r/
https://jhudatascience.org/tidyversecourse/
https://r4ds.had.co.nz/

The other option is to simply google your question directly, like "how do I plot a histogram in R". You will find answers, often on https://stackoverflow.com/.

Long format

The primary format for computer-aided statistical analysis is **long format**. At first, long format is not very intuitive, but you will very quickly learn to appreciate its logic.

rows: Each row is an **observation** or **trial** the experiment.

columns: Each column represents a **variable** (i.e., property!) of the observation. There are a lot of variables - the participant ID, properties of the participants (like age), the condition, the measurement, etc.

	participant	age	condition	measurement
trial 1	1	21	red pill	1
trial 2	1	21	blue pill	7
trial 3	2	19	red pill	4
trial 4	2	19	blue pill	5

Why is it called long format?

It is called long format because it grows longer as we expand the number of observations/trials in the experiment. (It also grows longer when we increase the number of participants, but all formats will do that, as we will see shortly.)

	participant	age	condition	measurement
trial 1	1	21	red pill	1
trial 2	1	21	blue pill	7
trial 3	1	21	green pill	3
trial 4	1	21	orange pill	2
trial 5	2	19	red pill	4
trial 6	2	19	blue pill	5
trial 7	2	19	green pill	2
trial 8	2	19	orange pill	3

Wide format

When humans enter experimental data into a table, they tend to do it in **wide format**. It is a very intuitive format for data.

rows: Each row is a **participant** (or whatever the relevant major experimental unit is).

columns: Each column is a property of the participant (or experimental unit). This can be the participant's inherent properties like age, or the participant's responses to each trial in the experiment.

	age	**trial 1**	**trial 2**	**trial 3**
participant 1	18	2	7	6
participant 2	22	2	6	5
participant 3	23	3	7	4

Why is it called wide format?

It is called wide format because it grows wider as we expand the number of observations/trials in the experiment.

	age	trial 1	trial 2	trial 3	trial 4	trial 5	trial 6
participant 1	18	2	7	6	6	3	4
participant 2	22	2	6	5	6	2	1
participant 3	23	3	7	4	5	3	5

Now we can see the naming scheme. Both long and wide format get longer if you add participants. But if you add trial/observations, long format gets longer, and wide format gets wider.

Choosing a data format

95% of the time, the correct format for your data is **long format**. It is also the format that nearly all of the functions in the tidyverse require.

So why do we even talk about wide format?

1. Data is often given to us in wide format. It is intuitive, so humans often store data this way. You should store your data in long format. But be aware that new data may come to you in wide format.

2. There are some scenarios where wide format is more convenient. There are some base R functions that we might use, like cor() (for correlations) that work with wide format. And sometimes it is easier to calculate a new variable from existing variables if the data is in wide format (though, it is not strictly necessary). So, you may find that you use wide format from time to time do calculations.

Learning to convert from wide to long

Tidyverse makes it very easy to convert between the two types of formats:

pivot_longer(): Converts from wide format to long format.

pivot_wider(): Converts from long format to wide format.

But before you jump in to using these convenient functions, I strongly recommend that you try it two other ways first:

1. Try doing it by hand in excel using copy and paste. This is a bit slow, but it really helps you to see the relationship between the two formats.

2. Try doing it in R using base functions (not tidyverse functions). This will take a decent number of lines of code. But it is a good exercise. The trick is to find R functions that mimic the steps you did by hand in excel. It teaches you that anything you can do in excel by clicking and pasting you can do in R. This is very freeing. It is the first step toward really adopting R as your primary data analysis tool!

Formatting factors

The final step is to make sure all of the factors in your data set are formatted correctly.

1. Check the types of each factor - character, integer, numeric, etc. Make sure that R guessed what you wanted correctly. You can use typeof() to see this. If the factor is not the right type, change the type using functions like as.numeric(), as.character(), etc.

2. Check the names of the factors. These will show up in your plots as the default names of the axes. You can change them with rename().

3. Check the names of the levels of each factor. These will show up in the legends of your plots. You can rename them with fct_recode().

4. Check the order of the levels of each factor (you can use levels() to make R print them in their current order). R will order the levels alphabetically by default. That is rarely the order you want. The order matters for plotting, and as we will see later, it matters for statistical tests. So be sure to set the order that you want. You can change the order with fct_relevel().

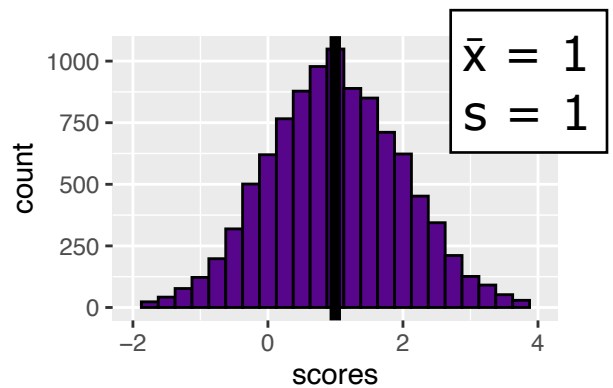
Plot the distributions

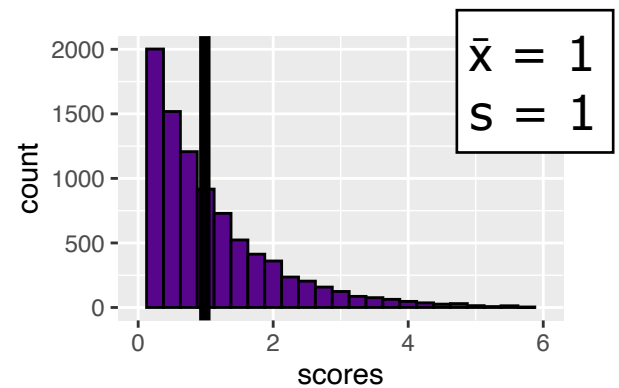
Always look at your data!

Always look at your data. Don't just calculate statistics like the mean and standard deviation without also plotting your data. Don't just run statistical tests without also plotting your data. This is for **quality control**. This will show you if your experiment worked the way you wanted.

