STAT 503 – Statistical Methods for Biology

Final Exam

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Please do not cheat.** Do not communicate with anyone other than myself or ITaP Help about the exam or the exam’s content until after 9:00 AM EDT on the due date.
2. For questions and clarifications, please email me with the words “STAT 503 exam” in the title.
3. If you would prefer to ask questions in person, I will be available via telephone (765-494-0030) and WebEx (<https://purdue.webex.com/meet/nlichti>) from 9:00 AM – 12:00 and from 3:00 PM – 5:00 PM on the exam date, and from 8:00 AM – 9:00 AM the next morning.
4. For technical support, contact the ITaP Customer Service Center at 765-494-4000 or itap@purdue.edu, and let me know as soon as you become aware of the technical issue.
5. If you experience problems with submission, please email your exam to me before the deadline.
6. The exam will be handled on Brightspace, similarly to a homework assignment.
7. Grades will not be curved.
8. This is an entirely open book, open note exam. Although you can use the internet (within the bounds of rule 1), I strongly advise you not to do so, since internet resources may not be correct or trustworthy.
9. You can use R, but you do not need access to R for the exam.
10. You may begin the exam at any time between 9:00 AM and 8:00 AM the next day. I strongly recommend that you plan to give yourself plenty of time.
11. The exam is designed to take < 2 hours.
12. Your answers should be typed in the fields provided.
13. All necessary statistical tables will be provided. You do not need to bring them.
14. Unless otherwise specified, round final answers to 3 decimal places and use α=0.05.
15. The answers to some questions may be used as input for subsequent questions. If you do not know how to find the first answer, make up a value to use in the follow-up questions, and clearly state the value that you are using. Your answer to the follow-up will be graded on the basis of your assumed value.
16. Complete sentences are NOT required. Please write only as much as required to completely answer the question that was asked.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question** | **Page** | **Points available** | **Subject** | **Points earned** |
| 1 | 3 | 9 | methods and assumptions |  |
| 2 | 4 | 9 |  |  |
| 3 | 5 | 9 |  |  |
| 4 | 6 | 9 |  |  |
| -- | 7 | -- | Instructions for Questions 7-13 |  |
| 7-8 | 8 | 11 | Checking assumptions |  |
| 9-11 | 9 | 19 | Interpretation and reporting |  |
| 12-13 | 10 | 9 | Study design and power |  |
| 14-15 | 11 | 10 | P-values, power, and sample size |  |
| **Total** | | **85/85** |  |  |

**INSTRUCTIONS FOR QUESTIONS 1-4:** Each question presents a biological research scenario. For each question, identify (**a**) the response variable and its **specific** type, (**b**) the explanatory variable and its specific type (if there is no explanatory variable, write "None"), (**c**) the type of inference (estimation or hypothesis testing) that most closely matches the goal of the analysis, and the target parameter(s) for that inference. Then (**d**) recommend the best option for an analytic method to make the inference, (**e**) list the conditions that must be met for the method to give valid results, and (**f**) identify a reasonable backup option that can be tried in case the assumptions cannot be met (more than one reasonable option may exist; you only need to identify one). **Define any symbols that you use.**

