The Mathematics Enthusiast

Volume 19
Number 3 *Number 3 (In Progress)*

Article 11

12-2022

Undergraduate Research in an Applied Probability & Statistics Course

Michael Smith

Recommended Citation
Smith, Michael (2022) "Undergraduate Research in an Applied Probability & Statistics Course," *The Mathematics Enthusiast*: Vol. 19 : No. 3 , Article 11.
DOI: https://doi.org/10.54870/1551-3440.1579
Available at: https://scholarworks.umt.edu/tme/vol19/iss3/11

Undergraduate Research in an Applied Probability and Statistics Course

Michael D Smith
Lewis University, Romeoville, IL

ABSTRACT: The author modified a course entitled *Applied Probability and Statistics* to engage students in performing their own hypothesis tests in a course-based undergraduate research experience (CURE). This paper discusses the structure of the course, what it does to encourage undergraduate research, and changes one could make to tailor this experience to their own institutional needs.

Keywords: undergraduate research, CUREs, probability, statistics

The Mathematics Enthusiast, **ISSN 1551-3440, vol. 19, no. 3**, pp. 807–821

Introduction

Prior studies that have showcased the benefits of undergraduate research ([WL], [Auch]). In an attempt to expose more students to the research process, the author has adjusted their Applied Probability and Statistics course to provide students with an introductory research experience. The purpose of this article is similar to that of [Ber] and [Cam], in that the author will be describing the experience that they provide to their students, why they have made the choices they have, reflecting on changes that could be made, and speculating on the impact those changes could have. There will also be discussion on how this course fits within the Course-Based Undergraduate Research Experience (CURE) framework provided by [Auch].

Lewis University is a private Roman Catholic and Lasallian university in Romeoville, Illinois. The enrollment is currently around 6,800 total students, with around 4,100 undergraduate students. This course is a required course for Mathematics, Computer Science, and Engineering majors, and is also taken by a variety of other STEM majors. It has a prerequisite of Calculus I, as the topic of continuous random variables requires knowledge of integration. Below is the course description from the course catalog.

"This course introduces the concepts of statistics and probability, including measures of center and spread, correlation coefficients, regression, random variables, discrete and continuous distributions, confidence intervals, and hypothesis testing. Students will also learn to use technology to complete statistical analyses."

It is worth mentioning that this course has gone through several iterations. It used to be that the material from this course was covered over two semesters as a full sequence, with some extra topics in hypothesis testing covered. Since then, the course has been designed to be a stand-alone course, covering more material in the first semester. In this new format, Applied Probability and Statistics started as an inquiry-based learning course, and then morphed into its current iteration with a focus on introducing students to performing their own analyses. There are now two separate courses that can be used to complete the sequence: Probability Theory, which focuses on a larger variety of random variables with a focus on preparing students for Actuarial Exam P; and Advanced Statistics, which focuses on a variety of statistical tests and using more technology in order to perform these tests. Applied Probability and Statistics is designed to prepare students for either course they may choose to take.

1 Description of Course Experience

This course is intended to give students from various backgrounds an introductory look at probability and statistics while also encouraging them to engage in the process of testing claims on their own. One of the main goals of this course is to develop students who are independent thinkers that can verify information presented to them. It is also a goal that students will become interested in research by having students engage in the whole process of hypothesis testing on claims they have chosen with data they have chosen.

Because there is much material to cover that is new to most of the students, this class could be described as mostly lecture-style with active-learning opportunities added where possible. The lectures for this class take two forms. Most lectures take the form of guided note packets, in which the professor leads students through concepts and definitions with some examples before having the students try some of the examples for themselves. The other type of lecture, which occurs before most projects, takes the form of iPython notebooks. In this style of lecturing, the instructor shows students how to use Python to import, analyze, and visualize data from data sources. This leads students to begin experimenting with programming, as not all of them have a background in it. There is nothing special about using Python. Other instructors have used R for programming in the past instead of Python, or have used Excel for basic functions instead of programming.

2 Course Assessments

This course is typically given in a 3 credit hour, 16 week format. In Appendix A, a timeline for the course can be seen. This timeline gives the order topics are covered, and shows when projects are due. The different categories of assessments are detailed below. The first three types of assessments listed below are fairly standard for most courses. The last type, projects, are how this course attempts to slowly introduce students to engaging in research.