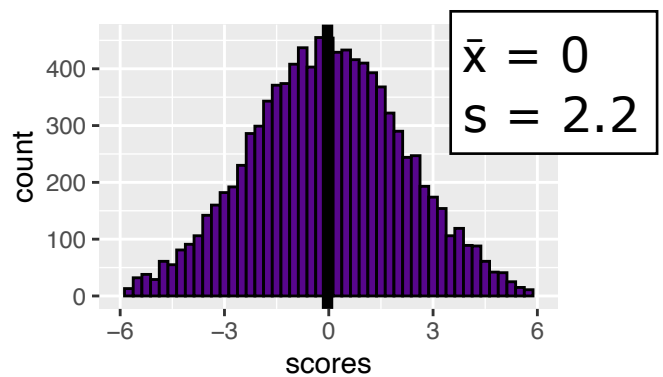
It is not enough to simply look at descriptive statistics like the mean and standard deviation. Those can be identical between distributions with very different shapes:

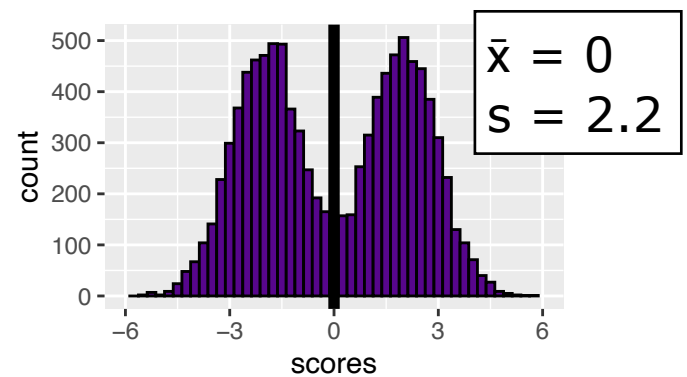
Same mean and sd, different distribution:





Same mean and sd, different modality:





Creating a useful distribution plot

1. Remember, bar plot for discrete variables and histogram for continuous variables.

2. For distributions, we place the outcome/dependent variable on the x-axis. But for mean plots (what we make later), we put the outcome/dependent variable on the y-axis. The logic is that we put the variable of interest on the y-axis — for distributions, that is the count.

3. For histograms, you must choose a bin width. The ideal is for there to be one that is meaningful for your theory. But if there is no obvious choice, you can try several different ones to see how it changes the distribution shape.

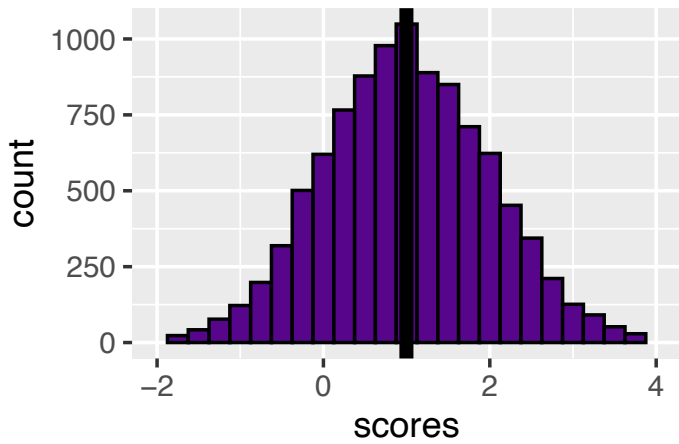
4. Set your x-axis and y-axis scales to scales that are meaningful for your variables, like 0 to 100 for a test score. If you don't set the scale, R will select a scale that is based on the range of your data set. This can distort things, so it is often better to select a meaningful scale.

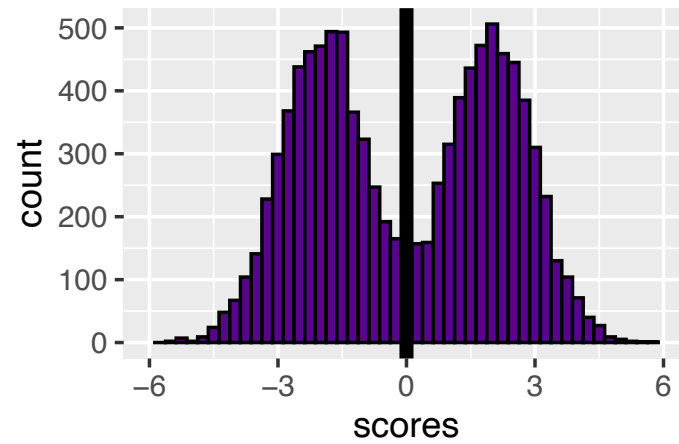
5. Set the labels to be helpful — the axis labels, the tick labels, etc.

6. Consider adding the mean and median as vertical lines - over time you may not need to these to know where they are, but they are helpful nonetheless.

Look at the shape!

The most important thing to look for when it comes to shape is unimodality versus bimodality.



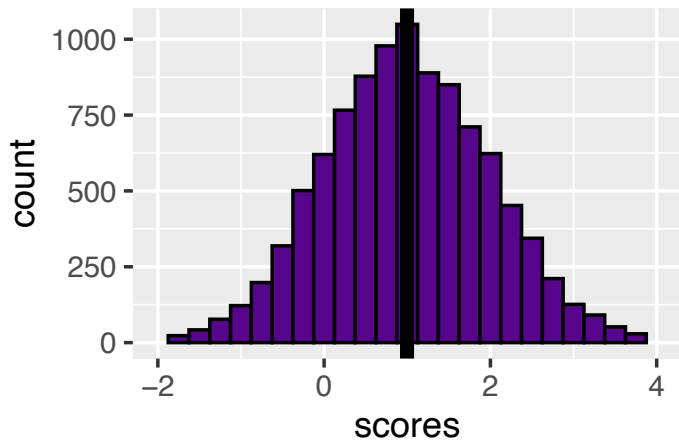


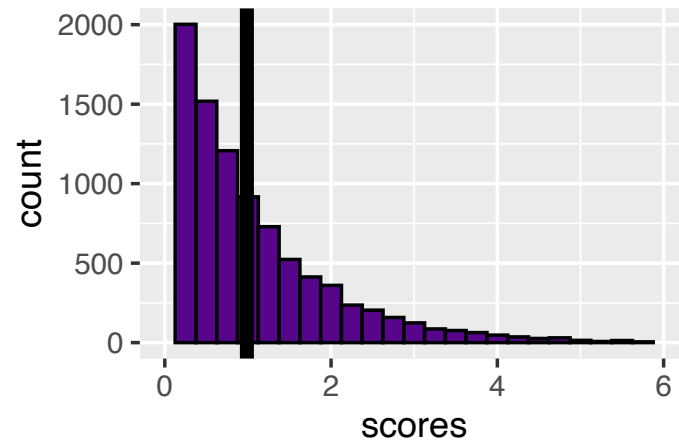
If there is evidence of bimodality, your sample may have been drawn from two different populations (instead of one population, as you assumed). The bimodal distribution above is a mixture of two distributions - one with a mean of -2 and one with a mean of 2.

