STAT 503 – Statistical Methods for Biology

Final Exam (practice)
Note: This exam may longer than the actual exam

Practice exam (answer key)

Name: ____

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. **You are allowed to bring four (4) double-sided, 8.5 × 11 sheets of notes**. Notes may be handwritten or typed in any font size, and can contain any material that you think will be helpful.
3. **All necessary statistical tables will be provided**. You do **not** need to bring them.
4. Write clearly and neatly. If I cannot read your answer, it will be counted wrong.
5. **Show your work** on all calculations. At a minimum, write the symbolic equations you are applying.
6. **Unless otherwise specified, round final answers to 3 decimal places and use $\boldsymbol{\alpha = 0.05}$**.
7. 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.
8. If you have a question, do not understand a question, or need scrap paper, raise your hand and I or a TA will come to you.
9. **Complete sentences are NOT required.** Please write only as much as required to completely answer the question that was asked.
10. 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.
11. You may use a calculator (you may **NOT** use your phone as a calculator). Graphing calculators are allowed.
12. Please bring your Purdue Student ID to the exam and be prepared to present it when you turn the exam in.
13. If you finish early, you may turn in your exam and leave. Please be as quite as possible to minimize distractions for your classmates.
14. You will have 2 hours to complete the exam.

Question	Page	Points available	Subject	Points earned
1	3	10	methods and assumptions	
2	4	10		
3	5	10		
4	6	10		
5-7	7	15	interpretation (checking assumptions)	
8-10	8	17	interpretation	
11-13	9	9		
14-15	10	10	sampling error, precision and power	
16-17	11	9	power (cont.), reporting results	
Total		**100 / 100**		

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"), and (c) the type of inference (estimation or hypothesis testing) and the target parameter(s). 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 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 medical researcher wants to use data on 200 patients to describe the relationship between age (in years) and systolic blood pressure (in mm Hg).

   a) [1 points] Identify the response variable and its type.

   Blood pressure, continuous numeric

   b) [1 point] Identify the explanatory variable and its type (if none, write "none").

   Age, continuous numeric

   c) [2 point] Identify type of inference and the target parameter(s) for this analysis.

   Estimation, intercept and slope of the regression line (regression coefficients, regression line, or equivalent is okay)

   d) [2 points] Identify the best (first) choice of method for this analysis.

   (simple linear) regression

   e) [3 points] What conditions must be met for inferences that use this method to be valid?

   1. Random sampling
   2. Independence of observations
   3. Normally distributed errors with constant variance
   4. Linear relationship between explanatory and response variables in the population
   5. No outliers

   f) [1 point] Recommend a backup method in case the method in part (d) cannot be used.

   Bootstrap

2. **[see instructions on page 3]** Two spectrometers will be used in an experiment. To check that they are consistent with each other, 24 specimens are prepared, and then each specimen is divided in half. The first half is analyzed with spectrometer A, and the second half is analyzed with spectrometer B. We want to confirm that there are no systematic differences between readings from the two machines.

a) [1 points] Identify the response variable and its type.

Spectrometer reading, continuous numeric

b) [1 point] Identify the explanatory variable and its type (if none, write "none").

Spectrometer ID (A vs. B), nominal categorical

c) [2 point] Identify type of inference and the target parameter(s) for this analysis.

Hypothesis test, population mean reading

d) [2 points] Identify the best (first) choice of method for this analysis.

Paired two-sample *t*-test

e) [3 points] What conditions must be met for inferences that use this method to be valid?

1. Random sampling
2. Independence of each pair of observations
3. Normally distributed errors (or normal distribution of pairwise differences)
4. No outliers

f) [1 point] Recommend a backup method in case the method in part (d) cannot be used.

Permutation, sign test, or Wilcoxin signed-rank;
Bootstrap (half credit)

3. **[see instructions on page 3]** In an effort to describe the prevalence of Lyme disease in the state, the department of public health captures 500 ticks and tests each of them for the disease (prevalence is the percentage of a population that is infected with or has been exposed to a disease).

   a) [1 points] Identify the response variable and its type.

