**This take-home final is worth 20 points and is due at noon on Thursday May 6th.**

**Insert text, tables and figures from R in the spaces provided, between the yellow highlight.**

**NOTE: Do not insert screen shots – copy and paste images and text from R.**

**Datasets:**

**You will use 2datasets for this exam, provided with this email and posted in the Bb portal, under the “Take Home” tab**

**You can also use your class notes and other materials (pdfs, panopto videos, textbook)**

**Turning in Exam: Due Date / Time: May 6th at noon**

**When you are done, save the file and rename by adding your name to the title: “TakeHomeFinal\_Sp21\_YOUR NAME”**

**Upload the file in the class Bb site, under the “Take Home” Tab**

**NOTE: Make sure you add your name to the file name, and use the correct email title.**

**Assignments that are incorrectly named will be fined 1 point.**

**Late Work:**

**Late submissions will be penalized 2 points per day (or fraction) starting on May 6th, at noon.**

**DISCLAIMER:**

**This exam recreates homework questions using tests you have seen and have practiced in class. There are no tricks – the computer will run these tests and you will get results.**

**Complete this assignment working ALONE, but use your class notes and reference materials provided by the instructor.**

**1) Long-tailed Monkeys: (3.50 points)**

You have data on the body lengths and the tail lengths of 25 endangered monkeys (use Dataset2) and want to determine whether these two variables vary together (e.g., larger monkeys have longer tails and smaller monkeys have shorter tails).

A) Use the Statistical Decision Tree (in lecture 29) to decide what test you will use to analyze these data:

- Are there any categorical predictor variables? If the answer is YES, what are they? (+0.25):

- Are there any continuous predictor variables? If the answer is YES, what are they? (+0.25):

- Are the data paired or unpaired ?– Explain your reasoning (+0.125):

- What test will you use? – Explain your reasoning (+0.125):

B) What is the null / alternate hypothesis for this statistical test? (+0.125):

C) Paste a scatter-plot of these data (+0.25). Describe the pattern you observe? (+0.125):

D) Perform two tests (one assuming normal distributions and one not assuming normal distributions) of these data and compare your results: Hint – use the nonparametric version that is identical to the parametric version, but uses ranks rather than the raw data.

* Parametric test name : \_\_\_\_\_\_\_\_\_\_ (+0.125) - Degrees of freedom (+0.125):\_\_\_\_
* Non-parametric test name : \_\_\_\_\_\_\_\_\_\_(+0.125) - Degrees of freedom (+0.125):\_\_\_\_

E) Copy and paste the results of parametric test (+0.125):

F) Copy and paste the results of non-parametric test (+0.125):

G) What are the results of these tests? Did the results agree? (+0.125)

H) Based on the similarity / lack of similarity of these two test results, would you guess that these distributions are normally distributed (Why / why not)? (+0.125)

I) You want to determine if the lengths of the tails and the bodies of these monkeys are the same, but are unsure about what statistical test to use. Which type of test should you use to account for the fact that larger monkeys have longer tails (+0.125) – Hint: are these data independent?

J) Perform two tests (one assuming normal distributions and one not assuming normal distributions) of these data and compare your results:

* Parametric test name : \_\_\_\_\_\_\_\_\_\_(+0.125) - Degrees of freedom (+0.125):\_\_\_\_
* Non-parametric test name : \_\_\_\_\_\_\_\_\_\_(+0.125) - Degrees of freedom (+0.125):\_\_\_\_

K) Copy and paste the results of parametric test (+0.125):

L) Copy and paste the results of non-parametric test (+0.125):

M) What was the outcome of these tests? Did the results agree? (+0.125):

**2) Monkey Model (3.5 points):**

Customs agents obtained two monkey tails from a wildlife trader and want to use your data to figure out the body length of the animals these tails were taken from. The tails are 11.0 cm and 12.0 cm long.

A) Use the Statistical Decision Tree (in lecture 29) to decide what test you will use to analyze these data:

- Are there any categorical predictor variables? If the answer is YES, what are they? (+0.25):

- Are there any continuous predictor variables? If the answer is YES, what are they? (+0.25):

- Are the data paired or unpaired ?– Explain your reasoning (+0.125):

- What test will you use? – Explain your reasoning (+0.125):

- Why would you not use a correlation? – Explain your reasoning (+0.125):

B) Explain how this method works (+0.125) and what are the assumptions (+0.125):

C) What are the null / alternate hypotheses of this test? (+0.125):

D) Perform this test and past the results (+0.125):

E) What are the degrees of freedom of this test? (+0.125):

F) What is the result of this test (significant / not significant) (+0.125) – Explain)?

G) Extract the residuals from this test and determine if they are, in fact, normally distributed.

