Section 8
Statistical Inference for Two Proportions

8.1 – Validity of a Statistical Study

Introduction

So far, we've developed the theory for conducting hypothesis tests about a single parameter. While there are some use cases for this, this method requires us to pick a null value for the sake of making a comparison of our data to a hypothesis, or there has to be a reason to use that particular null value. In the examples from last lecture, we used $p = 0.553$ to represent how many households had no children from the 2010 US Census, and we used $\mu = 14.5$ oz for the Starbucks coffee example because this is a standard set to ensure drinks don't overfill the cups. These examples seem to come about from previously known population parameters, or an existing threshold set for a particular scenario.

But what if we wanted to compare data across two groups or populations? Consider the dolphin therapy study we looked at in class this week where subjects with depression were sent to Honduras. Half were assigned randomly to go snorkeling with dolphins, and the other half did all of the same snorkeling activities but without the presence of dolphins. If we were to evaluate the effectiveness of dolphin therapy just based on a single sample of those that swam with dolphins, what would we compare our sample proportion to? Test against a null of $p = 0.5$? Maybe being better than 50% effective is reasonable for this context, but some conditions may be much harder to treat. What if this was done on subjects with severe depression symptoms? For an even more extreme case like cancer, we might be thrilled to find a treatment that's 10% effective. This is where it helps to have some baseline or control group for making comparisons, so that we can better isolate the effectiveness of a condition. For evaluating the effectiveness of dolphin therapy, we wouldn't be able to know if the presence of dolphins affected depression unless we had a control group to compare to – that is, a group that experienced all the same conditions but did not swim with dolphins.

Additionally, comparing two populations broadens the types of inferences that we can make. So far, the main method of inference we have learned is the _____ inference, which allows us to generalize results from our sample to the population. But when comparing two groups or variables, we may also be interested in drawing different kinds of inferences, potentially a _____ inference, that is, an inference that draws a cause and effect relationship. Before we dive into the statistical methods for analyzing these studies, let's first look at what allows us to draw these kinds of inferences!

Internal and external validity

When comparing two populations or groups, there are two ideal conclusions that we would like to draw:

- _____: being able to draw an inference from your sample to a larger population, that is, results seen in the sample apply to the population.
- _____: drawing a cause-and-effect conclusion – that is, the explanatory variable caused the outcome seen in the response variable.

Example: Suppose that researchers were interested in determining the effects of malnutrition in child development. Specifically, they wanted to see if it affects their height. How would we conduct a study that could determine a cause-and-effect conclusion (i.e. malnutrition causes you to be shorter) if there were **no ethical constraints**?

Example: What kind of study would you conduct now if there were ethical constraints?

These examples highlight two main study designs:

- _____: a study design where researchers have control of assigning individuals /units to their groups. Typically designed to determine causality.
- _____: a study design that can only observe groups that individuals/units are self-selected into or are intrinsically part of that individual/unit. Can identify only associations between variables.

It seems like experiments are the ideal way to conduct a study, as it best isolates your explanatory variable from other potential confounding variables. (We'll see an example of this impact later!) As we saw with the previous example, this can be ethically challenging to conduct in certain circumstances. In others, it may be impossible! Consider designing a study whose goal is to determine the impacts of autism on academic achievement among high schoolers. Here, the explanatory variable of autism is not something that can be randomly assigned. In other circumstances, it is often difficult and expensive to conduct an experiment, as controlling group assignment, monitoring compliance with the experimental condition, and incentivizing participation takes a lot of time and money!

We have discussed the first idea at length already, as this is the goal of taking a simple random sample of your population – so that the sample is most likely to be representative of the population you're studying. Of course, even perfectly random samples aren't always representative, you might get an exceptionally strange sample just by random chance. But without knowing every characteristic of your population (which would make looking at a sample rather pointless!), a randomly selected sample is often the best we can do.

When you have a study design that is centered around random sampling, this gives the study _____ validity, as the results go broader than just the sample you collected. Of course, as we discussed in previous sections, simple random sampling is difficult to conduct in practice. Any part of your study design that does not reflect this will introduce bias in some way. Some examples of these biases are:

- **Undercoverage bias:**

  Example:

- **Volunteer (non-response) bias:**

  Example:

- **Survivorship bias:**

  Example:

These biases represent threats to the generalizability of your study and should be reported as limitations of the statistical study you are presenting.