   Lyme disease infection, nominal categorical

   b) [1 point] Identify the explanatory variable and its type (if none, write "none").

   None

   c) [2 point] Identify type of inference and the target parameter(s) for this analysis.

   Estimation, population proportion

   d) [2 points] Identify the best (first) choice of method for this analysis.

   Maximum likelihood estimation (aka MLE, ML, likelihood profile); Agresti-Coull is also acceptable

   e) [3 points] What conditions must be met for inferences that use this method to be valid?

   1. Random sampling
   2. Independence of observations
   3. (All individuals have same probability of being infected)

   f) [1 point] Recommend a backup method in case the method in part (d) cannot be used.

   Bootstrap

4. **[see instructions on page 3]** A botanist hypothesizes that herbicide resistance in giant ragweed (*Ambrosia trifida*) is related to a specific genetic mutation. She collects 50 plants from a population that is known to be susceptible to herbicides, and 50 from an herbicide-resistant population. Then each plant is genotyped and the frequency of the mutation is compared between the two samples.

a) [1 points] Identify the response variable and its type.

Possession of the mutant genotype, nominal categorical

b) [1 point] Identify the explanatory variable and its type (if none, write "none").

Population ID (or herbicide resistance), nominal categorical

c) [2 point] Identify type of inference and the target parameter(s) for this analysis.

Estimation, Relative proportion of population made up of mutants

d) [2 points] Identify the best (first) choice of method for this analysis.

Odds ratio

e) [3 points] What conditions must be met for inferences that use this method to be valid?

1. Random sampling
2. Independence of observations

f) [1 point] Recommend a backup method in case the method in part (d) cannot be used.

Bootstrap might be used, although sample size is small

Use the following information and the R output on Page 13-14 for Question 5-13

A student is interested in the effects of soil moisture on the growth of a particular plant. He plants a total of 20 seedlings in individual pots and randomly assigns each of plant to one of 4 soil moisture regimes: 12%, 16%, 20%, or 24% water by volume. Then he randomly assigns the seedlings to locations in the greenhouse and grows them under similar conditions for 4 weeks. At the end of 4 weeks, he measures the height of each seedling and records its growth rate in mm/week (calculated as total height divided by 4). **R commands and output for the analysis of these data may be found on Page 12-13. Please use them to answer Questions 5-13.**

5. [1 point] What type of analysis is this?
   One-way Analysis of Variance (or ANOVA)

6. [12 points] Check all of the conditions that need to be met for this analysis to give valid results. For each condition, **either** cite specific evidence from the R output **or**, if you cannot test the condition with the given output, state that it is assumed. When citing evidence, you may use the letter codes in boxes on pages 13-14 to reference specific output. **There might be more rows than you need**.

Condition (assumption)	Met? (Yes/No)	Evidence
Random sampling	yes (or unknown)	assumed
Independence	yes (or unknown)	assumed
Normal errors (or normal populations)	No	Right skewed histogram (E) and convex QQ-plot (D), with asymmetric boxplot (C)
Equal variance	Yes	$\frac{s_{max}}{s_{min}} = \frac{2.56}{1.57} = 1.63 < 2$
No outliers	Yes	Boxplot does not contain any independent points

7. [2 points] Based on the checks you performed in Question 6, should we trust the model results? Why or why not?
   No. The error distribution is skewed and $n < 30$, so the CLT may not be reliable.

8. **Refer to the output labeled F on page 14**.

a) [3 points] State the null and alternative hypotheses tested in this output. Define any symbols that you use.

H_0: All treatment groups have the same population mean (or moisture has no significant effect on growth rate, or any other equivalent statement); $\mu_1 = \cdots = \mu_4$, where μ_i is the (true) population mean growth rate of plants grown at the corresponding moisture level.

H_a: At least one treatment mean is different (or moisture does have a significant effect)

b) [5 points] Write a sentence to formally report the result of this test, including test statistic, degrees of freedom, P-value, and conclusion.

