STAT 503 – Statistical Methods for Biology

Practice Exam 2

83 Points

Name: __Solutions_____

9:00 AM 10:30 AM Online (circle one)

Please:

1. Do not cheat (don't speak to anyone other than me or a proctor during the exam, and don't look at anyone else's paper).
2. Write clearly and neatly. If I cannot read your answer, it will be counted wrong.
3. **Show your work** on all calculations. At a minimum, write the symbolic equations you are applying.
4. **Round final answers to 3 decimal places**.
5. Answer questions in the space provided. If you need additional space, clearly mark where the continuation may be found on both the main exam and at the location of the continuation.
6. If you have a question or need scrap paper, raise your hand and I or a TA will come to you.
7. Complete sentences are **NOT** required. Please write only as much as required to completely answer the question that was asked.
8. Phones, computers, and other electronics that are able to access the internet are **not** allowed. Any use of such a device during the exam may result in a score of zero on the test.
9. **You are allowed to bring one (1) double-sided, 8.5 × 11 sheet of notes**. Notes may be handwritten or typed in any font size, and can contain any material that you think will be helpful.
10. You may use a calculator (you may **NOT** use your phone as a calculator). Graphing calculators are allowed.
11. Please bring your Purdue Student ID to the exam and be prepared to present it when you turn the exam in.
12. If you finish early, you may turn in your exam and leave. Please be as quite as possible to minimize distractions for your classmates.
13. You will have 1.5 hours to complete the exam.

Question	Page	Points available	Points earned
1-3	3	14	
4-5	4	7	
6-9	5	14	
10-11	6	10	
12-15	7	16	
16-18	8	8	
19-20	9	8	
21-22	10	6	
Total		**83**	/ 75 = _____%

Use the following information for Questions 1-11.

The phenomenon of "stalk lodging" occurs when the stem of a grain plant like corn or wheat breaks below the seed head, so that the stalk falls over. Lodging can be caused by a variety of agents, including wind storms, heavy rain, insect pests, and drought, and sometimes results in large crop losses. In subsistence farming communities, severe lodging events may even lead to famine. Crop breeders are working to develop varieties of corn and other grains with stronger stalks that can resist lodging.

To investigate the stalk strength of a new corn variety, 20 plants were subjected to stress tests. Specifically, the torque required to break their stems was recorded in Newton-meters (Nm). The mean torque required to break stems in this sample was 14.8 Nm, and the standard deviation was 3.2 Nm.

The corn varieties that are currently in use break when subjected to an average torque of 13.2 Nm. Please determine whether the new variety is stronger than these varieties, and estimate the mean torque required to break the new variety at a confidence of 95%.

1. [6 points] Identify an appropriate null and alternative hypothesis for this test (if you use symbols, please define them).

   There are two ways to answer this question. For the null hypothesis, you could either state that $\mu = \mu_0$ or that $\mu - \mu_0 = 0$. These are identical statements; in both cases, 13.2 Nm is a fixed value (it is not estimated from data)

   H_0: There is no difference between the mean torque required to break the new corn variety (μ) and the mean torque required to break the current varieties ($\mu_0 = 13.2$ Nm). That is, H_0: $\mu = 13.2$ OR $\mu - 13.2 = 0$.

   H_a: There is a difference between the mean torque required to break the new corn variety (μ) and the mean torque required to break the current varieties ($\mu_0 = 13.2$ Nm). That is, H_0: $\mu \neq \mu_0$ OR $\mu - \mu_0 \neq 0$. This could also be stated as a one-sided hypothesis, H_0: $\mu > 13.2$ OR $\mu - 13.2 > 0$

2. [4 points] Calculate the standard error of the average torque required to break stems in the new variety.

$$SE = \frac{s}{\sqrt{n}} = \frac{3.2}{\sqrt{20}} = 0.7155 \approx 0.716 \text{ Nm}$$

3. [4 points] What would be your first choice as a test statistic for this hypothesis test, assuming that the data meet the necessary assumptions? Briefly explain your choice.