Bimodality tells you that you missed an important difference between the members of your sample. They are responding to your condition differently. This a **confound** that you will need to eliminate.

Look at the shape!

A second, less important, question is whether there is skew in the distribution.





This doesn't matter all that much if the skew is mild; but if you have an extremely skewed distribution, you may need to be thoughtful about using the mean as your measure of central tendency.

The mean of a skewed distribution is impacted by the extremes in the tail. If you use the mean to identify the "effect", it is possible that the effect is driven by the extreme values alone.

Distribution plots are typically not published

Distribution plots are typically only about quality control.

You create them for yourself. They tell you if your experiment worked the way you expected. They help you see if you missed any confounding factors.

You will typically not publish distribution plots in the papers that you write. The only time you will publish distribution plots is if your theory makes specific predictions about the shape of the distribution (that is rare), or if something odd happened in your experiment and you want to report the odd event (that is also rare; in typical cases you will correct the experiment instead).

I say this because you won't see distribution plots very often as you read scientific papers. But you should **always** make them in your own studies. All careful scientists do this, but you don't see it. Please don't assume that you don't need to create distribution plots just because they aren't published.

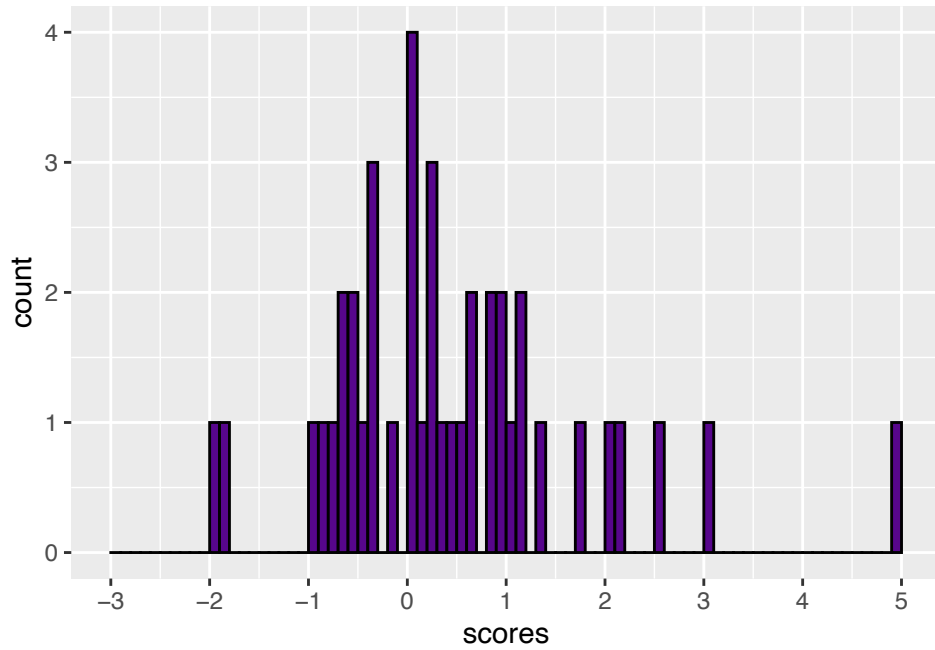
Decide how to deal with outliers

Identifying outliers

An **outlier** is an experimental unit (either a participant or a judgment) that is substantially different from other experimental units. Outliers add noise to your data, which can impact the inferential statistics that you will run later in this course.

The problem with outliers is that there is no hard and fast definition of what counts as an outlier.

To see this, take a look at this distribution of 41 observations. Which, if any of these observations would you call an outlier?



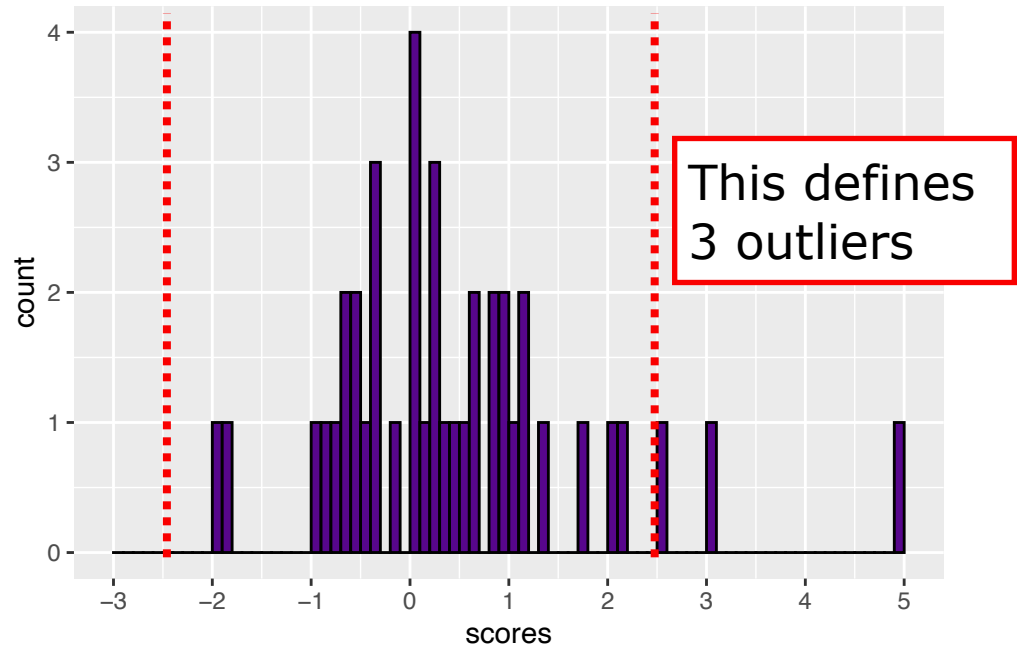
Because there are no hard and fast rules, the best we can do is talk about rules of thumb. These are suggestions. They are not actual rules. And, as always, you need to pay attention to the norms of the subfield of psychology that you are working in. The rules of thumb may be different!

A common rule of thumb -“fences”

The most common approach to outliers is to define “fences” - values on either end of the distribution that serve as a boundary, just a like a fence in the real world. Any value beyond the fence is considered an outlier, and any value within the fence is considered part of the distribution.

So how do we define a fence? The most common approach is to set the fences as some number of standard deviations away from the mean.

A a concrete example, we can set the fence at 2.5 standard deviations away from the mean. (The specific values chosen may vary by field.)



Fences are symmetrical around the mean (at least for symmetrical distributions). So we are cutting out both high and low outliers.

