PIG Project Report (Statistical Modeling)

What was the outcome of your project and how did it align with your original goals as stated in your application?

To enhance Math 248, we reviewed several textbooks and chose a course structure mirroring Math 247's successful flipped classroom and active learning approach. This update better aligns with Guidelines for Assessment and Instruction in Statistics Education (GAISE) and complements the use of *Intermediate Statistical Investigations* (Tintle et al.), the second textbook in the series that we currently use in Math 247. While we explored using GitHub and GitHub Classroom, technical integration issues with Canvas arose. Despite tabling this implementation for now, we updated the course labs with more current R code from the textbook's explorations and created a Chapter 1 exploration-style worksheet to facilitate material review from previous courses in a flipped classroom setting. The final project required only minor adjustments: a more detailed explanation of project goals, a dataset selection guide, and expanded resources to support a wider range of modeling techniques. Finally, the syllabus was modified to reflect the course's updated learning goals and the project's objectives while allowing for flexibility in instructor preference regarding course assessment practices (exams vs. quizzes).

What sense do you have at this point that your project has enhanced student learning and what will you be looking at in the future to know whether the project was successful?

We anticipate that Math 248's adoption of the familiar flipped classroom structure from Math 247 will ease the transition for students, aligning with pedagogical best practices and providing students with diverse introductory statistics backgrounds (such as ECON 227) with a valuable active learning experience. However, as this structure wasn't implemented this fall, its impact on student learning and understanding will be assessed during the next semester it is offered. Future evaluations include analyzing student progress in their conceptual understanding of statistical models, potentially through their performance on senior oral examinations in Econ-Math and Math-Stat.

What were the limitations or failings of this project, and how, in retrospect, might they have been better addressed or remedied?

The project's timeline was longer than anticipated, primarily due to more time spent evaluating textbook options than initially projected. Furthermore, we were unsuccessful

in integrating GitHub Classroom with Canvas in a test environment. Contacting WCTS earlier in the process might have avoided these integration challenges.

How do you envision sharing the results of your work with other colleagues at Whitman (or elsewhere)?

We met with Tim Parker to explore potential collaborations. We hope to continue this work and collaborate on future data analysis projects. To ensure consistent implementation and support for students, the revised course materials will be shared with all future Math 248 instructors.

MATH 248 Project Description

Due dates:

1. **Initial proposal: TBD by 11:59 pm.**
2. **Literature Review: TBD by 11:59 pm.**
3. **Project Report: TBD by 11:59 pm.**
4. **Project Presentation: TBD by 11:59 pm.**
5. **Letter of Learning: TBD by 11:59 pm**

The goal of the project is for you to research a statistical modeling technique or topic that has not been presented in class and won't be covered in this class. You will need to apply that method to a particular question of your choosing. You will need to first find the dataset of interest and then write up a project proposal describing the variables in your data as well as several potential research questions that you could answer with the data. After that, I will provide you with suggestions of what statistical technique could be used to answer the research question of interest. Make sure that the dataset is interesting to you! The data should have **at least 200 observations** and **at least 5 useful variables**.

The Initial Proposal should include:

- **Description of the Data:**
  - Identify the source of the dataset (the dataset will be submitted separately on Canvas).
- **Variable Descriptions:**
  - Provide numerical and graphical summaries appropriate for the variable types.
- **Research Questions:**
  - Present at least 2-3 distinct potential research questions.
  - Note that we will focus on one research question for the remainder of the project.
- **Helpful resources:**
  - Potential Data Sources
  - Library Data Resources Page

The initial proposal must be produced with R Markdown or Quarto document and submitted as a PDF file by TBD at 11:59 pm.

The Literature Review should include:

- **Everything from the Initial Proposal** (with edits as suggested or necessary) which will become your Introduction section in the final report.
    - Description of the data, including the source.
    - Description of important variables (numerical and graphical summaries appropriate for the variable type)
    - The specific research question we have identified for your methodology. Any resources/citations used to identify this question should be included.
- **Methodology section** where you will explain the modeling technique you have been assigned with citations to resources used to learn about it.