1. A researcher believes that a specific genetic mutation may be associated with a rare disease. To check, she samples 30 individuals who have the disease and 30 who do not, genotypes each of the 60 subjects to see if they have the mutation, and then tabulates the results.
   1. [1 point] Identify the response variable and its type.
      1. The response variable is the phenotype (the presence of lack of the disease)
      2. It is a Categorical/Nominal – either yes or no
   2. [1 point] Identify the explanatory variable and its type (if none, write "none").
      1. The explanatory variable is the genotype because you are trying to explain the phenotype with the presence of a genotype
      2. It is Categorical / Nominal
   3. [1 point] Identify the type of inference and the target parameter(s).
      1. Target Parameter is the statistical difference of the population of people with the disease.
      2. The type of inference is Hypothesis Test
   4. [1 point] Identify the best (first) choice of method for this analysis.
      1. Two-sample t-test
   5. [4 points] What conditions must be met for inferences that use this method to be valid?
      1. That the data was randomly sampled from the population of interest / no bias
      2. Data is distributed normally when plotted
      3. The Sample Size is reasonably large
      4. The std. deviations of the distributions of the samples are comparable
      5. The study actually makes sense
   6. [1 point] Recommend a backup method in case the method in part (d) cannot be used.
      1. Bootstrapping
2. **[see instructions on page 3]** Rice plants can absorb arsenic from the soil and incorporate it into their tissues, presenting a potential public health threat. To measure the relationship between soil arsenic concentration and arsenic concentrations in finished rice grains, 130 individual rice plants were randomly assigned to one of 26 soil arsenic concentrations ranging from 0 to 25 mg/kg (5 plants per concentration). After the plants had matured, arsenic concentrations were measured in 100 seeds from each plant and then averaged, yielding a single average tissue arsenic concentration per plant.
   1. [1 point] Identify the response variable and its type (be specific!).
      1. The response variable is the concentration of arsenic in each plant’s seeds (the grain that humans eat)
      2. The type of variable is numeric and continuous
   2. [1 point] Identify the explanatory variable and its type (if none, write "none").
      1. The explanatory variable is the concentration of arsenic in the soil samples
      2. The type of variable is numeric / discrete. With 26 samples ranging from 0-25 mg/kg you will have {0,1,2,3,4,5…. 25} mg/kg
   3. [1 point] Identify the type of inference and the target parameter(s) for this analysis.
   4. Estimation
   5. intercept and slope of the regression line
   6. [1 point] Identify the best (first) choice of method for this analysis.
      1. Simple Linear Regression
   7. [4 points] What conditions must be met for inferences that use this method to be valid?
      1. Independence of Observation
      2. Random Sampling
      3. No Outliers
      4. There has to be a linear relationship: Could be a straight or logarithmic or quadratic relationship
   8. [1 point] Recommend a backup method in case the method in part (d) cannot be used.
      1. Bootstrap (It’s always bootstrap)
3. **[see instructions on page 3]** A researcher wishes to determine whether an experimental treatment reduces tremor in a mouse model for Parkinson’s disease (“mouse model” means that the mice in question have been bred to show symptoms similar to the disease). Tremor is quantified using a specialized instrument that measures the frequency of shaking in each mouse, in Hz. A total of 12 mice are used, and each mouse is measured twice: once immediately before treatment, and once 30 minutes after the treatment is administered.
   1. [1 point] Identify the response variable and its type.
      1. The change in frequency of the tremor in Hz
      2. The type is numerical/continuous
   2. [1 point] Identify the explanatory variable and its type (if none, write "none").
      1. Boolean Value of Treatment or not treated
      2. Categorical / Nominal
   3. [1 point] Identify the type of inference and the target parameter(s) for this analysis.
      1. Hypothesis Testing, The change in frequency of the tremors in Hz
   4. [1 point] Identify the best (first) choice of method for this analysis.
      1. Paired two-sample t-test
   5. [4 points] What conditions must be met for inferences that use this method to be valid?
      1. Random Sampling
      2. Normally Distributed Errors
      3. No Outliers
      4. Independence of Observation
   6. [1 point] Recommend a backup method in case the method in part (d) cannot be used.
      1. Bootstrapping
4. **[see instructions on page 3]** Noting that her data are quite skewed, a horticulture researcher wishes to determine whether the median number of flowers produced by particular variety of coneflower (*Echinacea spp.*) is greater than 5. She has data on 30 plants.
   1. [1 point] Identify the response variable and its type.
      1. Discrete Numeric Variable
      2. The number of flowers produced by each plant
   2. [1 point] Identify the explanatory variable and its type (if none, write "none").
      1. None
   3. [1 point] Identify the type of inference and the target parameter(s) for this analysis.
      1. Estimation, Median Number of Flowers Produced by plant
   4. [1 point] Identify the best (first) choice of method for this analysis.
      1. Maximum likely-hood estimation
   5. [4 points] What conditions must be met for inferences that use this method to be valid?
      1. Random Sampling
      2. Observations are independent
   6. [1 point] Recommend a backup method in case the method in part (d) cannot be used.
      1. Bootstrapping