Paste a scatterplot of the residuals (+0.125):

Paste the results of the SW test of normality of the residuals (+0.125):

Based on this result, are you able to run this regression with these data? (Hint: is the key assumption met) Explain why / why not (+0.125):

H) Create a plot showing the linear regression line and the data (+0.25):

I) How much variance in tail length does the body length model account for (from 0% to 100%)? (+0.125):

J) Report and interpret the estimates of the intercept (+/- 2 SE) and slope (+/- 2 SE) from the linear model. Hint: use the rule of thumb (mean +/- 2 SE) to approximate their 95% Confidence limits) and determine if they are, in fact, significant (+0.125 for each cell):

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Parameter (mean +/- 2 SE)** | **Significant?**  **(using rule of thumb)** |
| Intercept |  |  |
| Slope |  |  |

K) Using these estimates, calculate the body length of a monkey that had a 11.0 cm long tail (+0.125). Please show your work (calculations) for full credit.

L) Does this value make sense? Explain why / why not (+0.125).

M) Using these estimates, calculate the body length of a monkey that had a 1.0 m long tail (+0.125). Please show your work (calculations) for full credit.

N) Does this value make sense? Explain why / why not (+0.125).

O) If you had monkey body lengths, what equation would you use to estimate their tail lengths? Use these same data to generate the equation (+0.125):

P) Paste a plot of this new linear regression line showing the point data (+0.25):

**3) Pig Weights (4.5 Points):**

You want to determine the effect of four different diets (feed types) on pig weights, but are unsure if you have normally-distributed data, so you will perform two tests: one parametric and one non-parametric using the following data: 20 pigs: 5 pigs assigned to four different diet treatments (see below). Enter data into Excel and import the file into R **(+0.5 POINTS FOR SETTING UP DATA)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feed 1** | **Feed2** | **Feed3** | **Feed4** |
| 60.8 | 68.7 | 102.6 | 87.9 |
| 57 | 67.7 | 102.1 | 84.2 |
| 65 | 74.0 | 100.2 | 83.1 |
| 58.6 | 66.3 | 96.5 | 85.7 |
| 61.7 | 69.8 | 97.2 | 90.3 |

A) Use descriptive statistics to summarize these four samples. Report the means / medians / SDs / ranges of the data below (+0.125 for each samples \* 4 samples). What is the grand mean? (+0.125):

B) Paste a graph illustrating the means and S.D.s of the four groups in the dataset (+0.250).

C) Based on this graph, do these four groups look different? (Why / Why not) (+0.125).

D) Before you do the parametric test, determine if the variances of the samples are equal:

Are the variances of the groups equal? Why / why not? – Paste and explain the test results (+0.125)

E) Use the Statistical Decision Tree (in lecture 29) to decide what test you will use to analyze these data:

- Are there any categorical predictor variables? If the answer is YES, what are they? (+0.25):

- Are there any continuous predictor variables? If the answer is YES, what are they? (+0.25):

Perform a test assuming normal distributions of these data

* Test name (+0.125) : \_\_\_\_\_\_\_\_\_\_
* Degrees of freedom (+0.125): \_\_\_\_\_\_
* State the null (+0.125) hypothesis: \_\_\_\_\_\_
* State the alternate (+0.125): \_\_\_\_\_\_

F) Copy and paste the tabular results of the parametric test you performed (+0.125):

G) Is this test significant? Explain why / why not (+0.125):

H) Next, do the nonparametric test:

Perform a test assuming these data do not follow a normal distribution:

* Test name (+0.125) : \_\_\_\_\_\_\_\_\_\_
* Degrees of freedom (+0.125): \_\_\_\_\_\_
* State the null (+0.125) hypothesis: \_\_\_\_\_\_
* State the alternate (+0.125): \_\_\_\_\_\_

I) Paste a graph illustrating the 25%, 50%, and 75% of the four groups in the dataset (+0.250):

J) Copy and paste the tabular results of the parametric test you performed (+0.125):

K) Is this test significant? Explain why / why not (+0.125):

L) After getting significance in the ANOVA test, are we done with the analysis? Why / Why not? (+0.125):

M) For the parametric test, report all the pair-wise comparisons of these four groups. Paste the tabular results and explain which groups are different from each other (+0.250):

N) For the nonparametric test, report all the pair-wise comparisons of these four groups. Paste the tabular results and explain which groups are different from each other (+0.250):

**4) Drug Treatments (5.5 Points)**

You want to test the effects of two factors (sex: male / female) and medical treatment (placebo /drugs) on hormone levels. Your sample size is 20 values from 20 individuals. Enter data into Excel and import the file into R. Perform the required parametric ANOVA test. **(+0.5 POINTS FOR SETTING UP DATA)**

|  |  |  |  |
| --- | --- | --- | --- |
| **No Drug** | **No Drug** | **Drug** | **Drug** |
| **Female** | **male** | **Female** | **Male** |
| 16.5 | 14.5 | 39.1 | 32 |
| 18.4 | 11 | 26.2 | 23.8 |
| 12.7 | 10.8 | 21.3 | 28.8 |
| 14 | 14.3 | 35.8 | 25 |
| 12.8 | 10.0 | 40.2 | 29.3 |