   The appropriate test statistic for these data is the Student's t statistic. We are interested in testing a hypothesis about and estimating the value of the mean of a continuous variable (torque required to break corn stems), and we do not know the value of the standard deviation for the population of breaking-torques in this variety of corn.

   Note that we almost never know the population standard deviation. Therefore, the t-statistic is almost always the first (best) choice for estimating the mean of a continuous variable.

4. [4 points] List the assumptions of the method you identified in Question 3. For the remaining questions, assume that all assumptions have been met.

   The assumptions for using the t-statistic are:
   1) The data represent a random sample from the population.
   2) All observations are independent of each other.
   3) The population is normally distributed (also including sample size requirement of the CLT here is okay).
   4) (Technically) The population standard deviation is unknown and must be estimated with s.

   There are technically no assumptions made regarding sample size. Results based on the t-statistic are robust to non-normality if the sample size is sufficiently large (≥ 10 for symmetric non-normal data, or ≥ 30 for mildly skewed data). That means we can get away with breaking the normality assumption. Not that we aren't making it. If the data <u>are</u> normal, then there is no formal sample size requirement, but answers with $n < 5$ are likely to be unreliable (with $n < 5$, answers using any method are probably unreliable).

   If you listed a different statistic, then its assumptions would be expected here.

5. [3 points] Are there any other potential problems that you need to check for, before proceeding with the test and estimation?

   Other than meeting the assumptions of the test, you should check for outliers before using the t-statistic. It is very sensitive to outliers.

   Various other answers were given at least partial credit, but failing to mention outliers resulted in some deduction of points.

6. [2 points] Calculate the degrees of freedom for the test statistic, if appropriate (if it does not have any, write NONE).

   This was graded on the basis of the proposed test statistic. For the single-sample t,

   $$df = n - 1 = 20 - 1 = 19$$

7. [5 points] Calculate the value of the test statistic.

   $$t_{19} = \frac{\hat{\mu} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{14.8 - 13.2}{3.2/\sqrt{20}} = 2.2362 \approx 2.236$$

   OR, if your null hypothesis was stated in terms of the difference between $\bar{x}$ and μ_0,

   $$t_{19} = \frac{(\bar{x} - 13.2) - \mu_0}{s/\sqrt{n}} = \frac{(14.8 - 13.2) - 0}{3.2/\sqrt{20}} = \frac{(1.6) - 0}{3.2/\sqrt{20}} = 2.2362 \approx 2.236$$

   If you set the calculation up perfectly but used $\hat{\mu} - \mu_0 = 14.8 - 0$ in the numerator, 3 points were deducted. This is a substantial error - always think about the meaning of the numbers that you are using.

8. [2 points] The two-sided P-value for the test statistic is 0.0375. Draw a conclusion regarding the outcome of the hypothesis test at $\alpha = 0.05$ (apply a two-sided test here, even if your alternative hypothesis in Question 1 is one-sided).

   At $\alpha = 0.05$ we would reject the null hypothesis, since $P = 0.0375 < \alpha$. Since we reject the null hypothesis, we conclude that the mean torque required to break the new variety of corn is different from the mean torque required to break the current varieties (specifically, because $\bar{x} > 13.2$, we can say it is greater).

9. [5 points] Use the P-value in Question 8 to calculate the P-value for each of the following one-sided alternative hypotheses:
   a. The new variety is stronger than the existing varieties, on average.

      $$P_{greater} = P(t_{19} > t^*) = 0.5P_{two-sided} = 0.5(0.0375) = 0.01875 \approx 0.019$$

   b. The new variety is weaker than the existing varieties, on average.

      $$P_{lesser} = 1 - P_{greater} = 1 - 0.01875 = 0.98125 \approx 0.981$$

      Note that if $\bar{x}$ had been *less than* μ_0, these would have been reversed.

