Midterm Exam - MATH 141

10/13/2021

Exam starts on 10/13/2021 9:00 AM and ends on 10/14/2021 at 9:00 AM (24 hours).

You must submit your midterm exam into Gradescope within the 24-hour window.

Instructions:

- Please provide complete answers/solutions for each question/problem.
- If it involves mathematical computations, please provide your reasoning and/or detailed solutions.
- Please abide by the Reed College Honor Principle.
- This exam is a take-home exam, you can use all of the notes you have, textbook, lecture slides, homework solutions, lab solutions, the course website, and the internet. Group work and collaboration is encouraged. Use all of the resources you have and what you need to answer the exam questions. Each student must take responsibility and ownership of their work and submit their work individually.
- If you have any technical questions, please send them to the instructor as a direct message on Slack. For the sake of fairness, the instructor will not respond to conceptual or clarifying questions. Because we may be unable to answer your question during the time frame in which you are taking the exam, document the issue on the exam and then proceed to the next problem.
- If you can't figure out why a code chunk is preventing you from knitting the document, replace "r" at the top of the code chunk with "r eval = FALSE, echo = TRUE". The code will not be executed, but it will be printed in your pdf, earning you some partial credit.
- Please follow the general exam guidelines written in the course website.
- **Please save your work as pdf file(s), don't put your name in any part of the document, and submit it to the Gradescope page for this course. Your document upload will correspond to your name automatically in Gradescope.**

R Libraries

```
# loading packages
library(tidyverse)
library(dplyr)
library(ggplot2)
library(gghighlight)
library(infer)
```

I. Dice to meet you, stranger!

You received two unknown six-sided dice labeled A and B from a stranger. The stranger wants you to determine which one is an unfair die and to what degree - or else! Note that a fair dice has equal probabilities on all six sides which is $\frac{1}{6}$.

Which dice is unfair? Dice A or Dice B (or both)?

1. Load the `dice.csv` data set. This data set has two columns. The first column is the label of the dice and the second column is the outcome when rolling the dice. Determine the number of rolls for each die.
2. Generate some data where you roll a separate die n times where you know it is a fair die. This generated data set will act as the control. Print out your generated data set.
3. Consider getting a "1" as a "success" and the rest as "failures". For each die, what are the population parameters and what are the sample statistics?
4. For each die, construct a null and alternative hypothesis where you take the difference in proportions against your control die generated data. Consider getting a "1" as a "success" and the rest as "failures". Make sure to use mathematical symbols and its corresponding sentences.
5. Perform a hypothesis test with randomization for die A. What is the point estimate(s) and the p-value? Plot your simulations. Make sure to label them properly and shade the appropriate region.
6. Perform a hypothesis test with randomization for die B. What is the point estimate(s) and the p-value? Plot your simulations. Make sure to label them properly and shade the appropriate region.
7. For significance level of $\alpha = 0.05$, make a decision for each die and interpret the p-value.
8. What type of errors could you make in each die? Explain your answer in the context of the problem.
9. Construct a 95% confidence interval for dice A - considering an outcome of "1" as "success" - using Bootstrapping and percentiles. Interpret the interval in this context. Plot your simulations. Make sure to label them properly and indicate the interval in your plot.
10. Construct a 95% confidence interval for dice B - considering an outcome of "1" as "success" - using Bootstrapping and percentiles. Interpret the interval in this context. Make sure to label them properly and indicate the interval in your plot. Make sure to label them properly and indicate the interval in your plot.

II. You are so so-fish-ticated! Fish be with you.

Pretend that you a sophisticated fish chef and you collected data on different species of fish from a fish market. You measured the weight, height, width, and different lengths for each fish in your sample.

Can you predict the weight of the fish?

1. Load the `fish.csv` data set. How many rows and columns are there in this data set? [1]
2. Determine the type of each variable in the data set if they are numerical (discrete or continuous) or categorical (nominal or ordinal). Note that the `Weight` variables is in grams, the `Species` variable has 7 species, and the rest of the variables are in centimeters.
3. Using the `Weight` variable as the response variable, write out two linear equations where `Height` is used as the predictor for the first linear model and `Width` as the predictor for the second linear model. Note that when writing the linear equations use the symbols for the slope and intercepts for the population parameters.
4. Use the least squares regression (use the `lm` function) to find the best fit linear model for each equation defined from problem 3. Write out each equation with their corresponding slope and intercept. Note that when we fit the data into our model, we use the "hat" symbols to indicate the response the variable as the estimate or prediction.
5. Interpret the slope and intercept for each linear model.
6. Produce scatter plots for (a) `Height` vs `Weight` and (b) `Width` vs `Weight`. Include in your scatter plot the linear fit you just produced in problem 4. Describe the scattplot and the linear fit. How well does the linear model fit when you look at it visually?
7. Perform a residual analysis for the linear equation with `Weight` as the response and `Height` as the predictor. Does the histogram of residual and the residual scatter plot meets the assumptions of the linear regression? Compute the coefficient of determination R^2 and interpret this value. Determine from your residual analysis whether or not you might need to use a more complicated model.
8. Perform a residual analysis for the linear equation with `Weight` as the response and `Width` as the predictor. Does the histogram of residual and the residual scatter plot meets the assumptions of the linear regression? Compute the coefficient of determination R^2 and interpret this value. Determine from your residual analysis whether or not you might need to use a more complicated model.
9. Produce a bar plot where it shows the number of fish in each species. Which species has the most fish samples? Please label your plot properly.
10. Produce a grouped box plot where it shows the `Weight` for each species. What are the top 3 species with the highest median weight? Please label your plot properly. Is there an association between the `Species` and `Weight` variables? Explain your reasoning.

[1] The original source of the fish data we are using in this part is from www.kaggle.com/aungpyaeap/fish-market/version/2.