Stat 102 Notes

Table of contents

Course Goals for STAT 102 **3**

1 Philosophy of Statistical Inference (Chapters 11-14) **4**
1.1 Randomization Tests (Chapter 11) . . . 6
1.2 Bootstrap Methods (Chapter 12) . . . 12
1.3 Inference with Mathematical Models (Chapter 13) . . . 14
1.4 Decision Errors (Chapter 14) . . . 18

2 Inference for Proportions (Chapters 16-18) **22**
2.1 Inference for a Single Proportion (Chapter 16) . . . 24
2.1.1 Bootstrap Tests for One Proportion . . . 24
2.1.2 Bootstrap Confidence Intervals for One Proportion . . . 28
2.1.3 Mathematical Model for a Proportion . . . 29
2.2 Inference for a Comparing Two Proportions (Chapter 17) . . . 34
2.2.1 Randomization tests for the difference in proportions . . . 34

Course Goals for STAT 102

STAT 102 is an introduction to formal statistical inference. We will carry out inference using both simulation-based approaches and classical, theory-based methods. By the end of the course, you will:

- Read an example where the research question is explicitly stated, and then translate what's stated into a statistical statement involving parameters or other simple distributional characteristics.
- Identify whether the ideal data collection strategy would involve random assignment, random sampling, or both and explain why.
- Work with an example where the research question is explicitly stated, along with an existing data set, and propose and carry out an appropriate analysis to answer the research question.
- Explain the terms/components of a given statistical model, and connect those terms to the research question at hand.
- Check basic assumptions of various (simple) analysis methods and justify the use of the method.
- Apply existing functions and point-and-click software for implementing basic data analyses.
- Use tactile simulation to carry out a simple resampling procedure.
- Identify the steps and perform the calculations required for routine statistical procedures to address a given problem.
- Calculate simple analyses (t-test, chi-squared test for proportions) by hand, to verify the validity of the computational algorithm.
- Recognize when computational results do not make sense in the context of the problem.

1 Philosophy of Statistical Inference (Chapters 11-14)

In STAT 101, you focused on Exploratory Data Analysis. Exploratory data analysis aims to investigate the characteristics of a data set through visualizations and numerical summaries. Visualizations may include:

- box plots
- histograms
- bar charts
- pie charts
- scatterplots
- heat maps
-
-
-

Numerical summaries used to explore a data set may include:

- sample mean
- sample variance/standard deviation
- five number summary, and other order statistics
- sample proportions
- calculated regression slope and intercept
-
-
-

More often than not, the data were collected to answer a research question about a larger population for which the data collected are a (hopefully) representative sample. This notion of drawing conclusions beyond the data collected is at the heart of statistical inference.

Example: Bred in the Bone

- If each baby is really guessing/choosing blindly, what proportion would you expect to choose the good guy? Why?

- Based on this, what randomizing device could we use to model this experiment?

- Experiment! Add your results to the plot on the board.

- Applet

- What do you observe in the plot?

- Real experiment

Take away:

In exploratory data analysis, the visualizations and numerical summaries you choose are driven by the type of data at hand. This is true for statistical inference as well. The type of data will drive the appropriate inference techniques. However, the goal of the research study will also impact the selected method, as will the underlying assumptions of the technique (we'll talk **a lot** more about this). That said, there are some overarching approaches to quantifying variability, and thus drawing conclusions beyond the data set at hand.

Approaches to quantifying variability

- Randomization methods (Chapter 11)
- Bootstrap methods (Chapter 12)
- Mathematical models (Chapter 13)

We'll start the semester by talking about these three approaches fairly generally. For (most of) the rest of the semester, we'll see how these approaches fit with different types of data.

1.1 Randomization Tests (Chapter 11)

The goal of hypothesis tests is to use an **observed** data set to answer a yes/no question about a characteristic of a larger population from which the observed data set was drawn. For example, is swimming with dolphins therapeutic for patients with clinical depression? That is, we want to assess whether or not the explanatory variable causes changes in the response variable.

To answer this question, Antonioli and Reveley (2005) recruited 30 subjects with a clinical diagnosis of mild to moderate depression. The subjects were required to stop all other treatments (therapy and/or pharmaceuticals) 4 weeks prior the experiment, and the 30 subjects were all taken to an island off the coast of Honduras. The subjects were randomly assigned to one of two groups. Both groups spent one hour swimming and snorkeling each day, but one group did so in the presence of dolphins and the other group did not. At the end of two weeks, each subject's level of depressions was evaluated, and whether or not the subjects had a substantial improvement in their depression was recorded.

Explanatory variable:

Response variable:

Is this an observational study or an experiment? What does that imply about inference?