Soil moisture has a significant effect on mean growth rate ($F_{3,16} = 29.906\,, P = 8.549 \times 10^{-7}$).

c) [1 point] What is the estimated standard deviation of the errors in this model (i.e., what is $\hat{\sigma}$)?

2.13 mm/wk (the mean square error)

9. [4 points] Report the test statistic, degrees of freedom, P-value, and your conclusion for a test of the null hypothesis that the mean growth rate for plants in the 12% moisture treatment is equal to zero.

The mean growth rate for plants in the 12% moisture treatment is significantly different from 0 mm/wk ($t_{16} = 7.758, P = 8.24 \times 10^{-7}$).

10. We are interested in all pairwise comparisons between treatments.

a) [2 points] Report the estimated difference in growth rate between the 24% and 12% moisture treatments.

The estimated difference (95% CI) between the means of the 24% and 12% moisture treatments is 5.862 (2.008, 9.716) mm/wk.

b) [2 points] Is there a significant difference between the 24% and 16% treatments at $\alpha = 0.05$? Please cite a specific result to support your answer.

No, there is no significant difference between the mean growth rates at 24% and 16% moisture (Tukey-Kramer $P = 0.0617$).

11. [1 point] What proportion of the variation in growth rate within these data is attributable to moisture treatment?

$$R^2 = 84.87\%$$

12. [2 points] What biological conclusions can you draw from this analysis (hint: if I want to grow these plants, what advice would you give me)? For the purposes of this question, assume the model is valid, regardless of your answer in Question 7.

    Soil moisture does have a significant effect on growth rate in this species. Specifically, the highest average growth rates occur at 16-20% soil moisture.

13. [6 points] If the results for this analysis are valid, can we conclude that there is a causal relationship between soil moisture and plant growth in this species? Please explain why or why not: specifically, what step in the study methodology allows you to infer a causal relationship (if you said no, then what step is missing), and why is that step important (what is its logical role)?

    Yes. A causal effect of moisture on growth rate can be inferred from this study because moisture treatments were randomly assigned to sample units. Random assignment minimizes the possibility that the observed pattern in the data could be due to an unobserved confounding variable.

14. [2 points] In your own words, please explain the concept of sampling error.

When we collect data, we can typically only observe a subset of the statistical population that we are interested in (i.e., a sample). Sampling error is the random variation that occurs in sample statistic (i.e., an estimate of a population parameter) from one sample to another, as a consequence of the fact that we are not observing the entire population.

15. In a pilot study of diet effects on fasting blood triglyceride levels, the sample standard deviation from a sample of 5 mice was 0.586 mmol/L.

a) [6 points] Suppose I have space to house 24 mice independently. **Please approximate the necessary sample size** to determine whether it is feasible to estimate the mean triglyceride concentration at a 95% confidence level to a margin of error of $\pm$ 0.2 mmol/L? Show your work.

Using the standard normal distribution,

$$n = \left(z_{0.975}\frac{s_x}{tol}\right)^2 = \left(1.96\frac{0.586}{0.2}\right)^2 = 32.98 \approx 33 \text{ mice.}$$

Alternatively, using 2 standard errors,

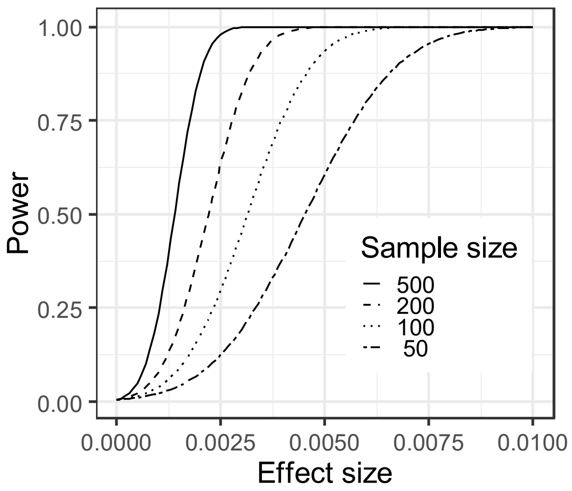
$$n = \left(2\frac{s_x}{tol}\right)^2 = \left(2\frac{0.586}{0.2}\right)^2 = 34.34 \approx 35 \text{ mice.}$$

In either case, the answer is no, it is not feasible.

b) [2 points] Will the approximation in 15(a) underestimate, overestimate, or correctly represent the sample size that would be needed for estimation using the Student's t distribution?

underestimate

16. The graph to the right shows a set of power curves for a two-sample t-test.



   a) [1 point] What is the approximate power at an effect size of 0.0025 units, with $n = 100$?