Notice that the last example on dolphin therapy did not leverage a method of random sampling. The subjects who participated in the study were volunteers who were recruited to be part of this study on depression and the presence of dolphins. What kind of useful conclusion can we make then if we didn't try to obtain a representative sample in any way? In this case, by the researchers using randomization to assign subjects to the two groups, we have done the best we can to make two groups as similar as possible. Thus, when analyzing the difference in depression between the two groups, we can isolate the source of this difference to the treatment, that is, the presence of dolphins with the one group that had a significantly greater improvement. By comparing groups that are as similar as possible in this way, we give our study _____ validity.

Regardless as to whether we conduct an experiment or an observational study (but especially for observational studies!), we need to assess the effect of other possible variables on the study. Specifically, we need to assess the impact of _____ variables. These variables are associated with both the explanatory/grouping variable and the response variable and are imbalanced within the explanatory variable.

> *Example:* Researchers want to investigate the effects of physical activity on the risk of cardiovascular disease. They compare two populations on the percentage of those who have cardiovascular disease: those who get at least 30 minutes of exercise daily and those who do not. What confounding variables might exist in this study?

In this case, we may want to _____ the results after we collect them. We may not know what variables could be confounding until we conduct the study, which is why collecting a great deal of demographic information from your participants is common in surveys or other observational designs. In this example, we might observe the following:

	Age < 40		Age ≥ 40		Total	
	CVD	No CVD	CVD	No CVD	CVD	No CVD
Sedentary	150 (10%)	1350	2000 (40%)	3000	2150 (33%)	4350
Active	210 (7%)	2790	375 (25%)	1125	585 (13.5%)	3915
Total	360	4140	2375	4125	2735	8265

In this table, we can see that CVD is far more present among those with sedentary lifestyles compared to those with more active lifestyles. However, when we break it down by age, we see that the overall effect is actually smaller within each group than it is when combining the groups. This indicates that the confounding variable is also related to the response, and that the observed difference in CVD cannot be fully attributed to one's physical activity. Even when stratified, physical activity had an impact within each group, but age is also seeming to have a big impact here too. This method can help to add to the external validity of your study, but always consider that other confounders that you haven't yet considered may be impacting your results too, and may be threats to making a causal claim. Here are some other aspects of studies that may impact causality:

- **Group selection (self-selection bias)**:

  Example:

- **Drop out differences**:

  Example:

- **Time/setting effects**:

  Example:

- **Independence**:

  Example:

8.2 – Statistical Inference for a Difference in Proportions

Theoretical background

To evaluate the difference in effectiveness for two populations or groups measured by a proportion, we need to understand how different our difference is relative to natural variation. This involves building a test statistic similar to what we did in the one population case. The general format of a test statistic is looking at your best estimate of the parameter $p_1 - p_2$, then dividing by the standard error of that estimate. First, we can quickly determine the best estimate for this parameter is the difference in sample proportions:

$$E(\hat{p}_1 - \hat{p}_2) =$$

Next, we then need to determine the standard error of our estimate. We can begin this by looking at the variance of our sample proportions:

$$\text{Var}(\hat{p}_1 - \hat{p}_2) =$$

Putting these two pieces together, we get a z-test statistic for testing two proportions:

$$z =$$

We can also use these two pieces of information to derive a confidence interval for the difference in two proportions, as shown below:

$$\hat{p}_1 - \hat{p}_2 \pm$$

Testing a difference in proportions by hand

Let's try an example using the equations we derived above.

> *Example*: Tolling on the bridge between Vancouver and Portland along I-5 is being proposed. To assess the differences in public opinion on tolling across state borders, a random sample of residents from Vancouver and a random sample of residents from Portland was taken. Of the 60 Portland residents sampled, 30 support tolling, and of the 50 Vancouver residents sampled, 20 support tolling. Do Portland residents support tolling more than Vancouver residents? Test at the $\alpha = 0.05$ level.

To conduct this test, we need to be sure that the assumptions are met for conducting this test. This primarily involves being sure that the z-test statistic's distribution is approximately normal. Those conditions that we need are as follows:

1.