2.1 Worksheets

Worksheets are designed to provide practice problems on the different topics covered. When students turn in a worksheet, its solutions will be made available to them. Since the purpose of these assessments is to give immediate feedback on whether or not a student is on the right track, worksheets are graded on completion instead of accuracy. In some instances, the worksheets are review guides for the exams.

2.2 Homework Assignments

Homework assignments are designed to assess student understanding on the different topics covered. These assessments are preceded by one or more worksheets, which should allow students to verify they are on the right track and have a basic understanding of the topics before working on more complicated problems on the homework. Unlike worksheets, these assignments are graded for accuracy.

2.3 Exams

The exams are intended to be summative assessments of the material learned up until that point in the semester. When traditional exams are given, the course has three exams: the first around week 5 covering the basic statistics, graphical summaries, and basic probability concepts; the second around week 10 covering various discrete and continuous random variable distributions, the normal distribution, and confidence intervals; and the third around week 15 covering various types of hypothesis tests. In this format, there is also a cumulative final exam given during week 16.

In the past, this course has used mastery-based testing instead of traditional examinations. In mastery-based testing, students are given multiple attempts to try each concept, but must answer the concept nearly perfectly in order to receive credit. The course is divided into 16 concepts, and students are given additional attempts to try concepts between each exam period. The main benefit of this style of testing is that students are encourage to improve upon topics they were not comfortable with previously, and will continue to study concepts that they have not yet passed. This type of examination works well for this course, but has not been implemented in multiple semesters since this type of assessment does not translate well to online format when required to make a sudden change during the semester. For more information on using mastery-based testing in mathematics courses, please see [Coll].

2.4 Projects

In this course, there are six different projects. The first five are meant to be more guided, with students being shown the ropes for the different parts of performing exploratory data analysis. The sixth and final project is meant to encourage students to go through the entire process of collecting and cleaning data, performing exploratory data analysis, and using the data to use hypothesis testing to test claims. Each project is designed to be an application of the topics learned in class. Some of the data sets are available online, and links to those data sets will be made available in Appendix C.

In the first project, students are tasked with finding summary statistics of given data. They must also import and graph stock data, and begin looking at data sets on their own to get some ideas of the types of questions they may want to answer later on. In the second project, students are given a data set and asked to create different least-squares regression lines predicting a particular variable with various explanatory variables. Students are encouraged in this assignment to then reflect on the benefits and drawbacks of each model they created. The third project is meant to be a way for students to put the ideas from the first two projects together. In this project, students are given a series of incorrect

statistics and graphs for a particular data set and are asked to verify and correct the information they were provided with. The purpose of this assignment is not only to review the programming they have previously completed, but to develop the ability to communicate their findings.

The fourth project has students graphing a variety of different random variable distributions. In particular, much time is spent on normal distributions. In the fifth project, students are given a data set and asked to run a hypothesis test on a category of their choosing. The data set given is purposefully messy, so that students need to engage in some data cleaning and explain the choices they made during that process.

The sixth project is a culmination of the work that students have done up until that point in the semester. An example of a students' final project can be found in Appendix B, along with the questions that are typically asked of students. The project is broken into three main sections. In the first section, students are meant to find their data and their claims, and perform some exploratory data analysis. In the second section, students are to use the data they have to perform a complete hypothesis test on their claims. In the final section, students are asked to reflect on their experience, commenting on how their results compared to their expectations and plans for what they would do differently in the future.

3 Analyzing the course through the CURE Framework

In [Auch], there is care taken to differentiate a CURE from traditional courses, inquiry courses, and internships. Auchincloss et al. provide a framework categorizing each different type of program based on five dimensions: use of science practices, discovery, broader relevance or importance, collaboration, and iteration. Using this article as a baseline, we will categorize how this course fits into the definition of being a CURE along the five dimensions given.

3.1 Use of science practices

This dimension focuses on the activities performed by the students, as well as who chooses which methods are used. In their final projects, students must collect, clean, and analyze data, as well as use it to perform a critical analysis of the arguments made by others. The students choose which of the statistical tests they have been taught to use in order to analyze their data. As such, this course falls under the "CURE" category for this dimension. Students in this course find themselves engaging in a number of statistical practices:

- Recognizing common sources of bias in surveys and experiments.
- Constructing and interpreting numerical summaries and graphical displays of data.
- Computing the least-squares regression line and using it to make predictions.
- Performing hypothesis tests and interpreting their results.
- Finding, importing, and cleaning data sets.
- Researching claims claim about data, and finding appropriate data to test those claims.