The question we will answer is whether the resulting data provide convincing evidence that subjects who swam with dolphins were more likely to see depression improvement than subjects who swam without dolphins.

If there really is no impact of swimming with dolphins, what does this imply about the explanatory and response variables?

If swimming with dolphins does improve depression, what does this imply about the explanatory and response variables?

This leads to two competing claims:

- **Null hypothesis**: H_0

- **Alternative hypothesis**: H_a

If the null hypothesis is true, how would this manifest in the observed data?

If the alternative hypothesis is true, how would this manifest in the observed data?

We will choose between the competing claims by assessing whether the data conflict so much with H_0 that the null hypothesis cannot be considered reasonable. If this happens, we'll reject the notion of H_0 and conclude that H_a must be true.

Up to now, we haven't seen the data! Here's a summary:

	Dolphin Therapy	Control Group	Total
Showed Improvement			
No Improvement			
Total			

We can see that

-
-

So,

The question remains…is this enough different from what we would expect under the null hypothesis to conclude that swimming with dolphins does make a difference in depression?

So far, nothing we've laid out is unique to a randomization test. Where does randomization come in?

Let's visualize these observations as a set of cards. Each card denotes a subject in the study. The color indicates the response: red for substantial improvement and black for no substantial improvement.

Any difference we see in the simulation is due to chance–the cards were randomly dealt into the dolphin/control groups.

It's not realistic to keep shuffling and dealing by hand...we need to turn to technology to do the randomization for us: Applet

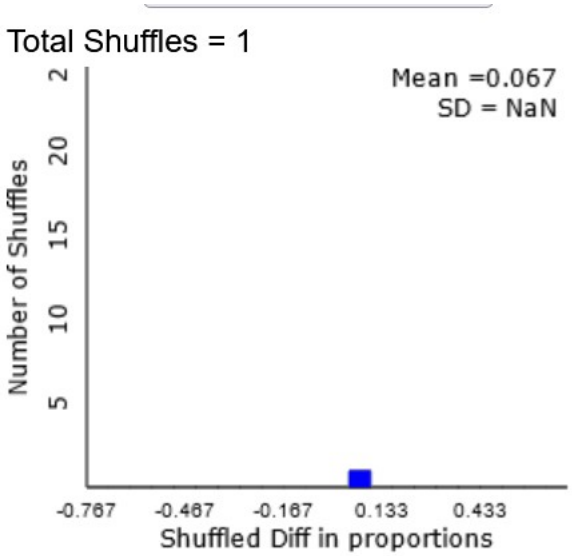


Figure 1.1: One Shuffle

We can do this over and over again to build up a **null distribution**. This distribution shows how we expect the variability to behave under the null hypothesis:

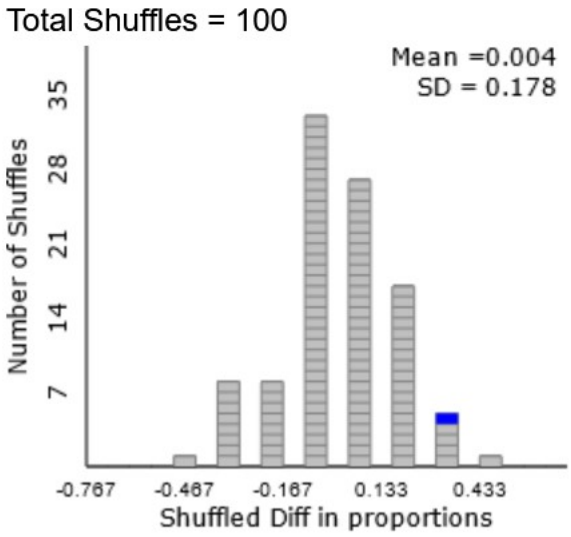


Figure 1.2: 100 Shuffles

What do you notice about this null distribution?

How rare is it to see our observed statistic **0.467** in this distribution? What does this imply?

So, we've just carried out a **statistical inference** technique! We might be wrong in our conclusions (more on this in Chapter 14), but we've made the best decision we could with the data available.

In summary:

Randomization Test Procedure:

- Frame the research question in terms of hypotheses
- Collect data from an observational study or experiment
- Model randomness that would occur if H_0 is true
- Analyze the data by comparing the observed data to the simulated distribution
- Conclusion

Now let's go to R!

1.2 Bootstrap Methods (Chapter 12)

Bootstrap methods are a relatively new statistical technique (proposed in 1979 by Efron), but they are based on a very simple idea. The goal is to characterize the variability of the statistic across many samples. One way we could do this is take lots and lots of samples from the population, and get a picture of how much variance there is among the samples. This is almost always impossible. So, rather than resample from the population, we could try resampling from the sample. This is the basic idea behind the bootstrap.

