Final Exam (Version A)

Stat 218: Applied Statistics for Life Science

Name: _____

2023-12-11

Read and sign the honesty pledge at the bottom of this page. Your exam will not be graded unless the honesty pledge is signed!

- Write legibly. You can lose points if I can't read your answers.
- You have 2 hours and 50 minutes to complete this exam.
- This exam is worth 100 points.
- You may use a calculator, and two (8.5" x 11") sheet of notes (front and back).
- I have provided you with the table of scenarios.
- Please write clearly in the spaces provided.
- Good luck!

Honesty pledge (please read and sign):

I understand that giving or receiving help on this exam is a violation of academic regulations and is punishable by a grade of F in this course. This includes using any unauthorized materials, a cell phone or other cellular device for any reason, looking at other students' exams and/or allowing other students, actively or passively, to see answers on my exam. This also includes revealing, actively or passively, any information about the exam to any member of Dr. Robinson's STAT 218 class who has not yet taken the exam.

I pledge not to do any of these things.

Signed: _____

Q1. Write a "C" if the variable is categorical or a "N" if the variable is numeric. (4 pts)

_____ Distance

_____ State of Residence

_____ Eye color

_____ Plant yield

Q2. In most statistical studies, the _____ is unknown and the _____ is known. (2 pts)

a. observational unit / variable

b. statistic / parameter

c. parameter / statistic

d. variable / observational unit

Q3. A student participates in a Coke versus Pepsi taste test. She correctly identifies which soda is which five times out of six tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You have studied statistics and you want to determine the probability of anyone getting at least five right out of six tries just by chance alone. Which of the following would provide an accurate estimate of that probability? (2 pts)

a. Simulate this on the computer with a 50% chance of guessing the correct soft drink on each try, and calculate the percent of times there are five or more correct guesses out of six trials.

b. Have the student repeat this experiment many times and calculate the percentage of times she correctly distinguishes between the brands.

c. Repeat this experiment with a very large sample of people and calculate the percentage of people who make five correct guesses out of six tries.

d. All of the methods listed above would provide an accurate estimate of the probability.

Q4. Tennis players often spin a racket to decide who serves first. The spun racket can land with the manufacturer's label facing up or down. A reasonable question to investigate is whether a spun tennis racket is equally likely to land with the label facing up or down. (If the spun racket is equally likely to land with the label facing in either direction, we say that the spinning process is fair.) Suppose that you gather data by spinning your tennis racket 100 times, each time recording whether it lands with the label facing up or down. Does this racket-spinning study call for a one-sided or a two-sided alternative? (2 pts)

a. One-sided because there is only one variable.

b. Two-sided because there are two possible outcomes.

c. Two-sided because the researchers want to know whether the spinning process is fair or biased in either direction.

d. One-sided because the researchers want to know whether the label is more likely to land face up.

Q5. [12 points] A large university is curious if they should build another cafeteria. They plan to survey their students to see if there is strong evidence that the proportion interested in a meal plan is higher than 40%, in which case they will consider building a new cafeteria.

i. State the parameter of interest, in words and assign a symbol (3 pts)

ii Circle the correct answer to complete the hypotheses. The hypotheses being tested could be stated as: (4 pts)

H_0 : The proportion of all students interested in a meal plan is (EQUAL / GREATER / LESS) _____ 0.40. Therefore, they (SHOULD / SHOULD NOT) _____ build a new cafeteria).

H_A : The proportion of all students interested in a meal plan is (EQUAL / GREATER / LESS) _____ 0.40. Therefore, they (SHOULD / SHOULD NOT) _____ build a new cafeteria).

iii. Suppose a researcher carries out this study and finds a p-value of 0.24. Which of the following errors could the researcher have made? (3 pts)

a. Type I

b. Type II

c. Type III

Briefly explain your decision.

iv. What would be the consequence of a Type II error in this case? (2 pts)

a. They don't consider building a new cafeteria when they should.