Literature review must be produced with R Markdown or Quatro document in the academic writing format discussed in class and resources available on Canvas course webpage and submitted as a PDF file by TBD at 11:59 pm.

Project Presentation should include:

- **Introduction**. An explanation of the data and questions related to the data. What problems or questions did you set out to investigate? How were the data collected?
- **Methodology**. Explain the statistical modeling technique. When is it used? What are the model assumptions? Can you put it in context with other techniques we have learned in class? (In other words, why do we need a new technique?)
- **Results and conclusions**, the summary, and the presentation of your data analyses. What did you find out? This might include tables, graphs, or verbal summaries.
- **Discussion and critique**. What did you learn about the problem or question you set out to investigate? What were the weaknesses and strengths of your analysis and this method?
- **Bibliography**

Project Presentation can be produced with R Markdown, Quatro, PowerPoint, Google Slides, etc. Make sure to submit a PDF file of your presentation by TBD at 11:59 pm.

Project Report should include:

1. **Introduction**. An explanation of the data and questions related to the data. What problems or questions did you set out to investigate? How were the data collected? This is where you include **Everything from the Initial Proposal.**
2. **Methodology**. Explain the statistical modeling technique. When is it used? What are the model assumptions? Can you put it in context with other techniques we have learned in class? (In other words, why do we need a new technique?). Include citations used in the **Methodology** section of the **Literature Review**
3. **Results and conclusions**, the summary, and the presentation of your data analyses. What did you find out? This might include tables, graphs, or verbal summaries.
4. **Discussion and critique**. What did you learn about the problem or question you set out to investigate? What were the weaknesses and strengths of your analysis and this method?
5. **R code to perform your modeling technique**. Make sure that all the code and output is visible within your report.
6. **Bibliography** with citations in paper.

Project Report must be produced with R Markdown or Quatro document and submitted as a PDF file by TBD at 11:59 pm for grading. You will also need to publish your final report on the web with Quatro/Markdown and include the link to it in the PDF submission of the report.

The breakdown of the project component grades is as follows:

Initial Proposal	**20%**
Literature Review	**20%**
Project Presentation	**30%**
Project Report	**30%**

Letter of learning:

This letter should summarize your experience during the project. You need to submit your individual letter of learning on Canvas *by TBD*. A missing letter of learning will result in a 5-point deduction from your written report grade. I reserve the right to take the information from your letter of learning and use it for your final project grade.

The letter of learning should include, but is not limited to the following:

- What was the most interesting/meaningful part of the project?
- What was the most challenging part of the project?
- What strategies have you employed in this project? What worked? What didn't?
- If you worked with a partner, evaluate the contributions of each of the members of your team

Math 248 Statistical Modeling

Course Description

In students' first course in Statistics, the focus was on learning about the process of conducting statistical investigations, most likely investigating research questions involving one variable or the association between two variables. This involved (1) identifying a research question, (2) gathering data to answer the research question, (3) examining the data, (4) testing hypotheses and estimating parameters using tests of significance and confidence intervals, (5) drawing conclusions about generalization and causation, and finally (6) reflecting on possible next steps in the research process. This statistical investigation method mirrors the scientific method and is also utilized throughout this second course in Statistics. However, now the focus expands to include explicit consideration of both study design and statistical analysis strategies for investigating the simultaneous impact of two or more explanatory variables on a response variable of interest. Throughout this course, the focus will remain on the concept of explaining variation in a response variable of interest and will emphasize a single, cohesive framework for thinking about explaining variation across study designs and variable types.