3.2 Discovery

This dimension focuses on how new the information discovered is, and to whom. Oftentimes, students are picking their own topics to pursue for their final projects. Many students pick topics that neither they nor the instructor have enough background in to know the results ahead of time. As such, the outcomes and findings are often unknown and novel to the students and instructor. Having said that, the information they discover is often not novel to the world. With that taken into consideration, this course falls under the "Inquiry" category for this dimension.

3.3 Broader relevance or importance

This dimension focuses on whether or not the results extend beyond the scope of the course. While some students will choose to continue to research the topics they have picked from their final projects, the results from those final projects are often not shared with the community or extend beyond the scope of the class. This means that this course falls under the "Traditional" category for this dimension.

3.4 Collaboration

This dimension focuses on how students are brought together and how the instructor engages with the students. During the final project, the instructor will provide insights to the students, but not guide them to a specific conclusion or line of reasoning. The students will occasionally bounce ideas off of one another, but each works on their own projects. This course falls under the "Traditional" category for this dimension.

3.5 Iteration

This dimension focuses on how messy the data is, and how often students end up repeating and revising their approaches. Because of the variety of topics and the various forms and amounts of information available for each of them, messy data is an inherent part of the process. While the students are asked to reflect on ways to iterate the process and make changes in the future, they often do not have the opportunity to do so in this course. With that being said, this course falls under the "Inquiry" category for this dimension.

4 Conclusion and Remarks

Many of the choices for this course are based on where it is situated within the overall curriculum at Lewis University. Because the students of this course widely range in major, there are some things that are not done that could be given the various backgrounds of the students. Given different constraints, there are many different changes that could be beneficial to those running a course similar to this one.

- In a course with more students in a particular discipline, it could be worthwhile to partner this course with other courses from that discipline in order to collect and analyze data from that domain. This would be beneficial to the students of this course, as it would give them extra practice with their skills and allow their results to stretch beyond the scope of this course.

- In a course that was part of a sequence, surveys could be designed by the follow-up course and then analyzed by the introductory course. This would allow students a more robust view of the overall research process over two semesters, as they currently only have access to data from popular data base websites. Another benefit to a course where you knew students would complete the sequence would be the opportunity for more time to iterate on their initial results and attempt to extend their work beyond the scope of the course.

- In a course that had more students with a programming background, more advanced web scrapping techniques could be added. This would result in students dealing with more messy data and develop their skills in wrangling data, which is a major part of the data analysis process.

- Instructors who wish to have their students' results extend beyond the course may opt to require students present their findings at a small local conference. This is something that some capstone courses at our institution require, but this course currently does not.

- Instructors may wish to find more ways to get the students working together. It could be that students are paired or grouped together on a particular topic and asked to analyze the information through different lenses. Such an endeavor may require more time for data collection, or require that instructors assign the topics so that they are verified to be robust enough for multiple lenses of analysis.

- Of the topics covered, ones that could reasonably removed involve continuous random variables (which could open this course up to being a general education course with no Calculus I requirement) and confidence intervals. These would still allow students to see the main concepts needed in order to perform hypothesis tests. This time could be spent on the final project, on providing opportunities for collaboration, building more iteration into the process, or requiring students to present their findings.
- One of the main topics this course is missing is analysis of variance (ANOVA). This would give students a valuable tool for analyzing multiple populations, and allow them to engage in deeper analysis of their data. If an instructor finds themselves with enough time, I would recommend adding this topic to their course.

Ultimately, this course hopes to introduce students to undergraduate research. Lewis University has a few opportunities for undergraduates to showcase their research with the S.T.E.M. Undergraduate Research Experiences (SURE) and annual Celebration of Scholarship. SURE is a summer research experience where students applied to engage with Lewis University professors over the course of a few months, culminating in a symposium at the end of the summer. The Celebration of Scholarship is an annual event where students present posters and talks on research they have been working on throughout the year. A major goal of this course is to get students to get an introductory sense for the tools they may use in their own research experiences, as well as give them a sense of the types of topics they may find themselves interested in researching.