Bootstrapping is used in many different applications. For this general introduction to the approach, we're going to consider a confidence interval for a proportion.

A **confidence interval** is

Note the goal of the confidence interval is different from the goal of a hypothesis test!

However, like with hypothesis tests, we need to understand the variability inherent to the statistic. To figure out how wide the range of plausible values should be, we need to know how a statistic varies from sample to sample in the population.

For example, let's think back to the Baby scenario and suppose our goal is to estimate the population parameter

The researchers collected one sample of 16 babies, and found that 14 picked the good guy. This is our observed data. What do you think would happen if we took a sample of 16 different babies? And then a different sample of 16 babies?

Idea of the bootstrap:

Infinite populations are pretty tough to work with, though. However, we can produce an equivalent bootstrap distribution by

So, we'll repeatedly draw bootstrap samples of size 16 (why 16?) and calculate the proportion of successes in each bootstrap samples. After we do this many many times, we'll have an idea of a range of plausible values for the population parameter. We'll set the **confidence level** by opting for a wider or a narrower interval, based on how certain we need to be in the results.

Bootstrap Process

- Frame the research question in terms of a parameter to estimate
- Collect data using an observational study or an experiment
- Model the randomness by using the observed data as a proxy for the population
- Create the interval (in future chapters we'll see there are multiple ways to do this)
- Conclusion

Let's go to R!

1.3 Inference with Mathematical Models (Chapter 13)

So far, we've seen computational methods like randomization and bootstrapping to characterize the variability of a statistic. The use of computational methods is relatively recent, due to the increase in computing power. In pre-computing days, re-sampling and randomization was very difficult. As a result, mathematical approximations were used and are still pervasive. If you took AP Statistics or a different intro statistics course, you employed mathematical models. However, to be clear, all of the methods we'll talk about (randomization, bootstrap, mathematical models) are techniques to get a **sampling distribution**.

The sampling distributions we've seen so far have been (mostly):

This isn't coincidence…it's guaranteed by a very important theorem, the **Central Limit Theorem**.

Central Limit Theorem

What are the requirements here?

- Independence:

- "Large enough":

Normal Distribution: Nothing follows it exactly–it's a mathematical construct. But, a lot of things follow it approximately, either:

- naturally:

- created to follow it:

The normal distribution depends on two parameters, μ = mean (where the distribution is centered) and σ = standard deviation (how spread out it is). μ shifts the distribution up and down the number line, σ stretches and contracts the curve. The **standard normal** distribution has $\mu = 0$ and $\sigma = 1$ (this is the distribution tabulated in normal tables in textbooks).

The standard normal gives us a convenient way to compare observations, and any normal distribution can be transformed into a standard normal. The **Z-score** is

If the Z-score is positive

If the Z-score is negative

Z-scores can be used to

- gauge the unusualness of an observation

- find probabilities

Helpful R functions:

- `pnorm(x, mean=0, sd=1)`
- `normTail(m=0,s=1,L=x)` or `normTail(m=0,s=1,U=x)` will draw pretty pictures–need to use the OpenIntro library
- `qnorm(prob, mean=0, sd=1)` gives a Z-score with area to the left

Pictures are super-helpful!

Example: Full-term birth weights for single babies are normally distributed with a mean of 7.5 pounds and a standard deviation of 1.1 pounds.

1. A baby is born weighing 9.1 pounds. What is the weight percentile for this baby?

2. Babies that weigh less than 5.5 pounds are considered low birth weight. What proportion of babies are low birth weight?

3. What weight would make a baby at the 25th percentile?

4. What is the probability a randomly selected baby weighs between 7 and 8 pounds?

The **Empirical Rule** (aka the 68-95-99.7 Rule) presents a general rule for the probability of falling within one, two, and three standard deviations of the mean in a normal distribution.

This rule is useful in a wide range of settings when trying to make quick estimate (we'll use it with bootstraps too!).

Some more definitions we'll use throughout the semester:

- **Standard error**:

- **Margin of error**:

Example (13.11): In 2013, the Pew Research Foundation reported that "45% of US adults report that they live with one ore more chronic conditions." However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used. Create a 95% confidence interval for the proportion of US adults who live with one or more chronic conditions. Interpret the confidence interval in the context of the study.

1.4 Decision Errors (Chapter 14)

Anytime we're using sample data to make decisions about a larger population we can potentially make a mistake. We can make an incorrect decision in a hypothesis test or calculate a confidence interval that does not capture the true population parameter. In a hypothesis test, there are four possible outcomes:

Type I error:

Type II error:

Examples:

- Doping in the Olympics

- Criminal trial

- Diagnostic test for a serious disease

Errors require a balancing act. We want to reduce the chance of making a Type I error but this will necessarily increase the chance of making a Type II error. The best we can do is to set the probability of a Type I error. We can do through setting the **significance level**.