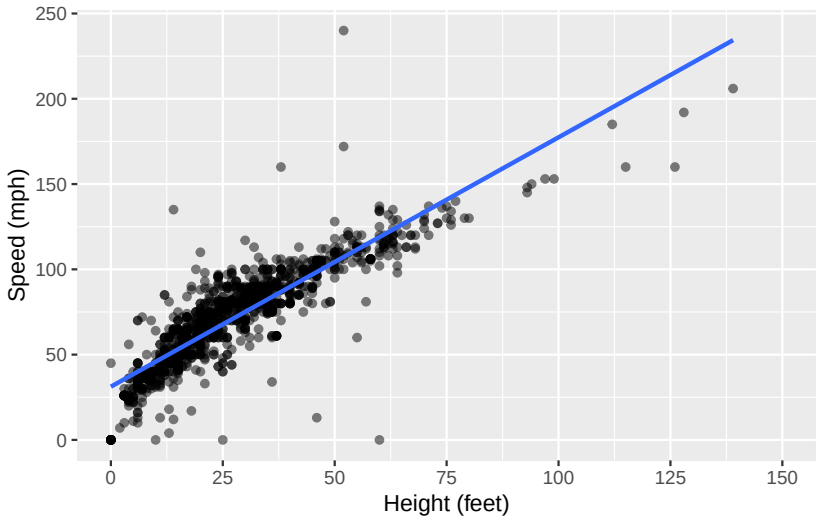
b. They don't consider building a new cafeteria when they shouldn't.

c. They consider building a new cafeteria when they shouldn't.

d. They consider building a new cafeteria when they should.

Q12 [24 points] This data set includes information on roller coasters from all around the world. It can be used to study the popularity of different types of roller coasters, the most popular parks and manufacturers, the relationship between height and speed, and much much more

Research Question: Is there a relationship between speed (miles per hour) of a roller coaster and the height (in feet)?



The R output for the least squares regression line is shown below.

```
# A tibble: 2 x 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl> <chr>      <dbl>     <dbl>
1 (Intercept)    31.2       0.67      46.4 <0.001     29.9       32.5
2 height          1.46      0.02      69.8 <0.001      1.42       1.5
```

i. Identify the variables and their data types. (4 pts).

Explanatory: Categorical / Numeric

Response: Categorical / Numeric

ii. Which of the following correlation coefficients is most reasonable for this data. (2 pts)

a. r = 0.881

b. r = 0.442

c. r = - 0.912

d. r = - 0.287

iii. Based on the R output, write out the equation of the line, in context of the scenario. (3 pts)

iv. What makes this line the "best"? (3 pts)

a. It takes the ratio of the sum of squared groups over the sum of squared errors.

b. It minimizes the sum of squared residuals

c. It maximizes the sum of squared residuals.

d. It connects the first and last point.

v. Interpret the estimated slope **in context**. *Note: If you did not find the estimated slope from above, interpret the value 2.* (3 pts)

vi. Use the equation to predict the speed of a roller coaster that is 50 feet tall. Show your work. *Note: If you did not find the equation of the line above, use* $30 + 2 \cdot x$. (2 pts)

vii. Use the equation to predict the speed of a roller coaster that is 200 feet tall. Show your work. *Note: If you did not find the equation of the line above, use* $30 + 2 \cdot x$. (2 pts)

viii. Which prediction should you have more confidence in? (2 pts)

a. A roller coaster that is 50 feet tall.