10. [6 points] Calculate a 95% confidence interval for the mean breaking-torque of the new corn variety.

At $n-1=19$ degrees of freedom, $t^*=2.09$

$$CI_{0.95}=\bar{x}\pm t^*_{n-1}SE=\bar{x}\pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right)=14.8\pm 2.09\left(\frac{3.2}{\sqrt{20}}\right)$$

$$\rightarrow (13.305, 16.295)\ \text{Nm}$$

11. [4 points] How many plants would we need to measure in order to estimate the mean torque with 95% confidence at a precision of $\pm$0.5 Nm? You may reuse the critical value from Question 10 in your calculation.

The margin of error for the confidence interval is the distance from the point estimate to one of the two confidence limits. Since the confidence intervals based on the *t*-statistic are symmetrical, it is easier to use the upper value. Therefore,

$$tolerance=\mathrm{t}^*_{\mathrm{n}-1}\left(\frac{s}{\sqrt{n}}\right)$$

$$0.5\ \text{Nm}=2.09\left(\frac{3.2}{\sqrt{n}}\right)$$

$$\sqrt{n}=2.09\left(\frac{3.2}{0.5}\right)$$

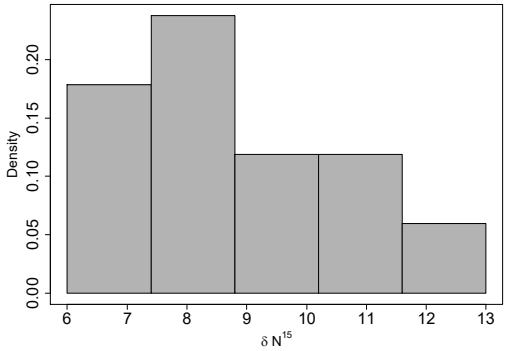
$$n=(2.09\times 6.4)^2$$

$$=178.91$$

We should always round target sample sizes up to be on the safe side, so the answer is 179 (or 180) plants.

Use the following information for Questions 12-17

Individuals who eat diets high in meat and marine fish have higher ratios of N^{15} to N^{14} (referred to as δN^{15}) in their tissues than individuals who mostly eat plants. Your friend is using δN^{15} to study meat consumption patterns, and is interested in estimating the average value in her study population. A histogram of her data is shown below, along with the results from her analysis. Please use these results to answer the following questions.




```
 One Sample t-test

data:  dN15
t = -2.2956, df = 11, p-value = 0.04236
alternative hypothesis: true mean is not
equal to 10
95 percent confidence interval:
 7.545231 9.948372
sample estimates:
mean of x
 8.746802
```

12. [4 points] Give the estimated mean and confidence interval from your friend's analysis, including the confidence level.

    From the R output, the estimated mean is 8.746. The 95% confidence interval is (7.545, 9.948).

13. [2 points] What sample size did your friend use?

    From the R output, there are 11 degrees of freedom for a single sample *t*-test.

    $$df = n - 1 \rightarrow n = df + 1 = 11 + 1 = 12$$

14. [6 points] Given your answer in Question 13, is your friend using an appropriate methodology? Please explain your answer.

    No. The choice of a *t*-test is reasonable because the data describe a ratio of N^{15} concentration to N^{14} concentration, which is a continuous random variable, and she is interested in making an inference about the mean δN^{15} value in the population. However, the histogram of the data shows a clear skew. While the skew is not severe (I would classify this as mild-to-moderate), it is sufficient that the *t*-test would be unreliable with only 12 data points.

15. [4 points] Suppose the sample size was 50. Would this change your answer in Question 14? Why, or why not?

    Yes. With a sample size of 50 (which is ≥ 30), we should be able to rely on the central limit theorem to ensure the approximate normality of the sampling distribution for the mean. As a result, the *t*-test will be robust to a relatively mild skew, such as the one seen in the histogram.

    No points were deducted if you said no but made clear that you understood the application and also explicitly described the skew as "strong," "severe," "heavy," etc.

