Save this homework file as: HW5\_LastName\_FirstName.docx

Save your R script file as: HW5\_LastName\_FirstName.R

You must use R for the analysis, but if you use Excel for anything else, save your file as: HW4\_LastName\_FirstName.xlsx

Be sure to include your name and homework 4 in the syntax comments. Add comments throughout your script so I will be able to follow your process better.

*-0.5 points will be deducted for not using comments in your syntax file.*

You may copy/paste R code in this homework to help you complete the questions.

Any R code will appear like this.

**Grading Rubric**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Q1 (a-d)** | **Q2a** | **Q2b** | **Q3a** | **Q3b** | **Q3c** | **Q3d** |
| **Points** | **2.0 (0.50 ea)** | **0.50** | **1.25** | **0.25** | **0.25** | **0.25** | **0.50** |

**Question 1**

Most people don't succeed in losing weight by dieting, and on average people don't lose any weight. Proponents of the quantized-self movement believe that by tracking eating behavior, using apps like MyFitnessPal, people will lose more weight. Pedro randomly selected 25 people from his university who were interested in dieting. He then had them use MyFitnessPal to track their eating behavior. He wants to understand whether using the app leads to greater weight loss.

1a) (Must answer all correct for points on question 1a.)

What is Pedro’s research question?

**Does usage of MyFitnessPal as a dieting technique lead to greater weight loss in comparison to dieting without the app?**

What is the predictor (independent variable)?

**The predictor is a participant’s use of the MyFitnessPal app.**

What is the outcome (dependent variable)?

**The outcome is how much weight a participant lost.**

What kind of dependent (outcome) measure is this?

**Ratio.**

1b) What should his null and alternate hypotheses be in words and in symbols?

**Ho: There is no effect or relationship between a participant’s usage of MyFitnessPal (x) and weight loss (y).**

**Ho: x = y**

**Ha: There is an effect or relationship between a participant’s usage of MyFitnessPal (x) and weight loss (y).**

**Ha: x ≠ y**

1c) Would he be more likely to conclude that there is a significant effect is he uses a one -tailed or two-tailed test?

**Pedro would be more likely to conclude that there’s a significant effect if he uses a one-tailed test. One-tailed tests are used when you’re interested in determining whether there’s a relationship between variables in a single direction (e.g., less than, greater than). Since Pedro is looking at whether the use of MyFitnessPal leads to greater weight loss, it makes more sense for him to use a one-tailed test.**

1d) Pedro found that after 6 weeks his sample of 25 had lost an average of 2 pounds. He set an alpha level of 0.10 at the start of the study and his p-value from the analysis is 0.12.

Are Pedro’s results statistically significant?

**No, Pedro’s results are not statistically significant because the p-value of 0.12 is greater than the alpha level set at 0.10 (0.12 > 0.10). Since the p-value is greater than the alpha, Pedro failed to reject the null hypothesis. This means his results suggests that there is no effect or relationship between MyFitnessPal use and weight loss.**

Is his group different from the rest of the population?

**No, Pedro’s group is not different from the rest population because his results are not statistically significant. Pedro’s group of participants were also selected randomly, so there’s not enough evidence to support the claim that his group his different from the rest of the population.**

What can you conclude about MyFitnessPal and weight loss?

**I can conclude that there is no effect or relationship between MyFitnessPal and weight loss. In other words, it’s not clear whether using MyFitnessPal as a dieting technique will lead to greater weight loss.**

*Do you believe Pedro’s findings to be substantively significant? (This question is not graded.)*

**No, I don’t think his results are substantively significant. Pedro only randomly selected 25 participants for his study. This is not nearly a large enough sample size to make substantive conclusions about a population, which in this case, is students at his university interested in dieting. Based on the prompt alone, I’m also not sure whether Pedro accounted for potential bias affecting his sample selection or data collection.**

**Question 2**

Professional athletes are often judged in terms of performance on a wide variety of metrics, which feed into a wide array of uses from fantasy leagues and gambling to AI development in video games. We will use the FIFA 2017 video game data as a proxy for actual soccer (football) player performance. Import the FIFA CSV file.