Significance level:

Another consideration that will impact the chance of making an error is the whether the test is one- or two-sided.

Two-sided hypotheses:

Example: Standard anticoagulant therapy to prevent blood clots requires frequent (expensive) lab monitoring. A new procedure called riva was tested because it did not require frequent monitoring. A randomized trial was conducted in 2012, with standard therapy randomly assigned to 2416 patients and riva randomly assigned to 2416 patients. A bad result was a recurrence of a blood clot in a vein. We want to know if the likelihood of a bad result is different between the two therapies.

Here are the results of the randomized trial

	Riva	Standard	Total
Clot	44	60	104
No Clot	2372	2356	4728
Total	2416	2416	4832

For two-sided tests, the p-value is the probability that we observe a result as least as favorable to the alternative hypothesis as the result we observe. That is, that we observe a result as extreme or more extreme in either direction.

When in doubt, use a two-sided test! Use a one-sided test only if you truly have interest in only one direction.

So, how can we control Type I error?

- Set up tests before seeing the data.
- Collect enough data that the test has sufficient **power**. We'll talk more about power later (and LOTS more in an experimental design course), but power is the probability of correctly rejecting a false null hypothesis. It's a function of how big the true difference is (which we don't know and can't control) and the sample size (which we can control).

2 Inference for Proportions (Chapters 16-18)

So far, we've discussed randomization, bootstrap, and mathematical models as methods to approximate/describe a sampling distribution and quantify variability. Now, we turn to how these three methods can be used to answer research questions for different kinds of data. The appropriate method will depend on both the type of data and the research question of interest.

During our class, we'll discuss two types of data: **categorical** and **quantitative**. Categorical data arise when the responses are categories. If you think about what is being measured on each unit in the sample, and could imagine checking a box to record the response, the data are categorical. For example:

We also have to consider the research question. The research question of interest will drive answers to the following:

1. Is a single variable being measured on each unit in the sample, or are two (or more) variables being recorded?

2. If two or more variables are being recorded for each unit in the sample, can one be considered the **response** variable and the other(s) be considered **explanatory**?

3. If a variable is categorical, does it have two possible outcomes (like yes/no) or more than two possible outcomes?

4. Is the research question focused around finding the answer to a yes/no question (like "Does a new teaching method improve student test scores?") or around estimating a value (like "By how much do student test scores change if a new teaching method is introduced?")?

5. Were the data collected obtained using random sampling, random assignment, both, or neither? (this is not typically driven by the research question, but can impact our analysis method and will **definitely** impact the conclusions we can draw.)

The answers to these questions will help us determine which method is most appropriate, as well as the specific analysis tool to implement that method. In this unit, we're going to focus on categorical variables. We're going to start with a single categorical variable measured on each unit in the sample, where that categorical variable has only two possible outcomes. From there we'll move to two categorical variables measured on each unit (one explanatory, one response), again with only two possible outcomes for each. Finally, we'll explore categorical variables with more than two possible outcomes.

2.1 Inference for a Single Proportion (Chapter 16)

Our first scenario involves a single categorical variable measured on each unit in the sample, with only two possible outcomes.

2.1.1 Bootstrap Tests for One Proportion

When we discussed bootstrapping earlier, we were sampling (with replacement) from our sample data, because we wanted to understand the variability inherent to our statistic, $\hat{p}$, assuming our sample is representative of all samples of the same size that could be drawn from the population. The goal of hypothesis testing is different: we want to understand the sampling distribution of $\hat{p}$ under the assumption

So, we need to repeatedly sample from a population with $p = p_0$. We can do this by simulating data sets of the same size as original sample, assuming that $p = p_0$. This is called a **parametric bootstrap**, because we are making an assumption about the value of the underlying parameter p and we are assuming a particular distribution to generate our simulated data sets. From each simulated data set, we could calculate the resulting $\hat{p}_{\text{sim}}$. Many simulated data sets will give us a good approximation for the distribution of $\hat{p}$ under our assumptions.

Example: Back to the babies picking the good guy or bad guy. We want to know if babies are more likely to pick the good guy puppet.

Under H_0, 50% of babies will pick the good guy. We'll assume this is true for all babies that could be tested. We'll simulate 16 babies undergoing this test to get a sample proportion from the null distribution.