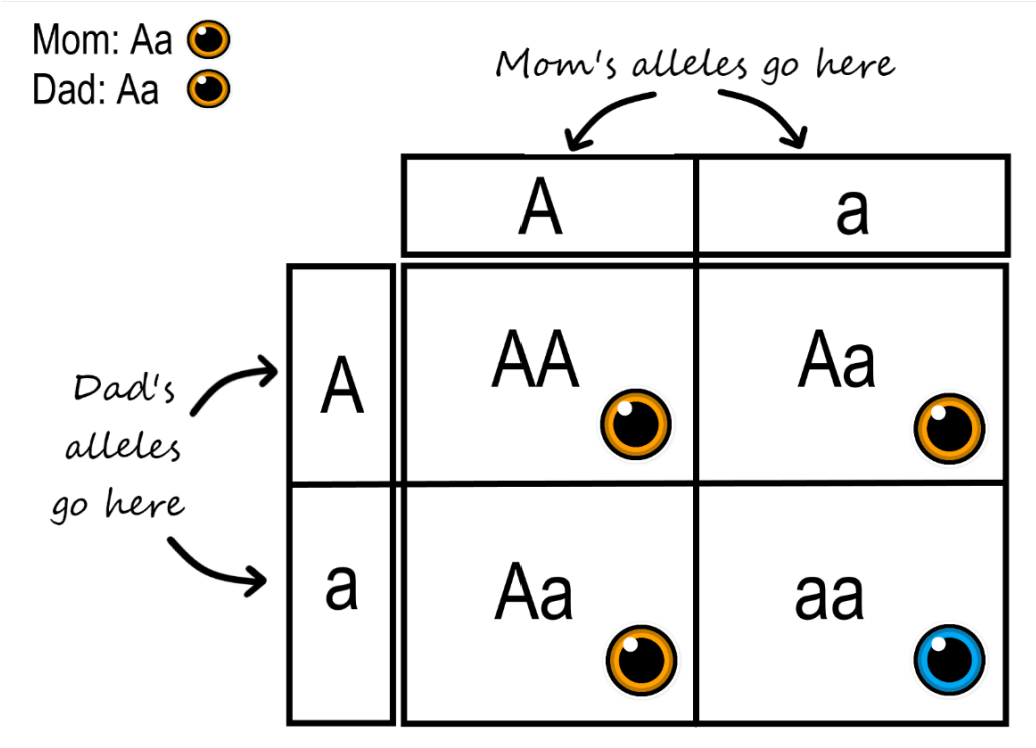
b. A roller coaster that is 200 feet tall.

ix. The observed value for a roller coaster that is 50 feet tall is 63 miles per hour. Compute the residual for this data point. (2 pts)

Q13. [18 points] Mendelian inheritance refers to certain patterns of how traits are passed from parents to offspring. These general patterns were established by the Austrian monk Gregor Mendel, who performed thousands of experiments with pea plants in the 19th century. Mendel's discoveries of how traits (such as color and shape) are passed down from one generation to the next introduced the concept of dominant and recessive modes of inheritance.

Mendelian inheritance refers to the inheritance of traits controlled by a single gene with two alleles, one of which may be completely dominant to the other. You can use a Punnett square to determine the expected ratios of possible genotypes in the offspring of two parents.

In the table below, we see an example of eye color inheritance. In this case, both parents are heterozygotes (Aa) for the gene. Half of the gametes produced by each parent will have the A allele, and half will have the a allele, shown on the side and the top of the Punnett square. Filling in the cells of the Punnett square gives the possible genotypes of their children. It also shows the most likely ratios of the genotypes, which in this case is 25% AA, 50% Aa, and 25% aa.



i. When Mendel crossed his pea plants, he learned that tall (T) was dominant to short (t). Suppose in your Biology course you carried out an experiment to test if the plot offspring would follow Mendelian inheritance. (2 pts)

Fill in the cells of Punnett square to give the possible genotypes for plant tallness.

	T	t
T		
t		

ii. If the Mendelian inheritance is true, what proportions would you expect for each of the following genotypes? Insert the corresponding values in each cell. (2 pts)

TT	Tt	tt
$\pi_{TT} =$	$\pi_{Tt} =$	$\pi_{tt} =$

iii. Actually, our table could be a bit simpler. Both the TT and Tt genotypes will present as "tall" plants, whereas tt genotypes will present as "short" plants.

Compress your previous table into a new table with only two levels of tallness. (2 pts)

Tall	Short
$\pi_{\text{Tall}} =$	$\pi_{\text{Short}} =$

iv. If the table above represents what Mendelian inheritance assumes to be true about tallness under H_0, state the alternative hypothesis using words. (2 pts)

v. After you cross your plants, you measure the characteristics of the 400 offspring. You note that there are 310 tall pea plants and 90 short pea plants.

Fill in the table summarizing these observed counts. (2 pts)

Tall	Short	Total
		400

vi. Fill in the table below, summarizing the expected counts for these 400 plants. (2 pts)

Tall	Short	Total
		400

vii. Calculate how far "off" was your observed number of tall and short plants were from what you expected if H_0 was true. Use these values to report the X^2 statistic for your experiment. (3 pts)

Tall:

Short:

X^2 statistic:

viii. The p-value associated with your X^2 statistic is 0.248. Your Biology textbook suggests you interpret this value as:

The large p-value proves that Mendelian inheritance is true.