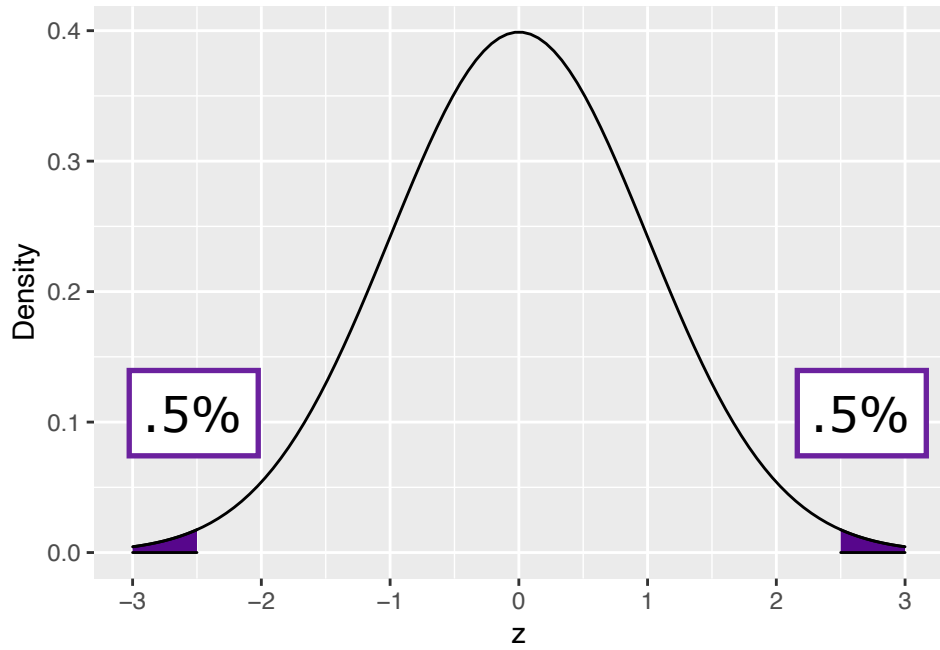
What are fences doing?

The fences are defining a portion of the tails of the distribution as “outliers”.

We know a lot now about what this means — we can identify the percentage of the distribution that is now defined as an outlier.

Based on the theoretical standard normal distribution, we know that z-scores (i.e., standard deviations) at +2.5 and -2.5 identifying percentile ranks of 99.5%. So each tail contains 0.5% of the distribution.

This means we are defining the most extreme 1% of scores (0.5% in each direction) as outliers!

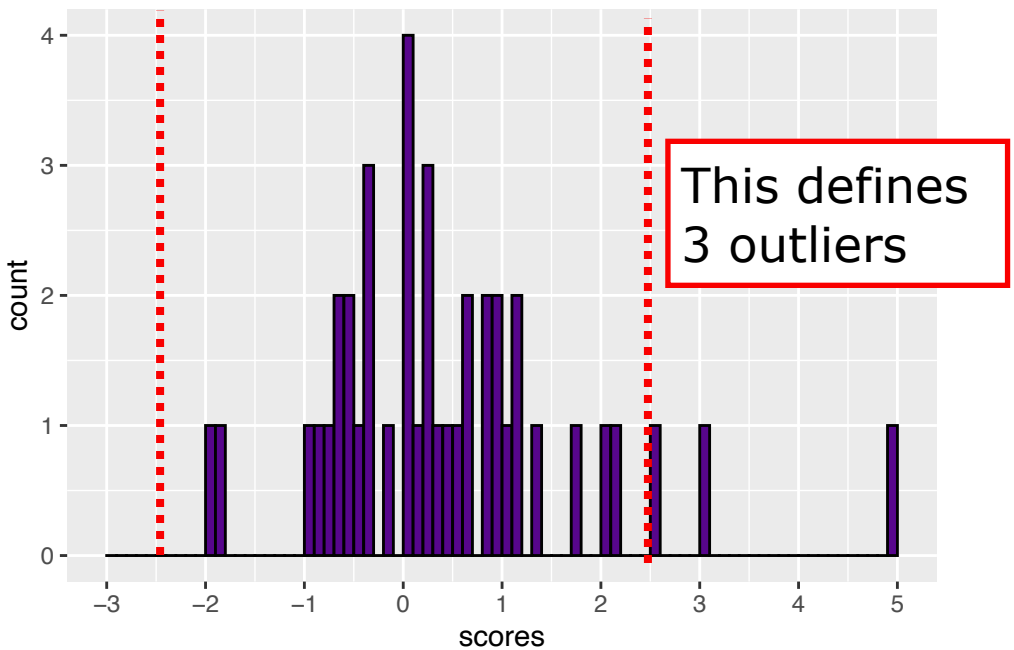


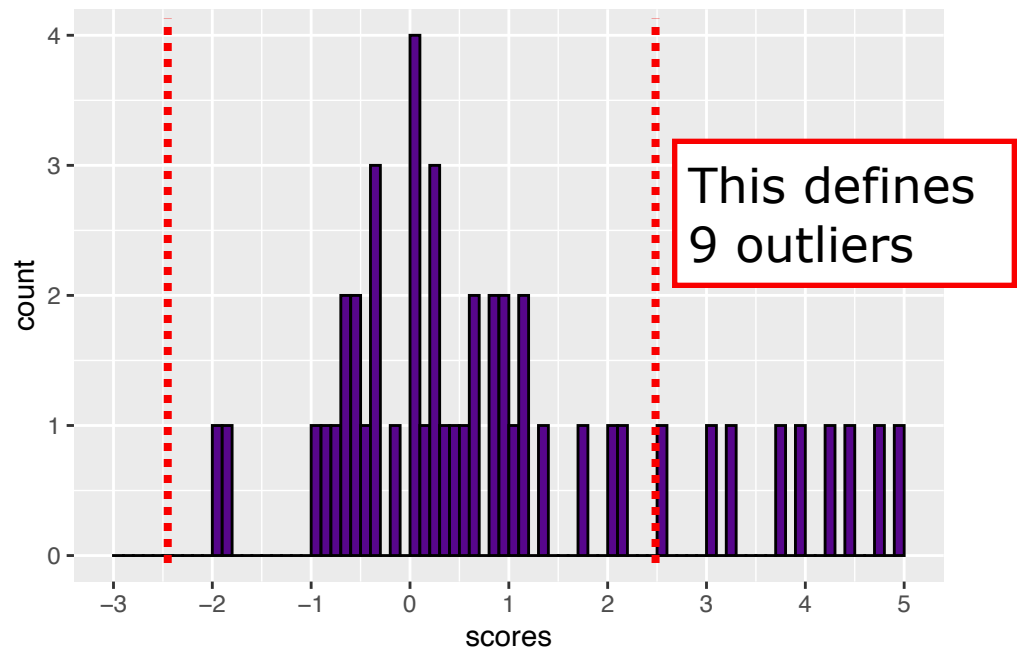
That is a very plausible approach to take. But please note that different subfields may set different criteria — such as 2 standard deviations (cuts 2.5% in each tail for a total of 5%), or 3 standard deviations (cuts only .3% total).