2. (for HTs)

2. (for CIs)

R code for testing a difference in proportions

We can also approach computing this test using R. One method for doing this involves creating a table of data. For the problem above, you could envision the data appearing in a two-way table:

	Support Tolling	**Against Tolling**	**Total**
Portland			
Vancouver			
Total			

This is more likely seen when working with a categorical data set in R and extracting counts (see next example), but we can also build this up manually using a structure with the table function in R. For this to calculate the percentages correctly, it is imperative that the groups represent the rows, and the response variable represents the columns. To enter this data into a table we first create a matrix structure, then cast it as a table:

```
tolling = as.table(matrix(c(30, 30, 20, 30), ncol=2, byrow=T))
```

If you would like to give the rows/columns names, you can use the following code to do that, although this is purely for aesthetic purposes. The names default to A, B, … if these are not specified.

```
colnames(tolling) = c("Support", "Against")
rownames(tolling) = c("Portland", "Vancouver")
```

This creates the following table in R if you run just the `tolling` variable in the console:

```
          Support Against
Portland       30      30
Vancouver      20      30
```

With this structure created, we can use the following code to conduct the test (or generate a confidence interval!)

```
prop.test(tolling, correct=F, conf.level=0.95, alternative="greater")
```

When the raw counts are presented as in this example, a simpler way to compute this test is to just enter both of the counts and sample sizes into the same fields we would have done on a one-sample test. Since there are two samples now, we just enter in a vector of *x* values and a vector of *n* values. See the example below:

```
prop.test(c(30, 20), c(60, 50), correct=F, conf.level=0.95,
alternative="greater")
```

Our use of the argument "correct" does not imply we are telling the interval to be incorrect, we are just turning off something called the Yates continuity correction that R does by default. This continuity correction is used for helping to correct the normal distribution approximating a binomial for smaller sample sizes. If the sample size you are testing is too small, it is better to use a more exact test like Fisher's Exact Test. While we used an exact binomial test for the case of one proportion, there is more controversy about the use of an exact test for two proportions due to the complications introduced by an additional variable, and how it potentially produces *p*-values that are too large.

We may also be presented with data that don't come in tallies, but inside an existing data set. In this case, it's even easier to conduct this test, as we can create a table of the counts in one line. Let's try an example using a data set directly.

> *Example*: Blood donations from volunteers serve a crucial role in the nation's health care system, so public health researchers are interested if the likelihood of donating blood in the USA has increased over time. To investigate this question, researchers examined data from the General Society Survey (GSS) in 2012 and 2014. The GSS is a national survey conducted every two years, which determines its participants through a random sample of adults in the USA. Data from this study are summarized in the data file **blooddonations.csv**. Determine if there has been a significant increase in blood donations at the $\alpha = 0.05$ level. Confirm this result by computing a 90% confidence interval.
>
> ```
> b.table = table(blooddonations$year, blooddonations$donation)
> prop.test(b.table, correct=F, alternative="less")
> prop.test(b.table, correct=F, conf.level=0.9)
> ```

8.3 – Conducting Tests with Simulation

Experimental design

This week, we focused on analyzing the dolphin therapy study. As a reminder, here is the design of that study:

> *Example:* Researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subject's level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed substantial improvement (reducing their level of depression) by the end of the study. The data from this study are summarized below and can be found in the **dolphin.csv** file.

	Improved	**No Improvement**	**Total**
Dolphin	10	5	15
Regular	3	12	15
Total	13	17	30

How might we evaluate the validity of this study? Well, we can first recognize that this study had an **experimental** design, by **randomly assigning** participants to each group. In TinkerPlots, we created a process that took the existing variables and put them in their own devices, then re-assigned them to each other. The assumption of the null hypothesis was key in doing this – we could just take the outcomes (improved or no improvement) to represent the subjects, as if the null were true, dolphin therapy would be just as effective as the regular therapy, and the subjects would have the same outcome regardless of group assignment. We can replicate this process in R too!