Let's see how this works in R. We'll start by setting up the original data, so we can calculate the test statistic.

```
data_baby<-c(rep(1,14),rep(0,2))

obs_prop<-mean(data_baby)
obs_prop
```

```
[1] 0.875
```

Next, we'll set up the null model with 50% successes and 50% failures:

```
para_boot<-c(rep(1,8),rep(0,8))

para_boot
```

```
 [1] 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
```

Now, we'll repeatedly sample from the null distribution, and collect the resulting $\hat{p}_{pb}$ from each sample.

```
numsim<-100
boot.sample<-data.frame(sim=1:numsim,stat=NA)

head(boot.sample)
```

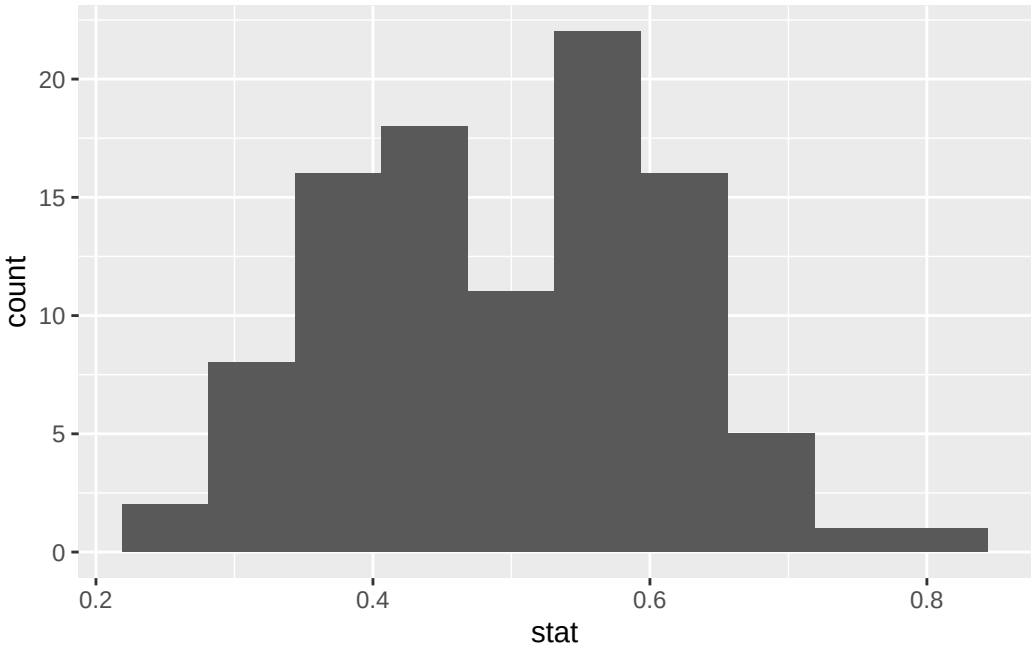
```
  sim stat
1   1   NA
2   2   NA
3   3   NA
4   4   NA
5   5   NA
6   6   NA
```

```
for(i in 1:numsim){
  boot.sample$stat[i]<-mean(sample(para_boot,size=16,replace=TRUE))
}

head(boot.sample)
```

```
  sim   stat
1   1 0.7500
2   2 0.5625
3   3 0.5625
4   4 0.4375
5   5 0.4375
6   6 0.5625
```

We can plot these $\hat{p}_{pb}$, and see how unusual our observed test statistic is.

```
library(ggplot2)
boot.hist<-ggplot(boot.sample, aes(stat)) + geom_histogram(bins=10)

boot.hist
```



We've got a plot, but how do we know exactly how many/what proportion of $\hat{p}_{pb}$ were greater than our observed test statistic, $\hat{p} = 0.875$?

```
count<-(boot.sample$stat >= obs_prop)

sum(count)
```

```
[1] 0
```

```
sum(count)/numsim
```

```
[1] 0
```

So, we have an estimated p-value of

What if we want to change the number of simulated data sets? Let's try 1000. This gives an estimated p-value of

Why is this an estimated p-value?

Why does this histogram not look bell-shaped?

Try Problem 5 in Chapter 16, and see if you can modify the Babies R code to re-create the histogram provided in the book. It might help to try the applet first and see what has to change there.

Why Bootstrap?

- Works for any sample size!
- Intuitive way to explain what the p-value is actually measuring.

2.1.2 Bootstrap Confidence Intervals for One Proportion

We've already done these! Recall that with a confidence interval, we're not assuming the null hypothesis is true. Rather, our goal with the bootstrap is to characterize the variability of our statistic $\widehat{p}$.

2.1.3 Mathematical Model for a Proportion

Sometimes, the sampling distribution of $\hat{p}$ can be well-approximated using a normal distribution. The conditions which must be met are:

-
-

If these conditions are met, then

[Note: This result is just another way of stating the Central Limit Theorem when dealing with proportions! The success/failure condition is playing the role of the "sample is large enough" requirement of the CLT.]