**PLEASE USE THE INFORMATION BELOW TO ANSWER QUESTIONS 7-13.**

**IMPORTANT:** On Brightspace, there are results from four possible analyses for the same data set. Question 7-11 ask you to select the analysis that you feel is most appropriate, to defend your selection, and to report and interpret the results. The data come from a real paper, but none of the analyses match the methods use in that paper. Therefore, reading the paper will **not** help you with the exam (if you’re curious about the real study, then I encourage you to read it after you complete the exam to avoid any possible confusion).

**Background Scenario:**

Running is a very inefficient form of transportation when compared with swimming, walking, or even flying, and running efficiency in humans is even lower than it is in many other animals. Less than 10% of the metabolic energy that we burn when we run is used to move our bodies forward. The remaining energy is used to maintain our posture and to swing our back foot into position for our next stride. In contrast, species that are more evolutionarily adapted for running, such as dogs and horses, possess spring-like, elastic tendons in their hips and spines. By alternately stretching and contracting, these tendons reduce the amount of work that must be done by the animals’ muscles and increase their running efficiency as a result.

Recently, Simpson *et al.* (2019, *J. Exp. Biol.* 222: jeb202895) devised a study to determine whether a device modeled on these tendons – basically, a rubber band tied between a runner’s legs – could be used to increase running efficiency in humans (you can watch a short video of a runner using the device by following the VIDEO link attached to this exam on Brightspace). They recruited 12 participants (5 women and 7 men) to take part in the two-day study. At the beginning of the first day, each participant was fitted with a device, and the researchers collected baseline metabolic data. Then, over the next two days, the runners each completed a total of 4 trials. In each trial, the runner completed two 10-minute runs: one run with the device and one without the device. The order of the runs was randomized for each runner and each trial, and all runs were conducted on a treadmill at a constant pace.

The data analyzed here describe the percentage change in energy use when running with the device relative to without the device, within a trial. That is, we are analyzing , where is energy used by runner in trial with the device, and is the energy used without the device. We would like to know the answer to two questions: (1) does running with the device require less energy than running without it? And (2) does the device require habituation (do runners need time to get used to the device)? If the device works and habituation is required, then the effect sizes in later trials will be more negative than in early trials.

**Along with the exam on Brightspace, I have provided R code and output for four different analyses of these data.** Please select the **ONE** analysis that you feel is most appropriate and use it to answer questions 7 and 8. Your answers in questions 9-11 will be scored based on the analysis that you chose to use in Question 7. You can examine all four sets of results before selecting one to use, but you do not necessarily need to look at them all. It is possible that none of the analyses presented here may be ideal; in that case, you should pick the one that is the least bad option.

CONTINUED ON THE NEXT PAGE

**ANALYSIS 1** presents a standard one-way ANOVA for the effect of trial on mean percent change in energy use.

**ANALYSIS 2** presents a nonparametric Kruskal-Wallis test for the effect of trial on mean percent change in energy use, combined with bootstrap confidence intervals.

**ANALYSIS 3** presents a nonparametric permutation test for the effect of trial on mean percent change in energy use, combined with bootstrap confidence intervals.

**ANALYSIS 4** presents a blocked one-way ANOVA for the effect of trial on mean percent change in energy use. We have not explicitly covered this type of analysis, but it is conceptually identical to a paired t-test. Notes in the results will explain unfamiliar outputs.