In the previous sections, like with bootstrapping the kissing study in Section 6, we created the data vectors by using the `rep()` function to build all the necessary duplicates. We could do the same based on the table above, but we can also easily do this by using the vectors from the csv file! Let's read in that csv file and use the sample function to randomly assign each person (and their outcome) to a new group, just as we did earlier!

```
treatment_rand = sample(dolphin$treatment, 30, replace=F)
result_rand = sample(dolphin$result, 30, replace=F)
```

By sampling both vectors without replacement at the full sample size of 30, we are effectively just reordering each vector. To complete the random assignment process, we can now create a data frame out of these two vectors – assigning each entry of the vector to the corresponding entry in the other vector.

```
dolphin_rand = data.frame(treatment_rand, result_rand)
table_rand = table(dolphin_rand)
table_rand
```

```
treatment_rand Improvement No Improvement
       Control           6              9
       Dolphin           7              8
```

Using this data, we could examine the results of this randomized test. In class, we looked just at the counts of each – the randomized table I got above would show a difference of just 1 between the two groups. Since we've formalized this section to be about proportions, let's examine this as a difference in proportions instead! To do this, we can extract elements from the table using square brackets. The counts in the table above are indexed by column, so the counts 6, 7, 9, 8 correspond to indices 1, 2, 3, and 4.

```
p_control = table_rand[1]/(table_rand[1]+table_rand[3])
p_dolphin = table_rand[2]/(table_rand[2]+table_rand[4])
diff = p_dolphin-p_control
```

Note: Generally, proportions are best to use universally, as the groups you are comparing may be different sample sizes – larger groups will have larger counts naturally, but proportions or percentages are always comparable!

With this whole setup, we now just need to place this code in a larger for-loop and collect these differences in percentages many times! We can then evaluate how many times we got a difference of 46.67% (10/15 – 3/15) or greater!

```
diffs = rep(0, 1000)
```

```
for (i in 1:1000) {
  treatment_rand = sample(dolphin$treatment, 30, replace=F)
  result_rand = sample(dolphin$result, 30, replace=F)
  dolphin_rand = data.frame(treatment_rand, result_rand)
  table_rand = table(dolphin_rand)
  p_control = table_rand[1]/(table_rand[1]+table_rand[3])
  p_dolphin = table_rand[2]/(table_rand[2]+table_rand[4])
  diffs[i] = p_dolphin - p_control
}
```

```
hist(diffs)
abline(v=0.4666, col="red")
```

```
mean(diffs > 0.4666)
```

Remember that mean here to find a percentage seems a bit counterintuitive – but using the ">" operator is effectively checking all 1000 differences to see if they are greater than 0.4666 or not, and assigning 1 (True) or 0 (False). Since there are only 1's and 0's here, finding the mean is the equivalent as the percentage of 1's present – both are found by adding up 1s and dividing by the sample size. Thus, we have the *p*-value! Check to see if this lines up with what you found in our activity, and with what you would find in the additional practice problem for this section.

Observational studies

The key aspect of an experimental design in the simulation process is the random assignment process. Because the researchers used a random process to generate their data, we could replicate

this random process under the null hypothesis to see what might happen just by random chance. But what if we were looking at an observational study, where random assignment isn't possible?

Assuming that the observational study had a good design, it would use random sampling from each of the two populations being compared. While we can't conclude cause and effect from a design like this, random sampling does promote good external validity – that is, it would show that the results in our study are generalizable to the larger populations.

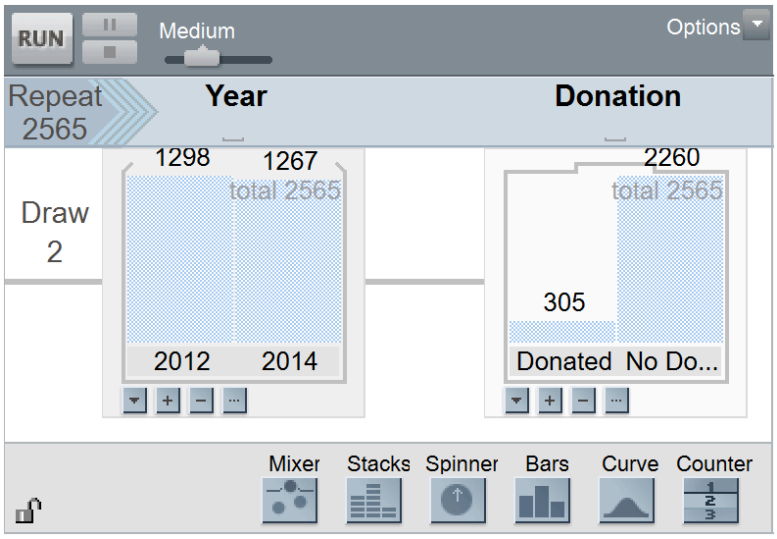
But how do we simulate what random sampling looks like? To accurately create a simulation that samples from our population, we would need to know what the population proportions would be for each population. But this isn't possible – and if we know what our population looks like, statistical inference wouldn't be necessary! However, we can make a guess as to what the population might be. This is what we did when we bootstrapped our data in Section 6, and assume that a random sample is the best representation we have for our population.