Class Location and Time

Tuesdays and Thursdays for 1:20

Textbook

Intermediate Statistical Investigations, Tintle, Chance, McGaughey, Roy, Swanson, Vanderstoep

Data analysis software

Applets and R. We will use R through Posit Workbench

Other technology

- GitHub Classroom
- Rmarkdown
- Quarto

Learning Outcomes

- Upon successful completion of this course, students will be able to:
- Choose, fit, assess, and apply appropriate statistical models to solve data-based problems.
- Understand and explain the limitations of statistical analysis and its interpretation.
- Employ statistical software effectively to perform data analysis.
- Present statistical analyses clearly and accurately in both technical and non-technical formats.
- Distinguish between statistical significance and practical significance and interpret both appropriately.
- Read, write, and critically evaluate statistical reports.
- Critically assess the quality of data, distinguishing between good data and "not-so-good" data.
- Independently learn how to use a new statistical modeling tool, demonstrated through a course project.

Course Content

This course provides a comprehensive introduction to statistical modeling techniques, emphasizing their application within the context of sound study design.

- Simple Linear and Multiple Regression. This section includes:
    - Residual analysis
    - Assessment of lack of fit
    - Tests for normality
    - Use of indicator variables
    - Variable selection techniques.
- Analysis of Variance (ANOVA). This section includes:
    - One-way classification
    - Two-way classification
    - Main effects and interactions
    - Residual analysis
- Controlling additional sources of variation. This section includes:
    - Paired data
    - Randomized Complete Block Design
    - Effects of covariation
- Categorical response. This section includes:
    - Comparing proportions
    - Logistic Regression

Course Assessment

1. Reading Quizzes

There will be reading quizzes based on the assigned textbook readings. These are meant to make sure that you complete the assigned reading and help you formulate questions about the material to ask your instructor

2. Explorations

Each class period we'll start working on an exploration designed to investigate new material. Most explorations are designed to take between 45 and 90 min on average to complete, so you'll most likely need to finish them outside of class. Explorations are due on Canvas the following day. Explorations are graded on completeness.

3. Midterms or Quizzes - depending on pedagogical needs or approaches
4. Final Project
    a. Proposal
    b. Presentation
    c. Report

Math 248: Worksheet: Review

Your Name Here

02/10/2025

Worksheet: Comparing Statistical Models - A Student Practice

Objective: To learn how to choose the appropriate statistical model for different scenarios and perform the analysis using R.

Instructions: For each scenario, answer the questions to determine the correct statistical model. Then, complete the R code to perform the analysis and interpret the results.

Part 1: Scenario 1 - Single Mean, Paired Data, or Two-Independent Samples?

Scenario: We measure the height of 50 students at the beginning of the school year and again at the end of the year. We want to determine if there's a statistically significant change in average height over the year.

Guiding Questions:

1. **Response Variable:** What is the variable being measured?
2. **Explanatory Variable:** Is there an explanatory variable influencing the response variable? If so, what is it?
3. How many groups of data do we have?
4. Are the measurements made on the same individuals or different individuals?
5. Are we comparing the means of two independent groups or assessing a change within a single group?
6. What model is appropriate? (Single-mean, paired data, or two-independent samples?)

R Code

Generate data

```
set.seed(123)
n_students <- 50
height_beginning <- rnorm(n_students, mean = 150, sd = 10)
height_end <- height_beginning + rnorm(n_students, mean = 2, sd = 3) # Simulate growth

# Create data frame
paired_data <- data.frame(
  StudentID = 1:n_students,
  HeightBeginning = height_beginning,
  HeightEnd = height_end
)
```

```
# Calculate the difference in heights
paired_data$HeightDifference <- paired_data$HeightEnd - paired_data$HeightBeginning
```

7. Perform the appropriate test

```
# Paired t-test
t_test_paired <- t.test(paired_data$HeightBeginning, paired_data$HeightEnd,
                        paired = TRUE)
print(t_test_paired)
```

```
##
##  Paired t-test
##
## data:  paired_data$HeightBeginning and paired_data$HeightEnd
## t = -6.3497, df = 49, p-value = 6.749e-08
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.211201 -1.667249
## sample estimates:
## mean difference
##       -2.439225
```

8. Check assumptions for paired t-test. Based on the assumption checks, is this test is appropriate for these data? Explain your reasoning.

```
## Normality of the differences
shapiro.test(paired_data$HeightDifference) # Shapiro-Wilk test for normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  paired_data$HeightDifference
## W = 0.99073, p-value = 0.9618
```

9. Interpret the p-value and confidence interval in the context of the scenario. What does your conclusion mean in terms of student height changes over the school year?