   0.25 (or 25%)

   b) [2 points] If the smallest effect that we care about detecting is 0.005 units, does it make sense to collect data on 500 individuals? Why, or why not?

   No. It would be a waste of effort. Using 500 sample units gives essentially the same power as 200 sample units. Moreover, we typically aim for a power of 0.8 at the minimum interesting effect size, and this is achieved with only 100 sample units.

17. Consider the following statement of results: "We found that blood pressure was significantly affected by exercise (two-sample independent $t_{12} = -1.912$, $P < 0.05$)."

   a) [2 points] Why is this style of reporting a significant P-value problematic?

   This statement does not tell us how close the P-value is to α. Without that information, we cannot tell how reliable the conclusion to reject the null actually is (or equivalently, how repeatable the outcome is likely to be).

   b) [4 points] What additional information should be reported, and what information should be reported differently, so that the results are summarized more clearly and completely?

   Additional information should include a point estimate and confidence interval. An exact P-value should also be given.

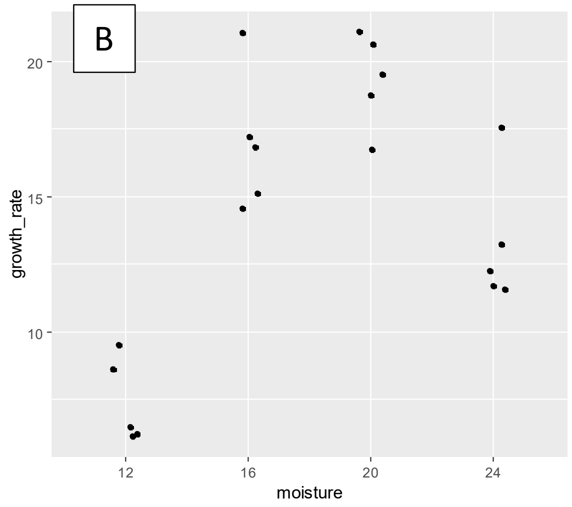
R output for Questions 5 - 13

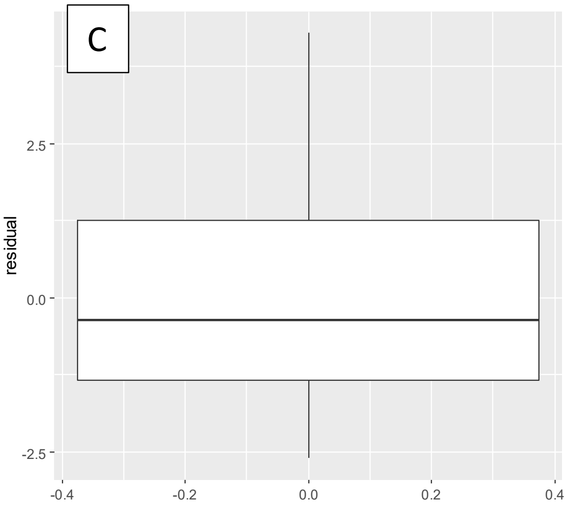
Each piece of output is labeled with a capital letter in a box.