What issue(s) to you have with this interpretation? (3 pts)

Q14. [24 points] The *Journal of Food and Agriculture* contained an article titled "Influence of hydroponic ad soil cultivation on quality and shelf life of ready-to-eat lamb's lettuce." In this article, researchers studied the effects of different hydroponic growing methods on the nitrate concentration of lettuce. In their study, the researchers randomly assigned 34 lettuce seedlings to one of three growing methods: soil, hydroponic A, or hydroponic B. At the end of the growing period (60 days), nitrate measurements of the lettuce were taken (mg / kg).

Research Question: Does the growing method affect the nitrate concentration in lettuce?

Results from the study are presented in the table below.

```
# A tibble: 3 x 4
  `Growing Method` `Mean Nitrate` `Standard Deviation of Nitrate` `Sample Size`
  <fct>                     <dbl>                           <dbl>         <int>
1 Soil                      3784.                            199.             9
2 Hydroponic A              4703.                            121.            12
3 Hydroponic B              3864.                            118.            13
```

i. One of the researcher's main questions was to determine whether the growing method affects nitrate concentration in lettuce. Considering how this study was executed, can they address this question? *Briefly justify your answer.* (3 pts)

Below is an incomplete ANOVA table, summarizing the data. You may use this information for the subsequent problems.

```
# A tibble: 2 x 6
  term           df       sumsq   meansq statistic p.value
  <fct>          <chr>    <dbl>    <dbl> <chr>     <chr>
1 Growing Method ___   5934436. 2967218. ___       <0.0001
2 Residuals      31     641263.   20686. <NA>      <NA>
```

ii. Use **symbols** to set up the null and alternative hypotheses investigated in the ANOVA analysis above. (3 pts)

H_0:

H_A:

iii. The alternative hypothesis investigated in the ANOVA table above is (2 pts)

$$H_A : \mu_{\text{Soil}} \neq \mu_{\text{Hydroponic A}} \neq \mu_{\text{Hydroponic B}}.$$

Circle one. (2 pts)

TRUE FALSE

iv. What are the degrees of freedom associated with `Growing Method`? (2 pts)

v. What is the value of the F-statistic? Show all work for credit. (3 pts)

vi. The value of the F-statistic would be larger if the nitrate standard deviations were smaller for each growing method. Circle one. (2 pts)

TRUE FALSE

vii. The value of the F-statistic would be larger if the nitrate means were more different across the growing methods. Circle one. (2 pts)

TRUE FALSE

viii. Which distribution was used to obtain the p-value presented in the table? Circle one.

a. t-distribution

b. F-distribution

c. Chi-square distribution

d. Binomial Distribution

ix. Citing values from the ANOVA table to support your answer, what conclusions could be drawn regarding the hypotheses stated above? (4 pts)

x. The table below presents all comparisons of soil treatment. What value of α^* should the researchers use from Bonferroni's adjustment to determine which of these tests produced "significant" results, so that the overall Type I error rate for these tests is less than 5%? (2 pts)

```
# A tibble: 3 x 3
  `Group 1`    `Group 2`    p.value
  <fct>        <fct>        <chr>
1 Hydroponic B Hydroponic A <0.0001
2 Soil         Hydroponic A <0.0001
3 Soil         Hydroponic B 0.2063
```

Q15. For each of the following, select the single most appropriate analysis for the situation described. Match the letter for the appropriate test needed to address each research question. (2 pts each)

a. Single proportion test (binomial test)
b. Confidence Interval for π
c. Chi-square goodness of fit test
d. Chi-square test
e. Single t-test (one mean)
f. Confidence Interval for μ
g. Two-sample independent t-test
h. Paired t-test
i. ANOVA (F-test)
j. Simple Linear Regression

i. _____ Researchers are interested in investigating how the number of visitors to Yellowstone National Park in a year impacts the local economy in Livingston. To do this they count the number of yearly visitors to Yellowstone and measure the dollars spent by tourists in Livingston for the year.

ii. _____ A study of honeybees looked at whether the species of honeybees varied by state. Ten states were used in the study, and 100 honeybees were randomly sampled in each state, and 7 different species were seen in the data set.

iii. _____ An attorney in Boston observes that some judges seem to select juries that contain few women. She collects data on 20 randomly selected juries from each of 10 judges, and the number of women on each jury for each judge.

iv. _____ A local physical education teacher wants to know if people can throw a softball farther than a baseball. He recruits 30 high school students and has each student throw a baseball and a softball.

v. _____ You are interested in deciding if you should rent a new apartment off campus. As this will be your first time living off campus, you are anxious to know the average amount of time it should take you to walk to campus. What is the best **method** to estimate the average time it will take you to walk to campus?

vi. A cigar salesman was interested in learning if the majority of men think cigars smell good. He randomly selected 25 men. He recorded a "yes" if a man thought they smelled good or a "no" otherwise.