Caution: be systematic!

Whatever approach you choose to identifying outliers, the most important thing is to be systematic. Choose a criterion and apply it systematically. Don't cheat.

Imagine that you have two conditions in your experiment. You choose +/-2.5 for the fences for the one on the left.





You **must** use the same criterion for outliers for the one on the right. You can't look at it and say "oh, well, I don't want to lose 9 data points" and then use a different criterion for this condition. The trick here is that there should be a reason for the decision that you make.

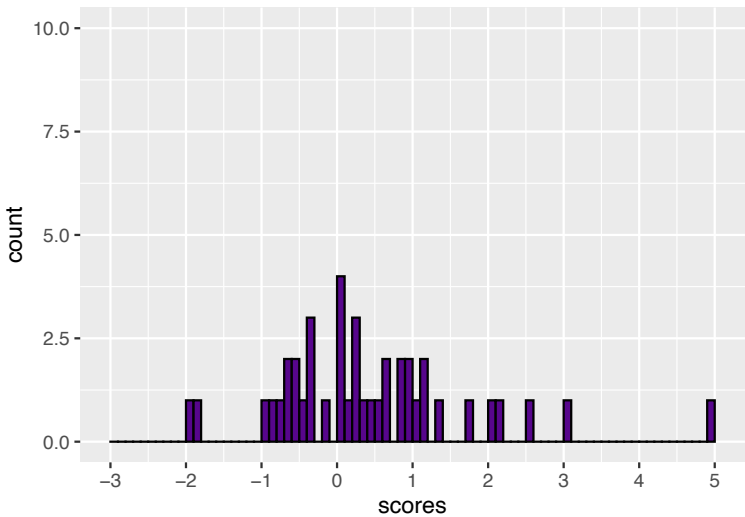
Strategies for dealing with outliers

Once you've identified outliers, the next step is to decide what you want to do about them. There is a wide range of possibilities. Here I will tell you about three that are very common.

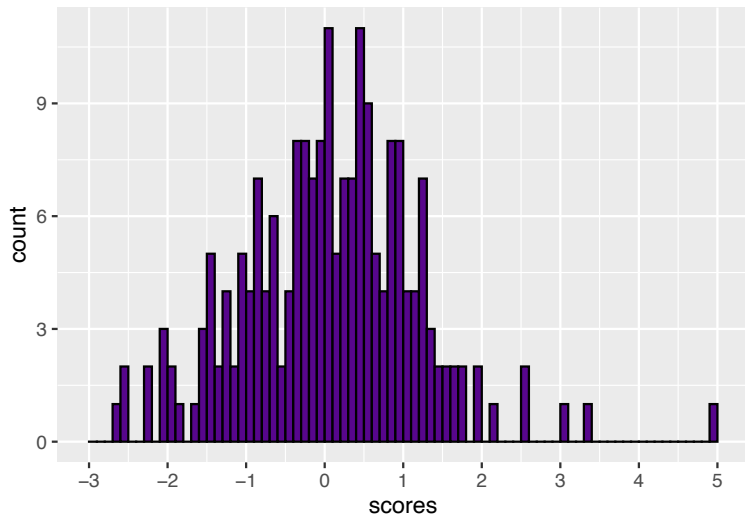
Each comes with a certain amount of risk that you will introduce patterns into your data that aren't really there. So I have color-coded these options according to that risk.

1. **Do nothing.** The best option is to run an experiment with a large enough sample size that you don't have to worry about the occasional outlier. We will talk about this when we discuss statistical power later in the course.

40 observations

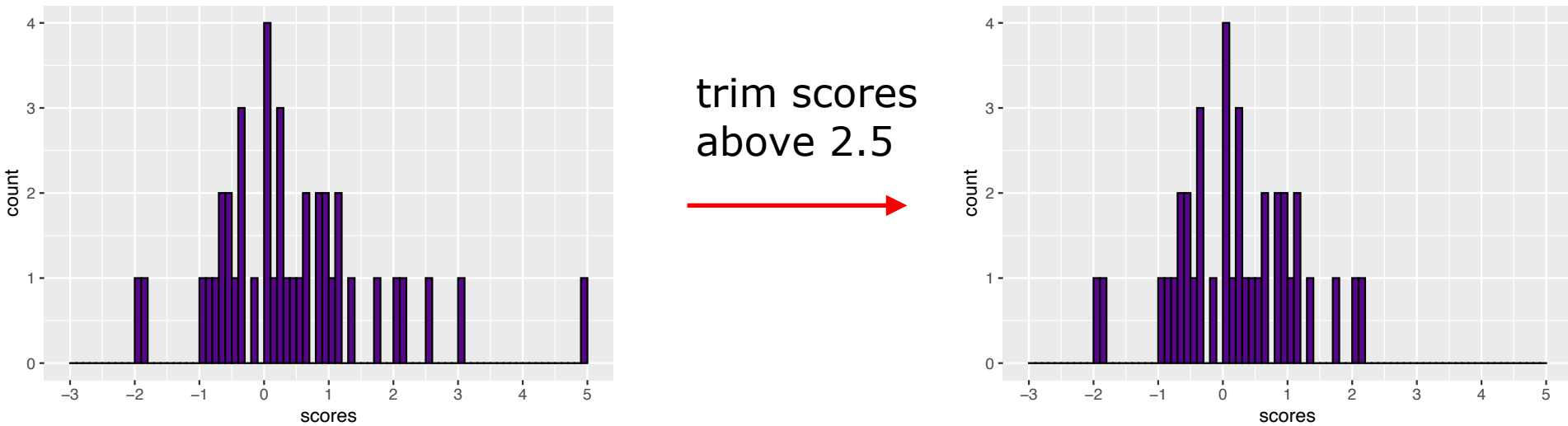


200 observations



Strategies for dealing with outliers with outliers

2. **Trim the data.** You can also look at the distribution of judgments for each experimental condition, and **remove** outliers. We call this trimming the data.



Trimming is the most direct way to deal with outliers. It is also the riskiest. You are directly changing your data. You are looking at your results, and choosing to reshape them based on what you think the should look like. That runs the risk that your personal biases could impact the data.

Use caution while doing this. Be sure that you can **justify** your criteria. And be sure that you are applying the criteria **systematically**.