A) Use the Statistical Decision Tree (in lecture 29) to decide what test you will use to analyze these data:

- Are there any categorical predictor variables? If the answer is YES, what are they? (+0.25):

- Are there any continuous predictor variables? If the answer is YES, what are they? (+0.25):

- What type of ANOVA would you perform? (highlight one with bold text):

dependent OR independent (+0.125)

one-way OR two-way (+0.125)

B) Before you start, answer these questions:

* Total degrees of freedom (+0.125): \_\_\_\_\_\_\_\_ (Explain Why?) (+0.125):
* Model degrees of freedom (+0.125): \_\_\_\_\_\_\_\_ (Explain Why?) (+0.125):
* Interaction (AXB) degrees of freedom (+0.125): \_\_\_\_\_\_\_\_ (Explain Why?) (+0.125):
* Error degrees of freedom (+0.125): \_\_\_\_\_\_\_\_ (Explain Why?) (+0.125):

C) Use the descriptive statistics to summarize these data: Report the means / medians / SDs / ranges of each dataset data below (+0.125 for each one \* 4 datasets):

D) Paste a bar graph (with error bars) illustrating the means of these four groups (+0.250).

E) What is the grand mean (+0.125)?:

F) What is the “male” mean (+0.125)?:

G) What is the “female” mean (+0.125)?:

H) What is the “no drug” mean (+0.125)?

I) What is the “drug” mean (+0.125)?

J) After looking at these data summaries, which factors do you expect to be significant in this analysis. Explain (+0.125):

K) Perform a parametric test of these data. Copy and paste tabular results of parametric test (+0.125):

L) What was the effect of “sex”: Explain the outcome using the results of the table above (+0.125). State the null / alternate hypothesis of this test (+0.125), and explain the result of the test (+0.125):

M) What was the effect of “drug”: Explain the outcome using the results of the table above (+0.125). State the null / alternate hypothesis of this test (+0.125), and explain the result of the test (+0.125):

N) What was the effect of “sex \* drug” interaction: Explain the outcome using the results of the table above (+0.125). State the null / alternate hypothesis of this test (+0.125), and explain the result of this test (+0.125):

O) Paste a figure illustrating the interaction of “sex” and “drug” (+0.250). Please interpret the figure (does it suggest there was a significant interaction – Explain: Why / Why not (+0.250).

**5) Testing for data normality (3.0 points)**

You have the following distribution (use Dataset1) of 100 fish lengths, which you want to test for normality using two criteria: (i) the kurtosis and skewness, and (ii) the Shapiro-Wilk (SW) test.

NOTE: For the skewness / kurtosis, use the criteria discussed in class (mean +/- 2 SE) to assess whether they are significant. For the Shapiro-Wilk test, use the p-value relative to alpha (0.05).

A) Start by summarizing the “length” data, using some summary statistics. Report the summary statistics below. NOTE: you will be graded on the statistics you select to report (+0.25):

B) Paste the boxplot. Do these data look normal – Explain why or why not (+0.125)?

C) Paste the Q-Q plot. Do these data look normal – Explain why or why not (+0.125)?

D) Report and interpret the kurtosis and the skew of this distribution (+0.2 for each cell):

Explicitly mention if they are significant or not. Fill out the information in the spaces provided:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Parameter2.SE** | **Parameter 95% C.I.** | **Significant**  **(Y / N)** | **Are data Normal (Y / N)** |
| skewness |  |  |  |  |
| kurtosis |  |  |  |  |

E) Based on these results, are these data normally distributed (+0.125)?

For full credit, explain what evidence you base your answer on (+0.25):

F) Next, perform the SW test. What are the null and alternate hypotheses tested by SW test?

Null Hypothesis (+0.1):

Alternate Hypothesis (+0.1)

G) Paste the SW test result (+0.125):

H) Describe the outcome of the test: Explain if the p value is significant and state whether this distribution is normal (+0.2):

**6. EXTRA CREDIT**

**Perform a data transformation of the original length data (in Dataset 1) using Rcmdr (2.0 points)**

A) What data transformation would you use to make these data normal?

Explicitly state why you selected this transformation – given the dataset characteristics? (+0.25).

Explain if you added any constants to the data before the transformation? (+0.25).

B) After the data transformation, report and interpret the kurtosis (+0.25) and the skew (+0.25) of the transformed data – like you did in question 1. Are they significant or not?

C) After the transformation, paste the SW test result (+0.25) and describe the outcome of the test (explain the meaning of the p value and state whether this distribution is normal?) (+0.25). Explicitly state if the data transformation worked?

D) What other transformation could you try to make these data normally distributed? (+0.25).

E) Explicitly state why you selected this transformation – given the dataset characteristics? (+0.25).