5 Appendix A - Weekly Schedule

Week #	Topics	Exams and Projects
1	Sampling, types of data, design of experiments, bias	
2	Graphical summaries of data, measures of center, spread, and position	
3	Correlation, least-squares regression line	Project 1
4	Basic concepts of probability, probability rules	Project 2
5	Discrete probability distributions	Exam 1
6	Continuous random variables	Project 3
7	Normal distributions	
8	Normal distributions (continued)	
9	Confidence intervals	Project 4
10	One-sample hypothesis testing	Exam 2
11	One-sample hypothesis testing (continued)	
12	Two-sample hypothesis testing	Project 5
13	Goodness-of-fit hypothesis testing	
14	Hypothesis tests for independence and homogeneity	
15	Review	Exam 3, Project 6
16	Finals Week	Final Exam

6 Appendix B - Example of a Student Final Project

Project 6 - Answering Your Own Questions

Applied Probability and Statistics

Directions: Use Python to answer the following questions. Type your answers to the questions in a Word document or PDF. **Include screen shots of any code used to answer your questions and of any graphs created.** Make sure to cite your sources (both where your data came from, as well as where you obtained your hypotheses).

At this point in the semester, you have found a number of different probabilities and learned how to perform statistical tests to answer questions that others have posed to you. Now, it's your turn to pose the questions and answer them!

I. Setting the Stage

1. Find a data set that interests you, using Kaggle or some other web site. Provide both the name of and the URL to that data set, and then download it.

   Data set: MLB Expanded Replay Reviews from 2014-2019. Link: https://www.kaggle.com/jacobgideon/mlb-expanded-replay-reviews-from-20142019

2. In previous tutorials and projects, we have needed to make small changes to the data set in order to be able to use it. For example, we have deleted columns, dealt with missing values, and turned columns that had numbers as text into columns that had numbers instead. Did you need to make any of those changes to your data after importing it? If so, describe them here. (**Note: It may be worthwhile to save this question until near the end, after you have completed the other parts of the project.**)

   Yes, I did make some changes to my data after importing it. One of the changes I made dealt with information found in the 'date' column. This column contained the full date that the replay took place (example: 3/21/2019), but I was only looking for the year in which these events occurred.

So, based on the values in the 'date' column, I created new columns that specified whether that play occurred during that specific year.

```
#Determining dates where a video review took place in each year.

count2014 = df['date'].str.contains('2014')
count2015 = df['date'].str.contains('2015')
count2016 = df['date'].str.contains('2016')
count2017 = df['date'].str.contains('2017')
count2018 = df['date'].str.contains('2018')
count2019 = df['date'].str.contains('2019')
```

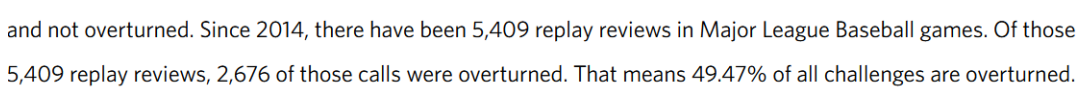
```
#Adding columns to specify whether each 'date' occured during the given year.

df2 = df

df2['is2014'] = count2014
df2['is2015'] = count2015
df2['is2016'] = count2016
df2['is2017'] = count2017
df2['is2018'] = count2018
df2['is2019'] = count2019
```

Another change I made to this data was removing the 'notes' column, as this column did not contain any data that was useful/necessary for this project.

```
del df['notes']
```

3. Find two claims about your data set on the internet, and include a screen shot of the claims and cite them. These could be official claims made in a paper or by a researcher, a tweet, a comment on social media, etc. Note that the individual(s) making these claims need not be an expert in the material.

Claim 1: From 2014-2017 seasons, 49.47% of challenges have been overturned.


and not overturned. Since 2014, there have been 5,409 replay reviews in Major League Baseball games. Of those 5,409 replay reviews, 2,676 of those calls were overturned. That means 49.47% of all challenges are overturned.

Source: https://www.samford.edu/sports-analytics/fans/2017/the-numbers-behind-replay-reviews-and-why-theyre-good-for-baseball

Claim 2: In 2017, the number of overturned calls dropped in comparison to 2016.


NEW YORK (AP) — The number of replays in Major League Baseball dropped this year along with the percentage of overturned calls.