But one key difference here is that we aren't using bootstrapping to estimate, we're using it to conduct a test. Thus, there is a null hypothesis we need to consider – that the population proportions for each population are the same ($p_1 = p_2$). If the population proportions are the same, then we could consider our samples from each population as coming from an identical population. So to see what might happen by random chance, we can take two samples from this population using bootstrapping, and see how different they are!

Let's try this with our previous observational study example – the blood donations. Remember that the table of the results for this study is as follows:

	Donated	**No Donation**	**Total**
2012	144	1154	1298
2014	161	1106	1267
Total	305	2260	2565

Thus, if we were to take this data and simulate this bootstrapping idea in TinkerPlots, the model might look like the sampler shown on the right.



Structurally, this looks nearly identical to the Dolphin Therapy study! We took our totals for each of the two variables and are re-mapping them to each other in some way. But this device makes one major change: the response or outcome variable is now set to with replacement, as it shows a closed top! This makes it so that we are taking one big bootstrapped sample from two identical populations (according to our null!), with 1298 being for the year 2012, and 1267 for year 2014.

Thus, if we wanted to replicate this simulation process in R, it will look very close to the dolphin study, but with one slight change:

```
year_rand = sample(blooddonations$year, 2565, replace=F)
donate_rand = sample(blooddonations$donation, 2565, replace=T)
blood_rand = data.frame(treatment_rand, result_rand)
table_rand = table(blood_rand)
p_2012 = table_rand[1]/(table_rand[1]+table_rand[3])
p_2014 = table_rand[2]/(table_rand[2]+table_rand[4])
p_2014 - p_2012
```

Here, we're still doing a random process on each vector and calculating the difference in sample proportions, but now, the donation variable is sampling with replacement, reflecting the change from random assignment to random sampling with bootstrapping.

To simulate this process many times now, we use the same for-loop structure that we have been using. We want to see how likely it is to get a difference of proportions bigger than the difference we observed in the table above between 2012 and 2014, which comes out to a proportion of 0.0161.

```
diffs2 = rep(0, 1000)
```

```
for (i in 1:1000) {
  year_rand = sample(blooddonations$year, 2565, replace=F)
  donate_rand = sample(blooddonations$donation, 2565, replace=T)
  blood_rand = data.frame(treatment_rand, result_rand)
  table_rand = table(blood_rand)
  p_2012 = table_rand[1]/(table_rand[1]+table_rand[3])
  p_2014 = table_rand[2]/(table_rand[2]+table_rand[4])
  diffs2[i] = p_2014 - p_2012
}
```

```
hist(diffs2)
abline(v=0.0161, col="red")
```

```
mean(diffs2 >= 0.0161)
```

What is the *p*-value that you get here? How does it compare to what we found with `prop.test` earlier?

8.4 – Additional Practice

Example: Let's revisit the dolphin therapy study we investigated last class: Researchers wanted to investigate a new form of animal therapy on depression. To do this, they recruited 30 participants aged 18-65 with a clinical diagnosis of mild to moderate depression. These 30 participants went to an island off the coast of Honduras, where they were randomly assigned to one of two groups (15 participants in each group). Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the dolphin therapy group) did so in the presence of bottlenose dolphins, while the other group (the regular therapy group) did not. When the experiment was completed each participant's level of depression was evaluated in order to determine whether or not a participant showed substantial improvement after the participant's therapy (dolphin therapy or regular therapy).

- Evaluate the design of this study in terms of its internal and external validity. What does this imply about the types of conclusions that can be drawn from this study?

- Determine if the dolphin therapy was effective in treating subjects with depression at the $\alpha = 0.05$ level using R to directly conduct the test (no simulation). The data from this study can be found in the **dolphin.csv** file.