1. [1] Identify the analysis that you have decided to use (give its number).
   1. Analysis #4
2. [10] Briefly justify your choice of analysis. Your answer should fully explain why you chose the analysis that you have selected. It should (i) address any aspects of the study design that influenced your choice, and should also (ii) list all of the assumptions that you checked, (iii) explain how you checked each assumption, and (iv) explain how the results of those checks led to your choice of analysis. Specific parts of the R results are labeled with capital letters in squares. You can use these letters to cite specific results. If there are assumptions that you could not check, please identify them and state that they could not be checked.
   1. Each person plays both sides of the experiment, so I would go with a paired t-test generally, which in this case it is stated that the blocked one-way anova is similar.
   2. Groups are not independent of each other
3. [12] Formally report the results of the analysis, including both the hypothesis testing and estimation results (you do not need to list the steps of the hypothesis testing procedure). If appropriate, your answer should explain the differences in percent change of energy use between trials, but you do not need to provide estimation results for contrasts that are not significant. You may use . Please round values round 3 decimal places.
   1. Fval is high so we can be confident that there is a large variation among the means of the group and that you wouldn’t expect to see it by chance
   2. F = 5.794
4. [4] From a biological point of view, how should we interpret the results of this analysis? Can we conclude that the device increases running efficiency? Does it require habituation? Explain your answer.
   1. The device increases running efficiency. The later trials show a more negative change and therefore you would assume that there is some level of habituation involved (Figure Section D) The figure also shows a lot of change values in the in the negative range as opposed to 0 or positive. This strengthens the argument for the devices.
5. [3] How confident are you in the conclusions that you drew from this analysis? Do you expect that they would be repeatable, meaning that if we redid the study using the same sample size, study design, and alpha, we would reach a similar qualitative conclusion? Please explain your answer.
   1. The P-value is 0.03 which is around 0.05 cutoff so we can be fairly confident.
   2. The sample size is fairly low, so there may be some differences, but we can be confident that the conclusion would be similar
6. [3 points] In a separate analysis using a different, randomly assigned set of subjects, the researchers found that a placebo device, which looked similar to the real device but was not elastic, caused no significant change in energy use across 4 repeated trials.
   1. Given this result, would it be reasonable to conclude that the device causes runners to become more efficient? Why, or why not?
      1. This strengthens the argument that the device helps increase efficiency of runners. This is the control trial that was needed to see if the weight of the device had any detrimental effect.
   2. There is a flaw in the approach that is described in this question. What is it?
      1. The flaw is that it’s a different set of subjects?
7. [6 points] For each scenario below, explain whether the power of this test will go up, go down, or remain unchanged, relative to the test you just performed, and explain **why**:
   1. We change the significance level from 0.05 to 0.01
      1. Decreasing the Power
      2. By lowering the significance level, you are more likely to accept the null hypothesis.
   2. We increase the number of runners from 12 to 20?
      1. Increase the Power
      2. Effects are easier to detect with a larger sample
   3. We use different runners for the control group (without the device) versus the treatment group (with the device) instead of collecting data on each runner both with and without the device (we still run 4 trials)?
      1. Decrease the Power
      2. Because you have no control over the natural variance between the abilities of the two sample groups.
8. [4 points] Briefly interpret each of the following -values using and power = 0.8 at the smallest effect size of interest. Include both your conclusion and your confidence in your conclusion (high, medium, or low).
   1. 072
      1. Pvalue is greater than alpha, but we have a high power, so we can have some low confidence in the conclusion.
      2. This value is garbage, just reject it outright. CONFIDENTLY REJECT
      3. Pvalue < alpha Therefore the results are statistically significant and we can have medium confidence since the p-value is fairly close to 0.05
      4. therefore, the results are statistically significant with HIGH confidence
9. [6 points] One of the major themes that we have seen repeatedly throughout the course has been the importance of sample size both to our selection of methods and to our interpretation of results. Briefly explain how sample size affects our use of statistics. Your answer should consider all stages of an analysis, from selection of methods, to checking assumptions, to interpretation of results.
   1. Selection of Methods: If you don’t have as many units in your sample as you would like, you could bootstrap. Otherwise, if you have the ability to reasonably collect multiple data points (100-1000) then you can just take multiple subsamples from the population.
   2. Checking Assumptions: Increasing the sample size allows you to be more confident that your sample is not biased.
   3. Interpretation of Results: By having a larger sample size, you are increasing the chance that you have a representative piece of the population and that your results are transferrable to the population as a whole.