Strategies for dealing with outliers

3. **Use gold-standard trials to identify uncooperative participants.** An uncooperative participant is one who is not doing the task the way you intended. The goal of a gold-standard trial is to identify those participants by giving them trials that 100% of cooperative participants will respond to identically.

What counts as a gold-standard trial can vary from experiment to experiment. So this takes some creativity and some pre-testing (to be sure that all cooperative participants will get it correct).

But the logic behind this is less risky than data trimming. In this case, you will be removing entire participants - that means their responses to all of the conditions in your experiment. So there is less risk that you will reshape one condition and not the others. All conditions should be impacted equally.

Furthermore, you aren't looking at the data that you care about - the critical conditions. You are only looking at data you don't care about - the gold-standard trials. So there is less risk that your personal biases about the study will impact the results.

These are not exhaustive

1. **Do nothing.**
2. **Use gold-standard trials to identify uncooperative participants.**
3. **Trim the data.**

There are a wide range of strategies for dealing with outliers. Here I just wanted to tell you about three fairly common, and very distinct, strategies out there. These are a good starting point.

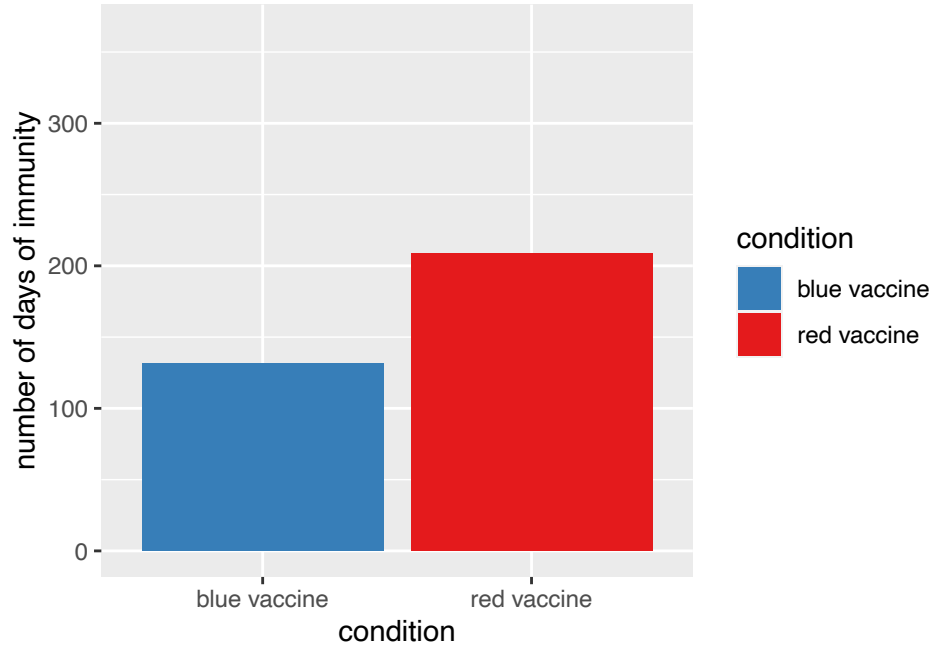
But as you learn about the subfields of psychology (or science more generally), you will learn the specific methods of those fields.

Plot the means of the conditions

The first step is to plot your means

This experiment has two conditions - a blue vaccine and a red vaccine. It studies how many days participants show immunity to COVID after taking this vaccine.

Because we only have two conditions, we will use a bar plot to plot the means. (Later in the course we will see line plots and scatter plots.)



<u>Rules for interpretable plots</u>

Independent variable on the x-axis, dependent variable on the y-axis. Add labels that clearly indicate what the axes are showing.

Choose scales for your axis that are meaningful for your measures. Don't "zoom in" to make effects look larger. Include the natural boundaries of the scale.

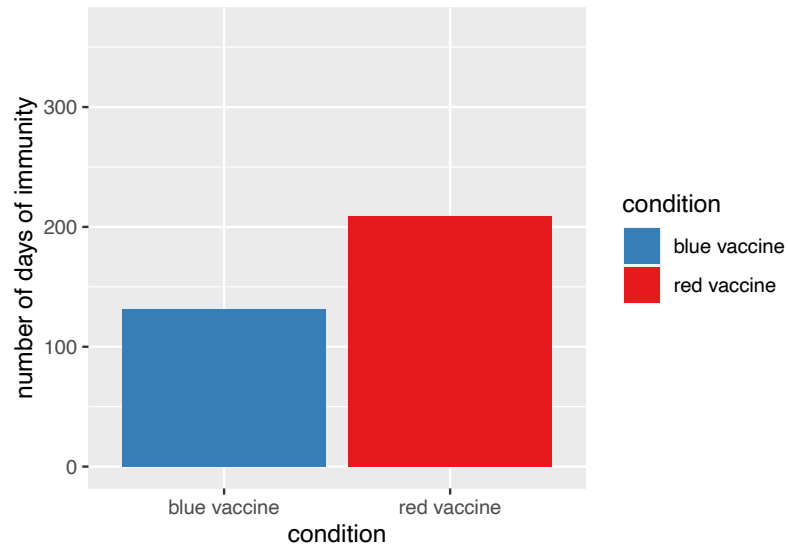
Only add color if it will help the readability. And make sure you use it in a way that aligns with how humans think. (Don't make the blue condition red!)

Always add error bars!

(In other words, add a measure of variability to your central tendency.)

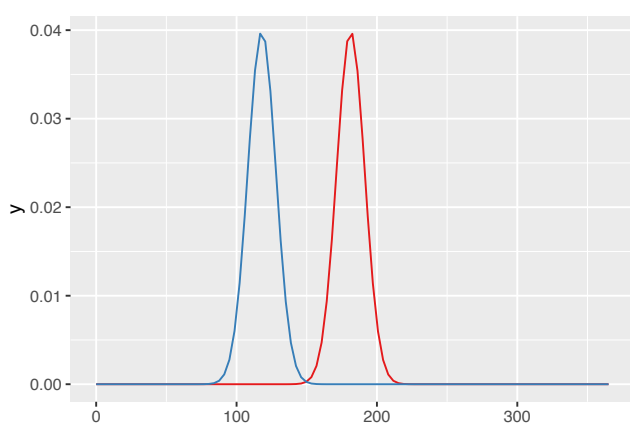
The problem with this plot is that it tells us nothing about variability

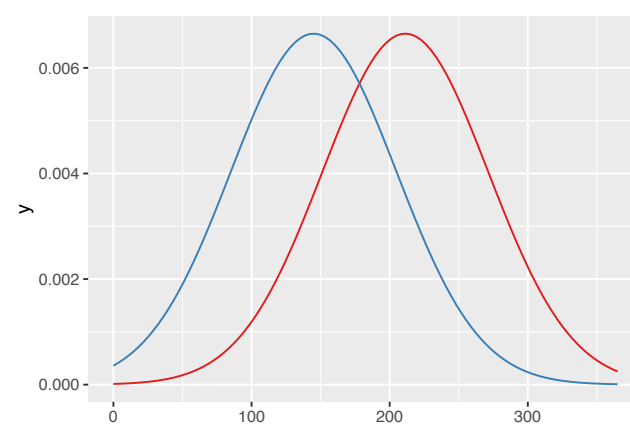
This plot only shows us the central tendency of the distributions of the two conditions.