Source: https://www.usatoday.com/story/sports/mlb/2017/10/01/number-of-replays-overturned-calls-drop-in-mlb/106223230/

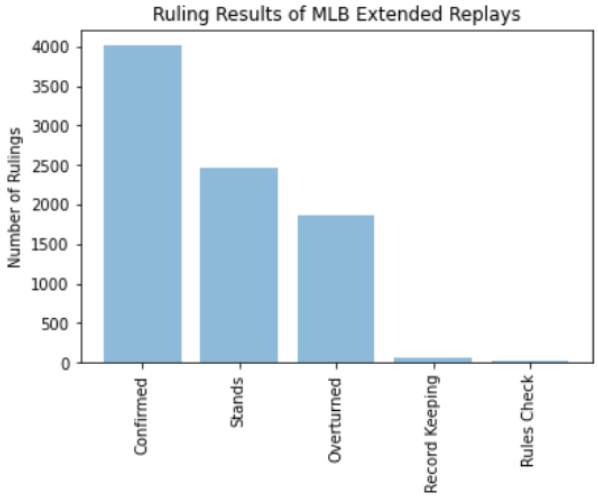
4. We have mentioned in the past that sometimes our sample's information cannot be extended to the population if it is not truly a simple random sample. For example, we have had an example where we wanted to use data collected about only third graders to make generalizations about all elementary school children, but were not able to do so because our sample was not truly representative of them all. Does your data set sample the correct population in order to make generalizations about the claims you found? How can you tell? (Regardless of the answer, you should use the data you have found to complete this project.)

   Yes, my dataset does sample to correct population for the claims that are made. My claims relied on data taken from the 2014 – 2017 Major League Baseball seasons. Since my dataset does contain information from these seasons, it can be used to make generalizations about this population.

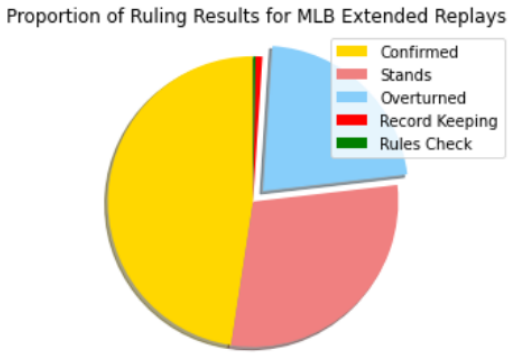
5. For each variable that has a claim, create a visualization that showcases it. This could be a box plot, a histogram, a bar graph, a pie chart, a scatter plot, or any number of other graphs. If you aren't sure if it should count, just ask!

   The variable that is looked at in both claims is the challenge ruling from MLB Expanded Replay Reviews. The following bar graph shows all of the results that are possible from a replay review, and the frequency of those results occurring. The proportion of these events are show using a pie plot.

```
#Bar Graph: Variable 'Ruling'

categories = df.ruling.unique()
amounts = df.ruling.value_counts()
axis_space = np.arange(len(categories))

plt.bar(axis_space, amounts, align = 'center', alpha = 0.5)
plt.xticks(axis_space, categories)
plt.xticks(rotation = 90)
plt.ylabel('Number of Rulings')
plt.title('Ruling Results of MLB Extended Replays')
plt.show()
```



```
#Pie Plot: Variable 'ruling'
labels = df.ruling.unique()
sizes = df.ruling.value_counts()
colors = ['gold', 'lightcoral', 'lightskyblue', 'red', 'green']

explode = (0, 0, 0.1, 0, 0)  # explode 3rd slice


patches, texts = plt.pie(sizes, explode=explode, colors=colors, shadow=True, startangle=90)
plt.legend(patches, labels, loc="best")

plt.title('Proportion of Ruling Results for MLB Extended Replays')
plt.axis('equal')
plt.show()
```



II. Testing the Claims

6. For **each** of the claims that you found, perform a complete hypothesis test. In other words, do the following for each claim:

   (a) State the claim again, for ease of reference.

   (b) Construct a null and alternative hypothesis using the claim.

   (c) Based on the information you have available, determine the appropriate test statistic.

   (d) Calculate your test statistic.

   (e) Calculate both a p-value and a critical value based on the significance level of $\alpha = 0.01$.

   (f) Based on the previous part, did you reject the null hypothesis?

