Save this homework file as: HW6\_LastName\_FirstName.docx

Save your R script file as: HW6\_LastName\_FirstName.R

You must use R for the analysis, but if you use Excel for anything else, save your file as: HW6\_LastName\_FirstName.xlsx

Be sure to include your name and homework 6 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.*

**Grading Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total** | **A** | **B** | **C** | **D** |
| **Q1** | **1.00** | **0.10** | **0.15** | **0.50** | **0.25** |
| **Q2** | **1.00** | **0.25** | **0.50** | **0.25** | **--** |
| **Q3** | **1.00** | **0.50** | **0.50** | **--** | **--** |
| **Q4** | **0.50** | **0.25** | **0.25** | **--** | **--** |
| **Q5** | **1.00** | **0.25** | **0.25** | **0.25** | **0.25** |
| **Q6** | **0.25** | **--** | **--** | **--** | **--** |

**Question 1**

A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of the 10 cars were first tested with regular gas and then with premium gas. The mileages on both types of gas was recorded and the value are as shown below:

(Use alpha=0.05)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Regular | 16 | 20 | 21 | 22 | 23 | 22 | 27 | 25 | 27 | 28 |
| Premium | 19 | 22 | 24 | 24 | 25 | 25 | 26 | 26 | 28 | 32 |

Q1a) What type of t-test is it? Give your reasoning (0.25 points)

**This is a paired t-test because there’s a small sample size (n = 10) and because the cars were tested twice, once with regular gas and another time with premium gas. This study is looking to analyze the difference between two variables (regular gas vs. premium gas) for the same object (cars).**

Q1b) State your null and alternate hypothesis in words and symbols (0.25)

**Ho: There is no difference in means between usage of premium or regular gas on a car’s mileage. Ho = 0**

**Ha: There is a difference in means between usage of premium or regular gas and a car’s mileage. Ha ≠ 0**

Q1c) Enter these data values in R as two variables. Perform a t-test in R to determine if there is a difference in means between using regular gas vs premium gas. Report the p-value, t-statistic and degrees of freedom using the APA format (0.50 points)

*Hint: you can create a variable by entering a vector of raw data. For example:*

VariableName <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 0) # Change VariableName to whatever you want

**t(9) = -4.47, p < 0.05**

**-- R --**

**# insert data**

**regular\_gas <- c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)**

**premium\_gas <- c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)**

**# perform paired t-test**

**t.test(regular\_gas, premium\_gas, na.rm=T, pooled=F, paired=T)**

Q1d) Based on your results, what would you conclude? Be sure to set your answer in the context of the problem. (0.25 points)

**Using an alpha of 0.05, I can conclude that the results favor the alternative hypothesis, allowing me to reject the null hypothesis. The p-value, which is 0.00155, is less than the alpha of 0.05. This means my results were statistically significant, further reinforcing that there is indeed a difference in means between usage of premium or regular gas on a car’s mileage.**

**Question 2**

DC and Marvel are two competing comic book properties from which many movies and tv shows have been made. Log into ELMS and from the Files/Datasets folder, download the comicmovie.csv data file. Import it into R/R Studio--note this is a csv file and requires the appropriate function.

Q2a) Create a PQ table reporting the summary statistics for Worldwide Box Office for DC movies, Marvel movies, and both combined. The table should include the following descriptive statistics for each group: n, min, max, mean, standard deviation, and first and third quartiles. Make sure your table title accurately captures the table contents and indicate the units of measurement somewhere, perhaps directly under the table or in the title. Do not carry decimals beyond two places. *Note: Missing values should not be part of your ‘valid’ n.* (0.25 points)

Table

Description automatically generated

2b) Perform a t-test in R to determine if there is a statistically significant difference between the means of worldwide box office sales for Marvel and DC. Report your findings using alpha = 0.10 and using APA citation. If success is measured in box office millions of dollars, can we conclude that one comic company is more successful making movies than the other? (0.50 points)

**t(48.14) = -0.91, p > 0.10**

**Based on these results, we cannot conclude that one comic company is more successful at making movies than the other. The p-value, which is 0.3663, is greater than the alpha of 0.10. This means that the results are not statistically significant and there’s no difference between the parameters of the two populations. We fail to reject the null hypothesis.**

**-- R --**

**# independent two sample t-test**

**t.test(dc$Worldwide, marvel$Worldwide, na.rm=T, pooled=T, paired=F)**

**# results: t = -0.91202, df = 48.142, p-value = 0.3663**

Q2c) What is the standardized effect size of the difference in means from Q2b? Perform a Cohen’s d analysis. Report the difference in means and the Cohen’s d score. Interpret the size of the effect. (0.25 points)

**Cohen’s d: -0.227309 (small effect size)**

**Difference in means: 90.03**

**-- R –**

**install.packages("effsize")**

**library(effsize)**

**# cohen's d**

**cohen.d(dc$Worldwide, marvel$Worldwide, na.rm=T, pooled=T, paired=F)**

**# difference in means**

**round(616.8957-526.8640, 2) # = 90.03**

**Question 3**

Table

Description automatically generated with low confidenceQ3a) Disney set out to purchase Marvel Entertainment in 2009. Using the same comic movie data, create a new dichotomous variable from the Studio variable. To represent the time before and after the Disney purchase, recode or subset the Studio variable in R so that all Marvel movies 2009 and earlier are in one category and all Marvel movies from 2010 to present day are in a second category. The Review variable measures critic and audience approval from Rotten Tomatoes. Report the summary statistics for Review, similar to Q1a, in a PQ table for Pre-Disney, and current day Disney-Marvel companies, as well as totals. (0.50 points)