Let's think more about the standard error of $\hat{p}$. Remember the **standard error** is

But this presents a problem. We don't know p (if we did, we wouldn't be doing tests or confidence intervals)!

How is this going to play out in hypothesis tests?

Example: Look at Problem 3 in Chapter 16. The journalist claims more than 1/5 adults living in Seattle support defunding the police. Is this true?

To find the p-value, we can use R.

```
library(openintro)
```

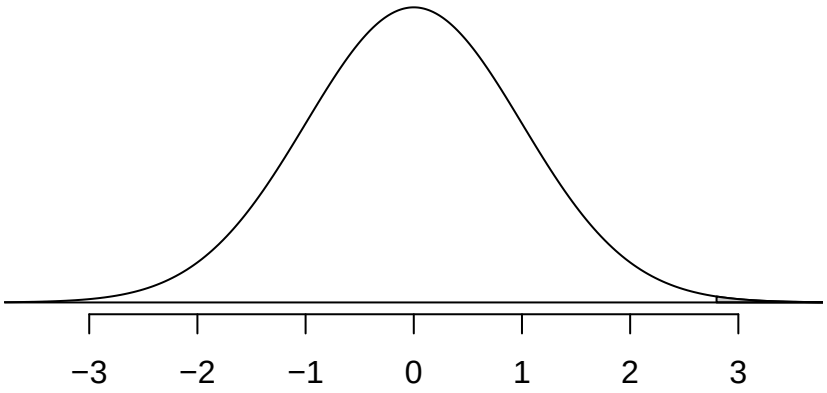
```
Warning: package 'openintro' was built under R version 4.4.3
```

```
Loading required package: airports
```

```
Loading required package: cherryblossom
```

```
Loading required package: usdata
```

```
normTail(m=0,s=1,U=2.79)
```




```
pvalue<-1-pnorm(2.79,mean=0,sd=1)
pvalue
```

```
[1] 0.002635402
```

Do you expect the p-value from the parametric bootstrap would be similar? Why or why not?

Let's see.

How is this going to play out in confidence intervals?

Example: Back to Chapter 16, problem 3. The journalist found that 159/650 Seattle residents support proposals to defund the police.

We can change the confidence level by changing z^*, and using the `qnorm` function.

- 90% confidence

```
qnorm(0.05,mean=0,sd=1)
```

```
[1] -1.644854
```

Back to the bootstrap confidence interval...

So far, we've seen **bootstrap percentile confidence intervals**. We calculated these directly from the bootstrapped $\hat{p}_{\text{boot}}$. If we want a 90% confidence interval, we can find the 5^{th} and 95^{th} percentile values of the $\hat{p}_{\text{boot}}$ values.

We can also use the variability of the $\hat{p}_{\text{boot}}$ to calculate an estimate of the standard error of $\hat{p}$, and then calculate the interval using the mathematical model approach. This is a **bootstrap SE confidence interval**.

This is a rough approximation, using the 68-95-99.7 Rule which says that 95% of the observed differences should be no farther than 2 SE from the true parameter (p). To do this, the bootstrap histogram must be roughly symmetric and bell-shaped. So, it works for Chapter 16, problem 3; it doesn't work with the babies.

Why Z-Test/Z Confidence Intervals?

- In many cases, works as well as simulation methods
- Easy to calculate without technology/can do "back of the envelope" analyses
- Z-scores and "estimate $\pm$ margin of error" are easily interpretable
- Classical methods, used by scientists across disciplines

2.2 Inference for a Comparing Two Proportions (Chapter 17)

We now move on to consider situations in which two categorical variables are measured on each unit in the sample, and each variable has two possible values. In cases like these, typically one variable is considered the response and one variable is considered explanatory. The explanatory variable may be randomly assigned (like whether or not a subject swam with dolphins) or it may be merely observed (like smoking status). The two possible values of the explanatory variable lead to two groups, and we're interested in comparing the population proportions that arise from these two groups. We'll focus on the function of parameters $p_1 - p_2$. The natural estimate of this is $\hat{p}_1 - \hat{p}_2$: the difference in the sample proportions. We'll be constructing hypothesis tests to compare p_1 to p_2 and finding confidence intervals to estimate $p_1 - p_2$.

2.2.1 Randomization tests for the difference in proportions

Example: Researchers are interested whether electrical brain stimulation will help with problem solving tasks. 40 volunteers were all trained to solve problems in a particular way. Half of the volunteers were randomly assigned to receive electrical stimulation and the other half received a sham stimulation (placebo). All volunteers were then presented with an unfamiliar problem and asked to solve it. The researchers are interested in testing whether the proportion able to solve the problem following electrical stimulation is greater than the proportion able to solve the problem without electrical stimulation.