16. [4 points] Propose an alternative methodology that your friend might want to try for calculating the confidence interval, and explain why you are proposing it.

    Two fallback techniques in this scenario would be the bootstrap and likelihood profile. The bootstrap does not make any distributional assumptions and is relatively easy to implement, so it would be a reasonable choice.

    The likelihood profile method can be used to estimate the mean with any parametric model, so it can be used with this data. However, using it would require you to specify a different parametric distribution for the data (by different, I mean not normal). This distribution would need to describe a strictly positive, continuous variable, and accommodate a skew. Here, options for appropriate distributions might include the Gamma distribution, the log-normal distribution, the Weibull, and the inverse Gaussian distribution (none of which we have discussed in this class).

    A third option would be to transform the data. In this case, a log-transformation might work reasonably well.

17. [2 points] Are there any potential drawbacks to the method you proposed in Question 16? If so, please describe them.

    The downside of the bootstrap is that it produces confidence intervals that are too narrow when sample sizes are small (as they are here). It will therefore underestimate P and overstate the precision of the estimate. However, they will be approximately correct and may be better than the t-based intervals in light of the skew and small sample size.

    The main downside of the likelihood profile method, other than the fact that it is more complex to calculate, is that you need to specify an alternative parametric model, and it may not be immediately obvious what that model should be. Neither the binomial distribution nor the Poisson would be appropriate in this situation.

    The main downside to transforming the data is that the results can be less intuitive to interpret. However, you could back-transform the confidence intervals after calculating them.

18. [2 points] Give an example of a situation in which it would be appropriate to make the significance level, α,
    a. Smaller than the standard level of 0.05 (e.g., 0.01): If α is less than the standard level, then you are setting a higher than usual standard for the level of evidence needed to reject the null. This would be appropriate if your example supposes a severe negative consequence of a false discovery.
    b. Larger than the standard level of 0.05 (e.g., 0.1): A low level for α increases the risk of a false discovery, but also increases power. It is appropriate is false negatives have worse consequences than false positives, and in pilot studies, where power is limited by small sample sizes and positive results will only be considered real if they are confirmed later.

19. [4 points] In your own words, please explain why it is necessary to know the power of a hypothesis test, as well as the significance level and P-value.

Together, the significance level (α) and P-value help to determine whether or not you should reject the null hypothesis. The significance level serves as a marker for the degree of inconsistency between the null hypothesis and the data that you consider to be sufficient to reject the null, and the P-value measures that inconsistency.

Power is the sensitivity of the test. You need to know how sensitive the test would be to a small but a meaningful deviation from the null hypothesis so that you will be able to interpret a negative result. If power is small (or if power is unknown), then failing to reject the null hypothesis cannot lead to any meaningful conclusion. Either the null hypothesis was not rejected because it is actually (or at least approximately) true, or it was not rejected because you lacked the power to be able to tell that it was false, but we have no way to tell which possibility is more likely. This ambiguity in the outcome of the experiment is something we want to avoid.

20. [4 points] What is the difference between statistical and biological significance.

If we say that a result has statistical significance, we are making a statement about the degree of confidence that we have when we reject a null hypothesis. This confidence depends on things like sample size (power) and our subjective opinion regarding what constitutes "strong" evidence against the null (this is encoded in our choice of α). Neither of these has anything to do with the actual magnitude of the effect that we are studying.