Q3b) Perform a t-test in R to determine if there is a statistically significant difference between the means of review scores for Marvel movies made before and after the Disney purchase. If success is measured by review scores, can we conclude that Disney has more success making Marvel movies than before the company buyout? Report your findings using alpha = 0.10 and using APA citation. *Hint: treat this data as independent, not paired; we would need before & after sales for each movie to be paired data.* (0.50 points).

**t(38.03) = -3.46, p < 0.10**

**Yes, these results suggest that after purchasing Marvel, Disney has had more success with raising review scores. The results specifically show that the p-value, which was 0.001365, are less than the alpha of 0.10, meaning the results are statistically significant and favor the alternative hypothesis. Furthermore, we can reject the null, which stated there would be no significant difference between the means of review scores.**

**-- R –**

**# independent two sample t-test**

**t.test(past\_2009$Review, present\_2010$Review, na.rm=T, pooled=T, paired=F)**

**# results: t = -3.4556, df = 38.034, p-value = 0.001365**

**Question 4**

4a) What is power of a test? (0.25 points)

**A power refers to the probability of correctly rejecting the null hypothesis when it’s false. The ideal value is a power of 0.8 or 80%.**

4b) Given a sample size of 126, effect size of 0.5 and significance level to be 0.05 calculate the power of a two-sample two-sided t-test in R and report your power value. (0.25 points)

*Hint: #1—you’ll need the “pwr” package in R.*

**install.packages(“pwr”)**

**library(pwr)**

*Hint: #2—the sample size (n) in the R function is for the number in each group, not total n.*

**Power value: 0.9768897 or 0.98 (rounded two decimal places)**

**\*\* This power value is better than our stated ideal of 0.8.**

**-- R –**

**install.packages("pwr")**

**library(pwr)**

**pwr.t.test(power=NULL, n=126, sig.level=0.05, type="two.sample", alt="two.sided", d=0.5)**

**# power = 0.9768897, 0.98 rounded**

**round(0.9768897, 2)**

**Question 5**

Q5a) Describe at least two major differences between an independent samples t-test and a paired sample t-test. (0.25 points)

**Two major differences between these two types of t-tests include:**

**Independent**

* **Independent samples t-tests tests for a significant difference in means or proportions between two different sample groups.**
* **These sample groups are independent of each other.**

**Paired**

* **A paired sample t-test, on the other hand, tests for a significant difference in means or proportions of one sample group that is measured twice. In other words, you compare one sample on two variables.**
* **Samples used in paired sample t-tests are dependent on each other.**

Q5b) What is a pooled estimate? Does the pooled estimate apply to independent sample tests, paired sample test, both, or none? (0.25 points)

**A pooled estimate is a way to estimate certain facts about a population such as, mean, standard deviation, or variance. Since independent sample t-tests should have relatively equal variance, pooled estimates apply largely to this type of test.**

Q5c) Is the following statement True or False? If false, rewrite the statement so it becomes true:Variance does not matter for t-tests; R uses the Student’s t for its default t-test, but I can use the var.equal=FALSE to specify a Kaiser Permanente test. (0.25 points)

**The statement above is false.**

**Variance does matter for t-tests. Prior to conducting a t-test, we assume that population variances are equal or fairly equal. R, however, assumes unequal variance by default and conducts the Welch’s two-sample t-test as a result. If you know the variances are equal, you can use var.equal=TRUE to tell R that they’re equal, thus allowing you to use the traditional two-sample t-test, Student’s t. A Kaiser Permanente test is not an actual statistical test, it’s the name of a healthcare company.**

Q5d) A Beta (or Type II) error occurs when you fail to reject a null hypothesis that is actually false. What are at least two ways to reduce the chances of experiences a Beta error? How does each of those methods reduce the chance of experiencing a Beta error? (0.25 points)

**One way to reduce the likelihood of experiencing a Beta error is increasing the sample size. By doing so, it reduces the variability and chances of failing to be in the rejection region. This also increases the power of a test. Another way to minimize the chances of a Beta error occurring is to also increase the significance level. A higher significance level means there’s a higher probability of rejecting the null when it’s true. While this method reduces the chances of a Beta error, it increases the chances of committing an alpha or Type I error.**

Q6) The number of data science classes taken and data scientist job salary in the USA in dollars has a correlation of 0.91, if we assume all other job market economy characteristics are the same, what would be the correlation be if that job salary was measured in Euros? Explain your answer. (0.25 points)

**If the correlation between the number of data science classes taken and data scientist job salary in the U.S. is 0.91, we can say that it has a high positive correlation. Since $1 USD converts to approximately 0.89 Euros, the correlation between these variables if job salary was measured in Euros would most likely lean toward a lower positive correlation.**