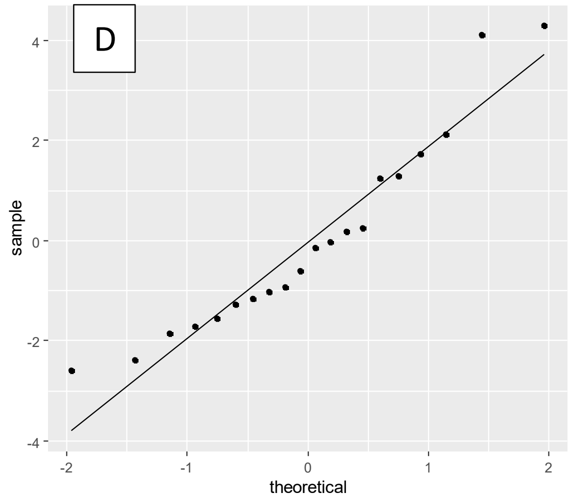
```
> plant_growth %>% group_by(moisture) %>%
+     summarize(mean = mean(growth_rate), sd = sd(growth_rate))

# A tibble: 4 x 3
  moisture  mean    sd
  <fct>    <dbl> <dbl>
1 12        7.39  1.57
2 16       17.0   2.56
3 20       19.3   1.72
4 24       13.3   2.49
```

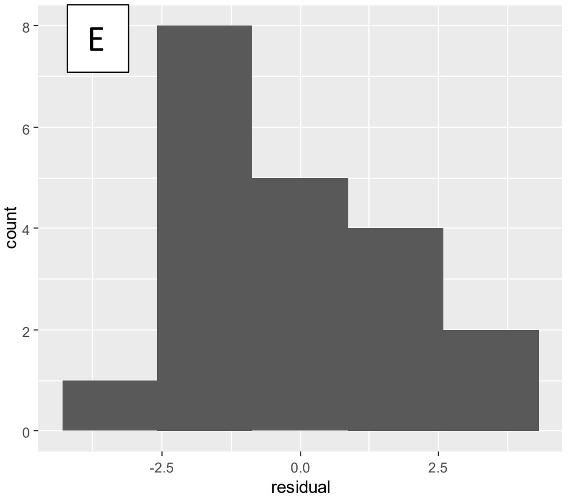
A

```
> plant_growth_model <- lm(growth_rate ~ moisture, data = plant_growth)
> plant_growth <- mutate(plant_growth, residual = resid(plant_growth_model))

> ggplot(plant_growth, aes(x=moisture, y=growth_rate)) + geom_jitter()     #B
> ggplot(plant_growth, aes(y = residual)) + geom_boxplot()                 #C
> ggplot(plant_growth, aes(sample =residual)) + geom_qq() + geom_qq_line() #D
> ggplot(plant_growth, aes(x = residual)) + geom_histogram(bins=5)         #E
```









OUTPUT CONTINUED ON NEXT PAGE

F

```
> anova(plant_growth_model)
Analysis of Variance Table

Response: growth_rate
          Df Sum Sq Mean Sq F value    Pr(>F)    
moisture   3 406.92 135.640  29.906 8.549e-07 ***
Residuals 16  72.57   4.536                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

G

```
> summary(plant_growth_model)

Call:
lm(formula = growth_rate ~ moisture, data = plant_growth)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5987 -1.3394 -0.3677  1.2485  4.2924 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.3890     0.9524   7.758 8.24e-07 ***
moisture16    9.5669     1.3469   7.103 2.50e-06 ***
moisture20   11.9592     1.3469   8.879 1.40e-07 ***
moisture24    5.8620     1.3469   4.352 0.000494 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.13 on 16 degrees of freedom
Multiple R-squared:  0.8487,	Adjusted R-squared:  0.8203 
F-statistic: 29.91 on 3 and 16 DF,  p-value: 8.549e-07
```

H

```
> aov(plant_growth_model) %>% TukeyHSD()

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = plant_growth_model)

$moisture
           diff       lwr        upr     p adj
16-12  9.566921  5.713346 13.4204972 0.0000136
20-12 11.959207  8.105631 15.8127832 0.0000008
24-12  5.862003  2.008428  9.7155793 0.0025033
20-16  2.392286 -1.461290  6.2458618 0.3198776
24-16 -3.704918 -7.558494  0.1486579 0.0616848
24-20 -6.097204 -9.950780 -2.2436280 0.0017581
```