   (g) State your conclusion in non-technical terms.

Claim 1

(a) From 2014 – 2017, 49.47% of challenges have been overturned.

(b) $H_0 : p = 0.4947; H_1 : p \neq 0.4947$

(c) Test statistic: z-score

(d) T = -1.4480525074021835

```
# Counting number of dates where video review took place 2014-2017
datecount = 0

for i in range(len(count2014)):
    if count2014[i] == True:
        datecount += 1

for i in range(len(count2015)):
    if count2015[i] == True:
        datecount += 1

for i in range(len(count2016)):
    if count2016[i] == True:
        datecount += 1

for i in range(len(count2017)):
    if count2017[i] == True:
        datecount += 1

print(datecount)
```

```
5614
```

```
overturnedCount = 0
for i in range(len(df2)):
    if df2.is2014[i] == True and df2.ruling[i] == 'Overturned':
        overturnedCount += 1
    elif df2.is2015[i] == True and df2.ruling[i] == 'Overturned':
        overturnedCount += 1
    elif df2.is2016[i] == True and df2.ruling[i] == 'Overturned':
        overturnedCount += 1
    elif df2.is2017[i] == True and df2.ruling[i] == 'Overturned':
        overturnedCount += 1
```

```
# Part d, calculating the test statistic

p = 0.4947
n = datecount
p_hat = overturnedCount / datecount

test_statistic_numerator = p_hat - p
test_statistic_denominator = math.sqrt((p*(1-p))/n)

test_statistic = test_statistic_numerator / test_statistic_denominator

print('Test statistic: ' + str(test_statistic))
```

```
Test statistic: -1.4480525074021835
```

(e) P-value = 0.14760236638177401, Critical value (c.v.) = -2.575829303548901

```
# Part e, calculating p-value and critical value

p_value = p_value = stats.norm.cdf(test_statistic, 0, 1) * 2
print("p-value: " + str(p_value))
critical_value = norm.ppf(0.01/2)
print("Critical value: " + str(critical_value))
```

```
p-value: 0.14760236638177401
Critical value: -2.575829303548901
```

(f) Since $p > 0.01$ and $c.v. < T$, we fail to reject the null hypothesis.

(g) We do not have enough evidence to conclude that the proportion of calls overturned from 2014 – 2017 differs from 49.47%.

Claim 2

(a) In 2017, the proportion of overturned challenges dropped in comparison to 2016.

(b) $H_0 : p_1 \leq p_2; H_1 : p_1 > p_2$, where p_1 is 2016 and p_2 is 2017.

(c) Test statistic: z-score

(d) T = 1.7580544268077558

```
overturned2016 = 0
overturned2017 = 0

for i in range(len(df3)):
    if df3.is2016[i] == True and df3.ruling[i] == "Overturned":
        overturned2016 += 1
    elif df3.is2017[i] == True and df3.ruling[i] == 'Overturned':
        overturned2017 += 1

count2016 = df3['date'].str.contains('2016')
number2016 = 0
for i in range(len(count2016)):
    if count2016[i] == True:
        number2016 += 1

count2017 = df3['date'].str.contains('2017')
number2017 = 0
for i in range(len(count2017)):
    if count2017[i] == True:
        number2017 += 1
```

```
p_hat1 = overturned2016 / number2016
p_hat2 = overturned2017 / number2017
n1 = number2016
n2 = number2017
p_hat = (overturned2016 + overturned2017) / (n1 + n2)

numerator = p_hat1 - p_hat2
denominator = math.sqrt(p_hat * (1 - p_hat) * ((1/n1) + (1/n2)))

test_statistic = numerator/denominator
print("Test statistic: " + str(test_statistic))
```

```
Test statistic: 1.7580544268077558
```

(e) P-value = 0.039369124513787046, Critical value (c.v.) = 2.3263478740408408

```
#Right-tailed test

p_value = stats.norm.cdf(np.inf, 0, 1) - stats.norm.cdf(test_statistic, 0, 1)
print("P-value: " + str(p_value))
critical_value = norm.ppf(1-0.01)
print("Critical value: " + str(critical_value))
```

```
P-value: 0.039369124513787046
Critical value: 2.3263478740408408
```

(f) Since $p = 0.039 > 0.01$, we fail to reject the null hypothesis.