It is possible that these come from very narrow distributions, or very wide distributions. We can't tell from central tendency alone.

If the distributions are narrow, it would suggest that the two vaccines are very different. If the distributions are wider, it would suggest that they are only mildly different. The means alone don't tell us this!

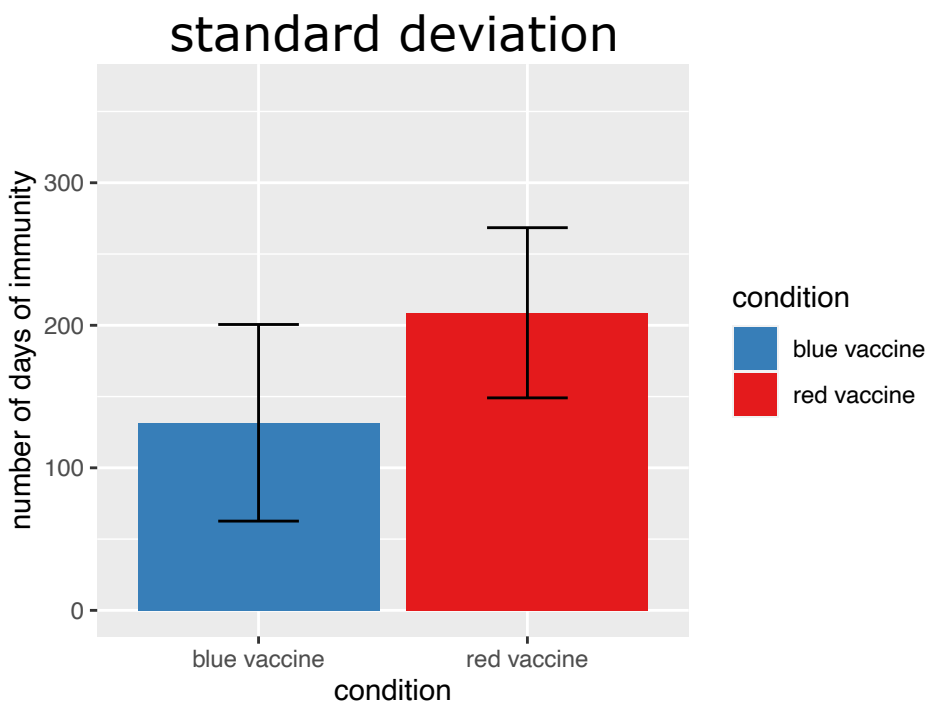


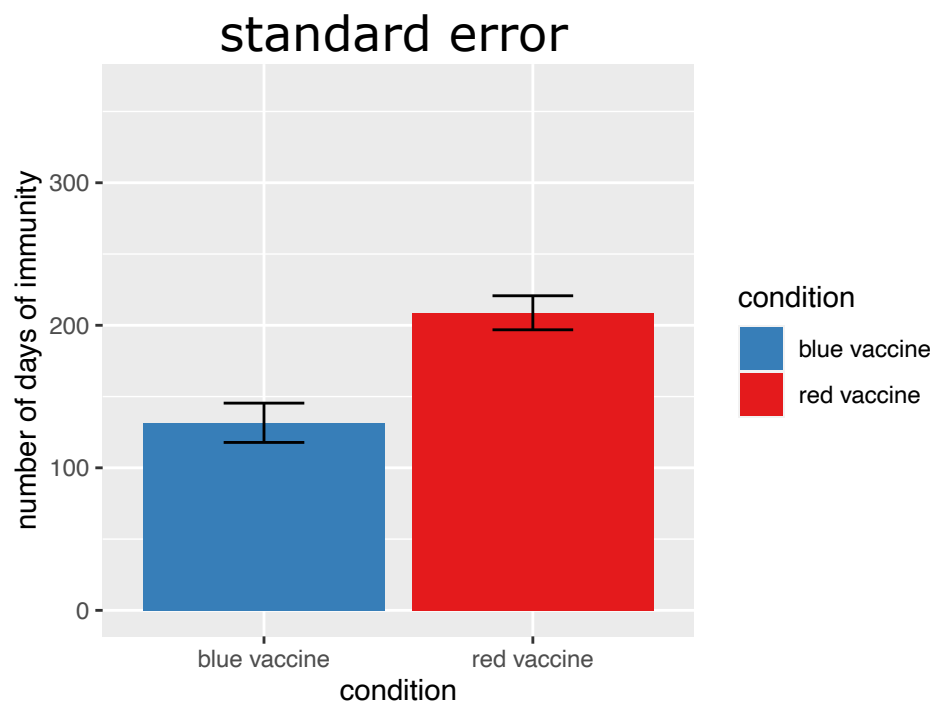


The solution is to add error bars

Error bars give us an indication of the variability that we should expect around the means.

At the moment, we have two measures in our toolkit that we could use: **standard deviation** and **standard error**.