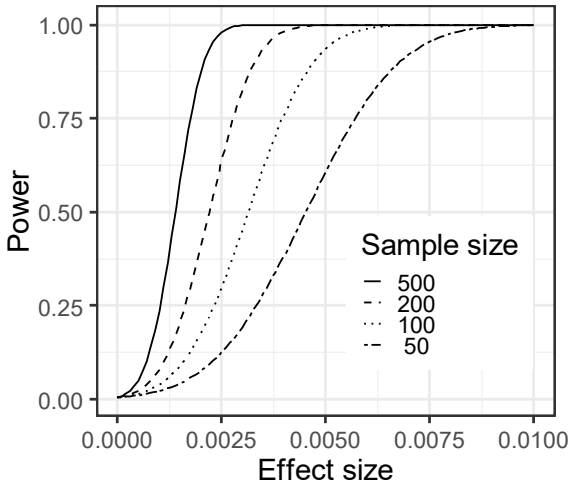
In contrast, if we say that a result is biologically significant, then we are saying that the estimated effect is large enough in magnitude to have a meaningful impact on some biological system or in some clinical setting.

For example, let's say a study finds that eating a certain food will increase the risk of developing cancer from 0.001 per year to 0.00102 per year. The result might be statistically significant, but the change is probably not sufficient to cause many people to alter their eating habits (they would not consider it to be biologically significant). On the other hand, an increase in risk from 0.001/year to 0.1 /year (i.e., from a 1 in 1000 chance to a 1 in 10 chance) very likely would be seen as biologically meaningful.

It's worth noting that "meaningful impact" is a subjective term. From an individual consumer's perspective, the change in personal risk from 0.001 to 0.00102 might not mean much. But for a public health policy maker whose decisions affect several hundred thousand people, it would mean a difference of hundreds of cancer cases, so in that context, it might be quite meaningful.

It is also worth noting that if we consider the result not to be statistically significantly different from zero, then what we are saying is that we are not confident that the true effect is bigger than zero (it is plausible eating this food has no effect on cancer risk at all). The key difference between statistical significance and biological significance is that the first refers to our confidence that an effect is real, while the second refers to the meaning of the effect, given that it is real.

21. [3 points] At work, you have been asked to collect data to check the calibration of a pH meter. The meter will be used to prepare buffers for use in future experiments, and it is very important that the meter be accurate to within $\pm$ 0.005 pH units. If it is not, then irreplaceable biological specimens might be used up without getting unusable data. On the other hand, you have studying to do, and testing the pH meter is *so b-o-o-r-r-r-ing* (imagine this in the voice of a 5 year-old who is being dramatic). So, you take an initial sample of 25 readings and use them to prepare the power curves shown here. Which sample size should you use, and why (explain what is wrong with the other options)?



Of the sample sizes shown here, the best option is $n = 200$. This is sufficient to almost guarantee that biases $\pm$0.005 pH units will be detected if they exist. Smaller samples cannot make this guarantee (the power is below 1 for n =100 or 50), and the larger sample size of 500 creates unnecessary extra work (and boredom) without generating any additional benefit (power is already essentially 1 with 200 data points, so it cannot get better). I am aiming at a power near 1 instead of 0.8 because the consequences of failing to detect a calibration error at ± 0.005 units is severe (loss of irreplaceable specimens).

If your answer made clear that you understood the basic logic of the tradeoff between gaining power and diminishing returns, but you applied a standard of 0.8, then only 0.5 points were deducted.

22. In a study of cellular calcium transport, we test the null hypothesis that mean uptake across the plasma membrane in 10 minutes will be equal to 4 nmol/mg. P =0.1793 and power at a minimum effect size of 0.5 nmol/mg is 0.4501. The 95% confidence interval is (2.4, 4.6) nmol/mg.

a. [2 points] Is it reasonable to conclude that the null hypothesis is supported? Why, or why not?

No, we cannot conclude that the null is supported. The power is too small - we have only a 45% chance of rejecting the null hypothesis, assuming that the true effect size is 0.5 nmol/mg. Therefore, assuming that the null really is true, we have a 55% chance of making a Type II error. Under these circumstances, a non-significant -value is uninterpretable and meaningless. The null might be true, or we might not have enough data to be able to tell that it is false.

b. [1 point] If we redo the study, what is the the single best thing we could do to make sure we get results that are more interpretable?

Use a larger sample size.

[THIS PAGE IS BLANK]