(g) We do not have enough evidence to conclude that the proportion of overturned challenges in 2017 was less than the proportion of overturned challenges in 2016.

III. Reflection

7. Did your findings match your expectations? Why or why not?

For Claim 1, the results did match my original expectations. Based on the hypothesis test, we were unable to reject the null hypothesis. This meant that we could not conclude that the proportion of

calls overturned was different from 49.47%, which is the result I expected since it falls in line with the original claim.

For Claim 2, the results did not match my expectations. For this claim, we were unable to reject the null hypothesis, meaning that we did not have enough evidence to conclude that the proportion of overturned challenges in 2017 dropped in comparison to 2016. However, based on the original claim, I was expecting that we would be able to reject the null hypothesis and conclude that the proportion of overturned challenges dropped in 2017.

8. Come up with at least one follow-up question that your answers made you want to know more about.

   Based on the result from Claim 2 (not enough evidence to conclude proportion of overturned challenges in 2017 dropped from 2016), I would like to run more tests to determine how this proportion compares to that of 2016. Specifically, I would like to run hypothesis tests that look at whether that proportion is higher than that of 2016, or whether they are different at all.

9. Were there any questions that your data set could not answer? Find at least one claim, or come up with your own.

   One question that this dataset cannot answer is whether games with a replay review were, on average, longer than games where no replay review took place.

10. If you were the person collecting the data on your own, what other pieces of information would you have wanted?

    If I were collecting this data on my own, I would want to have more information about the game following a review. Specifically, I would want information about whether a team scored after an overturned play, and whether a team went on to win or lose after an overturned call. This information could be used to help determine whether these replay reviews helped to change to outcome of the games.

    I would also like information on the length/amount of time the game took. With this information, I would attempt to determine whether games with a replay review are longer, on average, than games where no replay review took place.

7 Appendix C - Resources for Finding Data

In this section, the data sources that are typically used in the course will be shared, as well as the places that students often find their own data from. This is not meant to be an exhaustive list, but rather a starting point for those interested in incorporating data in their own courses.

- **Yahoo Finance:** This website offers instructions on downloading historical stock data into an CSV file. Students are asked to follow these instructions and create a time series plot for a stock of their choosing in Project 1. https://help.yahoo.com/kb/SLN2311.html
- **Kickstarter Data:** This data set from Mickaël Mouillé provides information on the success of various projects from Kickstarter, a crowdfunding website. In Project 5, students are asked to clean this data set and test hypotheses about a category of their choosing. https://www.kaggle.com/kemical/kickstarter-projects
- **Kaggle:** This website hosts a variety of publicly available datasets for students to obtain information on. Students will often use this website when finding data for their final projects. The data

sets available have varying levels of messiness (for an example, see the Kickstarter Data above), sometimes causing students to need to clean their data set before use. https://www.kaggle.com/

- **Data.gov** This website houses the U.S. government's open data. This gives a variety of data sets for students to choose for their final project, and may be useful in providing data sets for smaller projects leading up to the final project. The data sets here are usually fairly clean, but may not always have easy access to CSV files, leading students needing to copy and paste much of the information they find. https://www.data.gov/

References

[Auch] Auchincloss, L., et. al. (2014). Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. CBE-Life Sciences Education, 13, pp. 29–40. https://www.lifescied.org/doi/full/10.1187/cbe.14-01-0004

[Ber] Berkove, E. (2013). Service-Learning in a Capstone Modeling Course. PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies, 23(6), 507–518. https://doi.org/10.1080/10511970.2013.764367

[Cam] Camenga, K. A. (2013). Developing Independence in a Capstone Course: Helping Students Ask and Answer Their Own Questions. PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies, 23(4), 304–314. https://doi.org/10.1080/10511970.2013.764363

[Coll] Collins, J. B., et. al. (2019) Mastery-Based Testing in Undergraduate Mathematics Courses, PRIMUS, 29:5, 441-460, DOI: 10.1080/10511970.2018.1488317

[WL] Weston, T. J., & Laursen, S. L. (2015). The Undergraduate Research Student Self-Assessment (URSSA): Validation for use in program evaluation. CBE-Life Sciences Education, 14(3), ar33. DOI 10.1187/cbe.14-11-0206

Department of Engineering, Computing, and Mathematical Sciences, Lewis University, Romeoville, IL.
Email address: msmith42@lewisu.edu