👏 You are primo! Congratulations on completing STAT 218!

Have a great winter break!

_____ / 100

Variable(s) of Interest	Parameter of Interest	Statistic of Interest	Descriptive Method(s)	Inferential Method(s)	Assumption(s) for Inferential Methods
Single Categorical Variable **(Binary – 2 categories)**	True Population Proportion (π)	Sample Proportion ($\hat{p}$)	▪ Report $\hat{p}$ ▪ Bar chart	▪ Simulation ▪ Binomial test ▪ CI for π	Check the four conditions for the binomial
Single Categorical Variable **(More than two categories)**	True Population Proportion $(\pi_1, \pi_2, ...)$	Sample proportions $(\hat{p}_1, \hat{p}_2 \ldots)$	▪ Report $\hat{p}_1, \hat{p}_2$ ... ▪ Stacked bar chart	▪ Chi-square Goodness of Fit	▪ Observations are independent ▪ EXPECTED counts should be greater than 5
Two Categorical Variables **(in general)**	True Conditional Population Proportions $(\pi_{1\|G1}, \pi_{1\|G2}, ...)$	Sample Proportions $(\hat{p}_{1\|G1}, \hat{p}_{1\|G2} \ldots)$	▪ Report sample proportions ▪ Contingency table ▪ Stacked/Dodged/Filled Bar Plot	▪ Chi-square test	▪ Observations are independent ▪ EXPECTED counts should be greater than 5

Variables of Interest	Parameter of Interest	Statistic of Interest	Descriptive Methods	Inferential Methods	Assumptions for Inferential Methods
Single Numerical Variable	True Population Mean (μ)	Sample Mean ($\bar{x}$)	▪ Report measures of center and variation ▪ Dotplot, boxplot, histogram ▪ Describe shape ▪ Identify outliers	▪ One-sample t-test ▪ CI for population mean	▪ Either the sample size is fairly large or the data reasonably follow a normal distribution
Comparing Numerical Variable across Two Categories of a Categorical Variable (DEPENDENT samples)	True Mean Difference (μ_d)	Sample Mean Difference ($\bar{x}_d$)	▪ Report measures of center and variation for the differences ▪ Dotplot, boxplot, histogram of the differences ▪ Describe shape, identify outliers	▪ paired t-test ▪ CI for population mean difference	▪ Independent differences ▪ Either the number of pairs is fairly large or the differences reasonably follow a normal distribution
Comparing Numerical Variable across Two Categories of a Categorical Variable (INDEPENDENT samples)	Difference in True Population Means ($\mu_1 - \mu_2$)	Difference in Sample Means ($\bar{x}_1 - \bar{x}_2$)	▪ Report $\bar{x}_1, \bar{x}_2,$ and s_1, s_2 ▪ Side-by-side boxplots, facetted histograms	▪ Two-sample t-test ▪ CI for $\mu_1 - \mu_2$	▪ Observations are independent between groups ▪ Either both sample sizes are fairly large or the data from each group reasonably follow a normal distribution

Variables of Interest	Parameter of Interest	Statistic of Interest	Descriptive Methods	Inferential Methods	Assumptions for Inferential Methods
Comparing Numerical Variable across 2 or more categories of a Categorical Variable			▪ Group means, group standard dev. ▪ Side-by-side boxplots, facetted histograms	Analysis of Variance (ANOVA) F-test statistic	▪ Independence *between* and *within* groups ▪ Equal variances ▪ Normality
Comparing Two Numerical Variables			▪ Correlation (r) ▪ Scatterplot ▪ Regression line ($\hat{y} = b_0 + b_1 x$)	Linear Regression Analysis Slope = 0?	▪ Linearity ▪ Independence ▪ Normality ▪ Equal Variance