Part 2: Scenario 2 - Single Mean, Two-Independent Samples, or Paired Data?

Scenario: We want to compare the average weight (in kg) of adult men and adult women. We have independent samples of men and women.

Guiding Questions:

1. **Response Variable:** What is the variable being measured?
2. **Explanatory Variable:** What is the explanatory variable that might influence the response variable?
3. How many groups of data are there?
4. Are the groups independent or related?
5. What is the objective of the analysis (to compare the means of two groups or assess change within a group)?
6. What model is appropriate? (Single-mean, two-independent samples, or paired data?)

R Code

Generate Data

```
set.seed(123)
n_men <- 100
```

```r
n_women <- 100
weight_men <- rnorm(n_men, mean = 75, sd = 10)
weight_women <- rnorm(n_women, mean = 60, sd = 8)

weight_data <- data.frame(
  Weight = c(weight_men, weight_women),
  Gender = factor(rep(c("Male", "Female"), each = c(n_men, n_women)))
)
```

7. Perform two-sample t-test

```r
# Use if assumptions are met
t_test_weight <- t.test(Weight ~ Gender, data = weight_data, var.equal = TRUE)
print(t_test_weight)
```

```
##
##  Two Sample t-test
##
## data:  Weight by Gender
## t = -14.011, df = 198, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -19.12400 -14.40486
## sample estimates:
## mean in group Female   mean in group Male
##             59.13963             75.90406
```

8. Check assumptions for two-sample t-test. Based on the assumption checks, is this test is appropriate for these data? Explain your reasoning.

```r
#Normality
shapiro.test(weight_data$Weight[weight_data$Gender == "Male"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight_data$Weight[weight_data$Gender == "Male"]
## W = 0.99388, p-value = 0.9349
```

```r
shapiro.test(weight_data$Weight[weight_data$Gender == "Female"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight_data$Weight[weight_data$Gender == "Female"]
## W = 0.97289, p-value = 0.03691
```

```r
# Homogeneity
leveneTest(Weight ~ Gender, data = weight_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  3.0866 0.08048 .
##       198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. Interpret the p-value and confidence interval. What does your conclusion mean in terms of

average weight differences between men and women?

Part 3: Scenario 3 - Single Mean, Two-Independent Samples, or Paired Data?

Scenario: A researcher wants to know if the average score on a cognitive test is significantly different than a known population average of 50. A sample of 25 participants takes the test.

Guiding Questions:

1. **Response Variable:** What is the variable being measured?
2. **Explanatory Variable:** Is there an explanatory variable?
3. How many groups of data are there?
4. Are the groups independent or related?
5. What is the objective of the analysis (to compare the means of two groups or assess change within a group, or compare to a known value)?
6. What model is appropriate? (Single-mean, two-independent samples, or paired data?)

R Code

Create the dataset

```
set.seed(123)
n_participants <- 25
test_scores <- rnorm(n_participants, mean = 55, sd = 5)
```

7. Perform a one-sample t-test to determine if the mean of the test_scores is significantly different from 50.

```
t.test(test_scores, mu = 50) # One-sample t-test
```

```
##
##  One Sample t-test
##
## data:  test_scores
## t = 5.1053, df = 24, p-value = 3.184e-05
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  52.87939 56.78731
## sample estimates:
## mean of x
##  54.83335
```

8. Check assumption of normality for the one-sample t-test. Based on your assumption check, is a one-sample t-test appropriate here? Explain your reasoning.

```
# Normality
shapiro.test(test_scores)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  test_scores
## W = 0.96745, p-value = 0.5812
```

9. Interpret the p-value. What does your conclusion mean in terms of the average cognitive test score compared to the population average?