Mean composite scales are variables that represent a mean score built from other variables. Use R to construct a mean composite scale that represents a player’s kicking abilities. Use the following variables: Shot\_Power, Finishing, Long\_Shots, Curve, Freekick\_Accuracy, Short\_Pass, and Long\_Pass. Attach your new variable to the FIFA 2017 dataset. *See the R setup file for more on how to construct a mean composite scale with an example.*

The following questions will have you evaluate the new kicking ability scores of all FIFA 2017 players, with those of the top and bottom ranked football clubs: FC Barcelona and Longford Town respectively. [[1]](#footnote-1) The club names are the same in the R data. You will need to use subsetting techniques to isolate the scores for each team. You may use whichever subsetting approach you prefer.

2a) Report summary statistics below for this new kick scale variable for all FIFA players, FC Barcelona, and Longford Town clubs (PQ table with n, min, max, 1st & 3rd quartiles, median, mean, and sd).

Table

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated

2b) Calculate a 95% confidence interval for the mean for all FIFA players, FC Barcelona, and Longford Town. Do this by hand below and confirm your work using R. Round your answer to 2 decimal places.

**\*\* I think my calculations vary slightly because I rounded all of my calculations by two decimal places to make it easier to read/calculate by hand. \*\***

Text, letter

Description automatically generated**-- R –**

**# checking confidence intervals**

**t.test(fifa$kicking\_ability.mean)**

**# 95 percent confidence interval: 49.65762 50.12021**

**t.test(fifa\_barcelona$kicking\_ability.mean)**

**# 95 percent confidence interval: 53.84434 67.87860**

**t.test(fifa\_longford$kicking\_ability.mean)**

**# 95 percent confidence interval: 32.87401 42.54314**

What can you conclude from your confidence intervals?  *Not graded, be descriptive.*

**Based off of my confidence intervals, I can conclude that all FIFA players average a kicking ability of 49.89 +/- 0.24. FC Barcelona Club players, on the other hand, average a kicking ability of 60.86 +/- 6.86. Lastly, Longford Town Club players average a kicking ability of 37.71 +/- 4.68.**

**Question 3**

The General Social Survey has 6 questions that pertain to optimism. The variables are named LOTR1, LOTR2, LOTR3, LOTR4, LOTR5, and LOTR6.

3a) What type of variables are LOTR1 through LOTR6? (scale of measurement). Look up the full survey question asked for one of the six LOTR variables and enter it below. *Hint: This will be easier if you use the GSS online data explorer rather than the pdf codebook.*

Table

Description automatically generated**LOTR1-LOTR6 are factor variables.**

3b) Use the R code from the setup file to create an item scale, which allows us to do quantitative analysis on it. The new variable will be called LOTR.scale. *Examine the setup code to learn more about how we constructed this new variable.*

Using a PQ table, report the n, min, max, 1st & 3rd quartiles, median, mean, and sd.

3c) Use R to produce a histogram of lotr.scale with PQ labels and title and insert it below this question. Describe if you believe the data to be relatively normally distributed.

Chart, histogram

Description automatically generated

**This data seems to be normally distributed, but I would also note that there’s a slight left skew.**

3d) Use R to calculate a confidence interval for LOTR.scale. Use α = 0.01. Round your answer to 2 decimal places. You may find the confidence interval by hand or R, but do not need to do both.

*Hint: If you use an R function to find the confidence interval and you get NA in your output, then you’ll need to somehow tell the function to ignore the missing values; however, not all R functions can use the na.rm=TRUE option, so you might need to get creative with your code.*

**α = 0.01 = 2.576**

**Text, letter

Description automatically generated**

1. As reported using the top and bottom non-zero average rating on <https://www.futhead.com/17/clubs/?sort=-average_rating>. [↑](#footnote-ref-1)