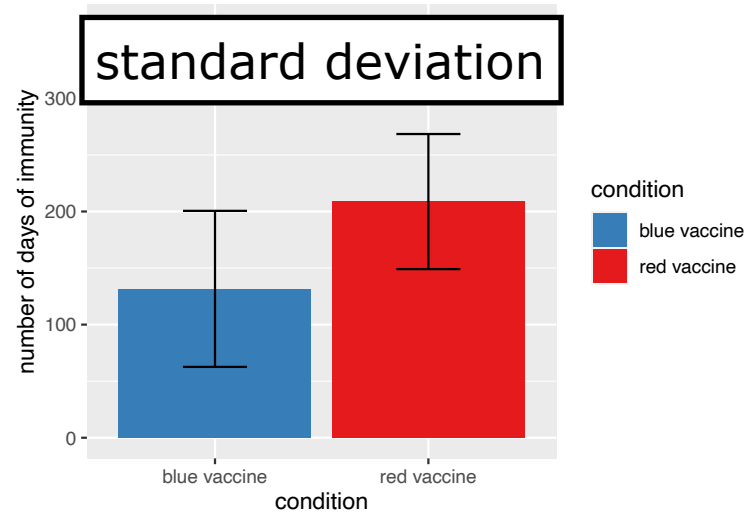




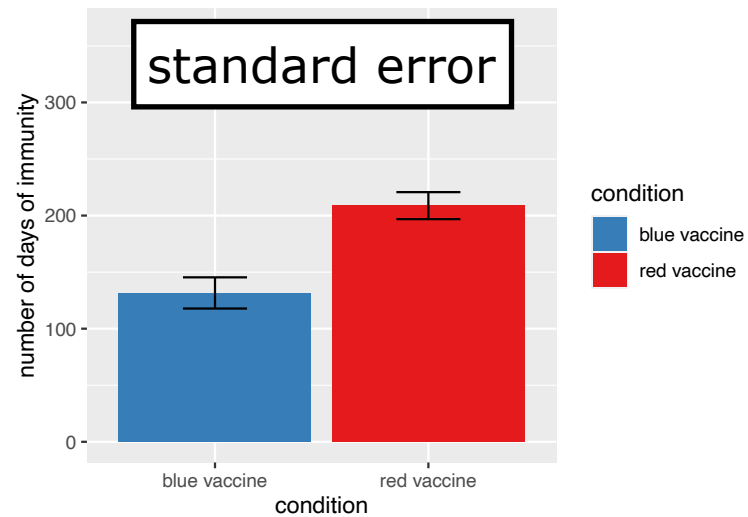
In both cases, we draw the error bars such that they extend **1 unit above** and **1 unit below** the mean.

The choice between standard deviation and standard error is meaningful

Standard deviation is a measure of the **variability of individuals** (in a sample or in a population). So, if you show us this, you are showing us how the individual observations vary around the mean. You are telling us about the individual variation within the samples that we collected.

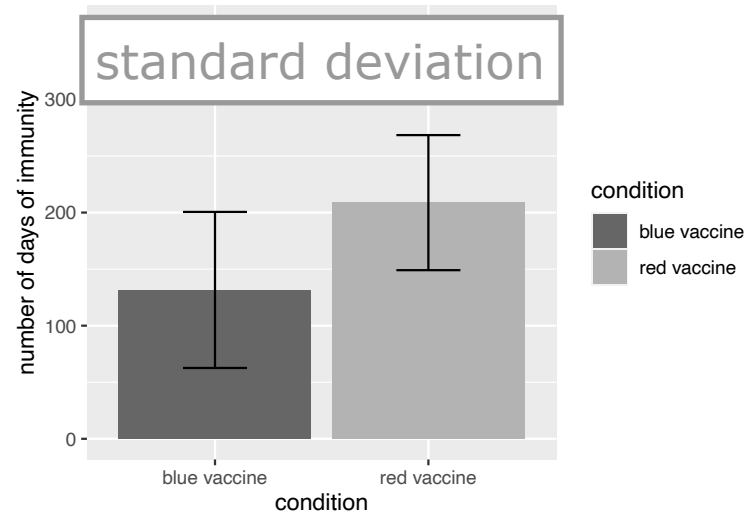


Standard error is a measure of the **variability of means** (in the sampling distribution of the mean). So, if you show us this, you are showing us how the means of this particular sample size vary around the population mean. You are telling us how much we should expect means to vary if we repeated the experiment over and over.



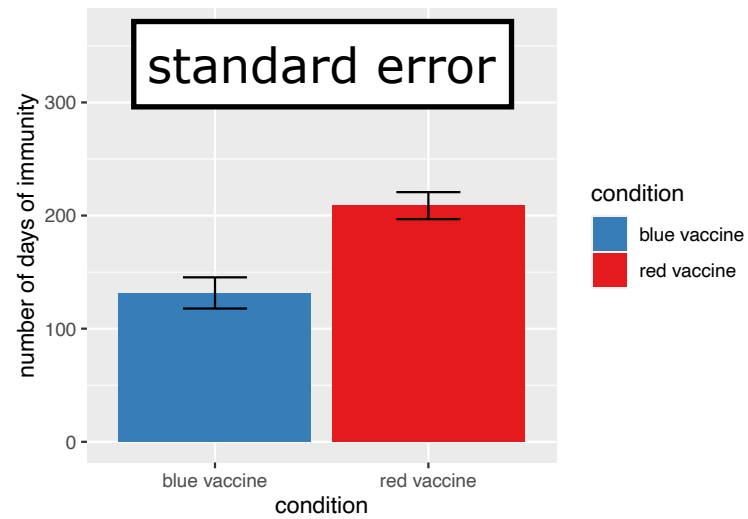
The most common choice is **standard error**

Standard deviation is a measure of the **variability of individuals** (in a sample or in a population). So, if you show us this, you are showing us how the individual observations vary around the mean. You are telling us about the individual variation within the samples that we collected.



Standard error is the most common choice, at least in areas of psychology that I work within.

This makes some sense. SE lets us begin to make inferences about the sample means. For example, if the error bars of these means overlap, it tells us that they are within 2 SE of each other - so perhaps within the bounds of random variation among means.



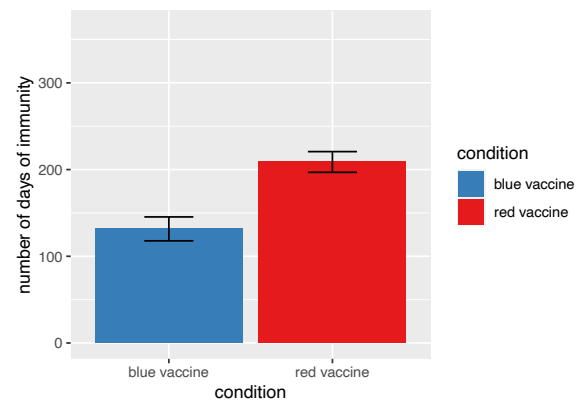
One piece of advice about write-ups

The description of the results comes first

Write-ups are covered in more detail in research methods and in advanced and capstone courses. But it is important to review this a bit in statistics too.

The biggest piece of advice that I can give you is to always remember to **describe your results** before jumping into the inferential statistical tests.

Begin your results section with either a plot of the means and variability, and describe what you see in the plot. These are your results.



The inferential analyses (t-tests, ANOVAs) are not your results. They are analyses. Report them after.

I am saying this because, as humans, we will tend to want to skip to the more complicated parts. And it is very common to view the inferential analyses as more complicated than these descriptive analyses (with all the t's, and F's, and p-values…). But I want us to build good habits from the beginning. Let's fight this urge, and always remember to describe our results.