There are a couple of different ways we could state the hypotheses of interest:

- H_0 :

- H_a :

Recall that hypothesis tests work by assessing how unusual our observed data are, if the null hypothesis is really true. A very unusual result implies that observed data are not likely to have occurred under the null hypothesis. Randomization tests allow us to assess that unusualness by estimating the null distribution–a simulated distribution of what we could expect the distribution of $\hat{p}_1 - \hat{p}_2$ to look like if H_0 is true. We assume H_0 is true by recreating the randomization that occurred in the experiment.

Here are the data:

	Solved	Not Solved	Total
Sham			20
Electrical			20
Total			40

To demonstrate what the randomization test is doing, we need 40 cards. Why 40?

Of these cards, _____ are red and _____ are black. What do these represent?

We'll shuffle, and deal into two stacks.

A randomization test is going through this shuffling/dealing over and over again, find the difference in proportions for each simulation.

Let's look at this in the applet.

What do you notice about the null distribution? How unusual is the observed $\hat{p}_1 - \hat{p}_2$?

Example: Try 17.2. Set up the hypotheses and describe how a randomization test would work.

- How many cards are needed?
- How many red? How many black?
- How many should be dealt into each stack?
- What would you calculate from each shuffle/deal?

Let's do this in R!

Bootstrap confidence interval for the difference in proportions

As we saw with a single proportion, bootstrapping will allow us to estimate the variability of $\hat{p}_1 - \hat{p}_2$ without assuming the null hypothesis is true. With a single proportion, we drew repeated samples (with replacement) from our sample data, and from each bootstrap sample calculated $\hat{p}_{\text{boot}}$. The distribution of the $\hat{p}_{\text{boot}}$ provided an estimation of the sampling distribution of $\hat{p}$.

Now, with two samples, our observed statistic of interest is

Let's go to R, and see how this works with the electrical stimulation example.

Now, by re-using this example, we are ignoring (maybe) the research question. But, we're doing both a randomization test and confidence interval so that we can compare the resulting sampling distributions of $\hat{p}_1 - \hat{p}_2$. What is different? What's the same?

Bootstrap percentile confidence interval:

Bootstrap SE confidence interval:

Mathematical model for the difference in proportions

For a single proportion, we needed two conditions to be met to ensure the sampling distribution of $\hat{p}$ is approximately normal:

-
-

If these conditions are met, then

We must meet similar conditions to ensure the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal:

-
-

If these conditions are met, then

Like before we don't know p_1 and p_2, so we'll use our best guess. And, like before, our best guess will change depending on whether we're constructing a confidence interval or carrying out a hypothesis test.

How is this going to play out in a hypothesis test?

Example (17.5): (Aside: why can't we do the electrical stimuation example again?). A 2021 Gallup poll surveyed 3941 students pursuing a bachelor's degree and 2064 students pursuing an associate's degree. The survey found that 51% of the bachelor's students (2010) and 44% of the associate's students (908) said that COVID-19 will negatively impact their ability to complete the degree. We want to decide whether the proportion of bachelor's students who believe the pandemic will negatively impact degree completion is different from the proportion of associate's students who believe they will be negatively affected. Let p_B be the proportion of bachelor's students who believe they'll be negatively affected and let p_A be the proportion of associate's students who believe they'll be negatively affected.

To visualize the p-value: `normTail(m=0,s=1,L=-5.15, U=5.15)`

To get the p-value: `pnorm(-5.15,mean=0,sd=1)+(1-pnorm(5.15,mean=0,sd=1))`

How is this going to play out in a confidence interval?

Example(17.9): A Kaiser Family Foundation poll for US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan.'' There were 347 Democrats, 298 Republicans, and 617 Independents surveyed (Foundation, 2019). We want to estimate the difference between the proportion of Democrats and Independents who support a National Health Plan.

What impacts the width of a confidence interval?

There are three main things that impact the width of a confidence interval:

- Confidence level

- Sample size

- Standard Error

No matter what method we use to calculate the confidence interval, the **confidence level** is a statement about the long run percentage of confidence intervals that would succeed in capturing the true value of the parameter. What does this mean? Applet

Hypothesis tests vs. Confidence Intervals

While we should be matching analysis method to the research question, there is a nice relationship between hypothesis tests and confidence intervals. Recall that confidence intervals give a set of plausible values for the unknown parameter.

Inference for Two-Way Tables (Chapter 18)

So far, we've considered categorical variables with only two possible outcomes: success and failure. Many categorical variables have more than two possible outcomes, so we can't easily define the proportion of "successes.'' Instead, we'll summarize categorical data with more than two levels using two-way tables. In this class, we're still going to restrict ourselves to only two variables (often explanatory and response, but not necessarily), both with two or more levels. However, there are certainly statistical methods for more complicated situations.

Typically, research questions focus on how the proportions of the possible outcomes in the response variable change (or don't) across the levels of the explanatory variable. However, we can also consider questions about a single variable with more than two outcomes (are the possible outcomes all equally likely? do the possible outcomes follow a particular pattern?) or just whether the two categorical variables are independent or dependent without assigning an explanatory/response relationship. Due to the structure of the variable(s), there really isn't a population parameter of interest. We can't (usually) make a function of proportion of successes that makes sense to estimate, like we can with $p_1 - p_2$. That means we'll be considering only tests, not confidence intervals. We'll focus on the randomization test and the mathematical model approach. Both methods start with the same set-up.

Example: When surveys are administered, we hope that the respondents give accurate answers. Does the mode of survey delivery affect this? Schober et al (2015) investigated this question. They had 147 people who agreed to be interviewed on an iPhone, and they were randomly assigned to one of three interview modes: human voice, automated voice, text. One question asked was whether they exercise less than once per week during a typical week (a yes is mostly likely considered socially undesirable). The explanatory variable here is survey mode and the response is whether or not the respondent said yes. Here are the data:

		Text	Human voice	Automated voice	Total
Exercise	Yes	34	21	20	75
	No	124	139	139	402
Total		158	160	159	477

Based on these data, it looks like the answer to the question does change depending on survey mode, with respondents more likely to say yes via text. However, we don't know if this result could have happened by chance.

Expected Counts

We don't expect the proportion of 'yes' to be exactly the same across all survey modes, but we want to know if these vary enough to convince us that survey mode and answer are not independent. To do this, we need to find **expected counts** for each cell in the table.

		Text	Human voice	Automated voice	Total
Exercise	Yes	34 ( )	21 ( )	20( )	75
	No	124 ( )	139 ( )	139 ( )	402
Total		158	160	159	477

So now the key question…are the observed and expected cell counts different enough?

- Cell(1,1) obs - exp = 34 -
- Cell(1,2) obs - exp = 21 -
- Cell(1,3) obs - exp = 20 -
- Cell(2,1) obs - exp = 124 -
- Cell(2,2) obs - exp = 139 -
- Cell(2,3) obs - exp = 139 -

New test statistic!

In our example:

- Cell(1,1) (obs - exp)^2/exp = $(34 - 24.84)^2/(24.84) = 9.16^2/24.84 = 3.3778$
- Cell(1,2) (obs - exp)^2/exp = $(21 - 25.16)^2/(25.16) = (-4.16)^2/25.16 = 0.6878$
- Cell(1,3) (obs - exp)^2/exp = $(20 - 25)^2/(25) = (-5)^2/25 = 1$
- Cell(2,1) (obs - exp)^2/exp = $(124 - 133.16)^2/(133.16) = (-9.16)^2/133.16 = 0.6301$
- Cell(2,2) (obs - exp)^2/exp = $(139 - 134.84)^2/(134.84) = 4.16^2/134.84 = 0.1283$
- Cell(2,3) (obs - exp)^2/exp = $(139 - 134)^2/(134) = 5^2/134 = 0.3731$

To see if this is 'big' we need the sampling distribution of our new test statistic. We can estimate that sampling distribution using either a randomization test or the mathematical model approach.

Randomization Test

The randomization test for a two-way table works just like it does with two samples. We'll randomize by shuffling and dealing/assigning the 75 yes answers and 402 no answers to the three survey modes at random.

- How many colors of cards?

- How many stacks to deal them into? How many in each stack?

- What do we find for each deal/shuffle?

Applet

Conclusion:

Mathematical Model

Based on what we just observed in the applet, the normal distribution is not going to be a good approximation to sampling distribution. It turns out this test statistic follows a different mathematical distribution, the **chi-squared distribution** (proof: see STAT 262). The normal distribution has two parameters that determine its shape: the mean (μ) and standard deviation (σ). The shape of the chi-square distribution is determined by a parameter called the **degrees of freedom (df)**. Figure 18.2 on page 339 shows how the shape of the distribution changes depending on the df.

So how can we use this?

Again, we have conditions that need to be met for the mathematical model to be a good approximation:

-
-

Example: Let's go back and do the survey mode example using the mathematical model approach. First, we'll need to check the conditions are met:

-
-

To find the p-value, we can use the R function `pchisq()`. Like `pnorm` it gives area to the left. So,

p-value = `1-pchisq(6.1971,df=2) =`

We can also do this directly in R.