Review Guide for Final Exam

STAT 218: Applied Statistics for Life Science

What to Expect

- You may bring **two** $8\frac{1}{2} \times 11$ standard sheet of notes (both sides). I will not provide you with formulas, so put what you think you need on here.
- Additionally, I will provide you with the table summary of the different situations.
- You may bring any calculator to use. I will have a handful of simple calculators. You may **not** use your phone as a calculator.
- The exam is mostly multiple choice, but will have a couple of short answer questions mixed in.
- You will have 2 hours and 50 minutes to complete the exam.

I have not finalized point assignment for the final yet, but I would expect approximately:

- 40% new material (I would fully expect an ANOVA scenario and Regression scenario on there, who knows, it may even be one from this review!).
- 30% cumulative material.
- 20% determine the appropriate statistical analysis.

> Canvas Discussion Board
>
> Post any logistic or studying questions on the Canvas discussion board. Please respond to each other!

Key Concepts to Review

Note: this does not guarantee every topic on the exam is on the list.

Chapters 0-2: Statistical Thinking + Single Proportion

Visualize with a bar plot of the proportion / count of successes

Summarize with a single proportion / count of successes

- Sample, population, statistics, parameter, sample size
- Symbols for proportions (statistic vs. parameter)
- Test for a single proportion.
  - What impacts the strength of evidence?
- Basics of Simulation
  - Remember we are simulating given the null hypothesis is true
  - How to calculate p-value from simulated null distribution
    - Remember we can never prove the null true
- Confidence intervals one proportion
- What affects confidence intervals?

Chapter 3: One Categorical Variable with more than two levels

Visualize with a stacked bar plot of the proportion / count of each category

Summarize with the proportion/count of each category

- What is a test statistic? How do we use it in general?
- Chi-square goodness of fit test
  - How to find expected values given proportions and a total count.
  - How to calcualte the X^2 test statistic.
  - Determine degrees of freedom?

Chapter 4: Two Categorical Variables

Visualize with a stacked/dodged/filled bar plot of the proportion / count of successes

Summarize with a contingency table

- Explanatory variables & response variables
- Observational study versus randomized experiment
  - Association vs causation
  - Confounding variables / Bias (What are they? How do they affect our study? How might we reduce the risk of having confounding variable?)
  - What do we need to have an experiment? Why can't we always have an experiment?
- Chi-square Test
  - How do we calculate the X^2 test statistic?
    - Expected Values?
  - How do we determine the degrees of freedom for the Chi-square distribution?

Chapters 5/6

Visualize with a histogram/boxplot/dot plot

Summarize with a mean & standard deviation / median and IQR

- Sample, population, statistics, parameter, sample size
- Symbols for means (statistic vs. parameter)
- Test for a single mean
- Confidence intervals one mean

Chapter 7: Comparing Means across two groups

Visualize with faceted histograms / side-by-side boxplots

Summarize with a mean & standard deviation / median and IQR for each group

Summarize Independent with the difference in the two means

Summarize Paired with the mean of the difference between the two groups

- Two-sample Independent t-test vs Paired t-test
  - When should we use each?
    - How are they different from each other?
    - How is the random assignment different?
    - Picking which scenario is two sample and which is paired
- Paired t-test: Can use test for one mean again. (like tests from chapter 6)
  - Find p-value (from simulation)
  - t-statistic
  - Confidence interval (calculation, interpretation)
  - Making conclusion from the confidence interval (i.e. is 0 inside the confidence interval, what does that mean about the two groups)

Chapter 8: ANOVA

Visualize with faceted histograms / side-by-side boxplots

Summarize with a mean & standard deviation / median and IQR for each group

- How do you write out the hypotheses carefully?
- What does the F-test measure?
  - What happens if you increase/decrease variability between groups? Within groups?
- How to find the different parts of the ANOVA table?
- What is multiplicity and why do we need it? How do you use Bonferroni's adjustment?

Chapter 9: Simple Linear Regression

- Regression: quantitative response and explanatory
- Scatterplot: Form, direction, strength, unusual
- Correlation coefficient (r)
  - What are the possible values for r?
  - What values of r denote a strong relationship? Weak relationship?
  - For a large value of r, do points fall very close to the best fit line, or far away? What about for very small values?
- Line of Best Fit (how is it found?)
  - y versus $\hat{y}$
  - Interpreting the slope and y-intercept
  - Making predictions
  - What makes an observed point influential?
- Extrapolation
- Residuals
  - Calculation
  - Predict whether observed point is above or below the best fit line

Scope of Inference

- Random sampling versus random assignment
  - What's the difference between the two? Can we have both?
  - Benefits of both

Practice Problems

General Questions

1. Given the following, when would you find evidence to support the alternative? Think through why.

a. P-value and significance level α.

b. Confidence Interval. What do the values of a confidence interval tell us?

2. What impacts the strength of evidence and how?

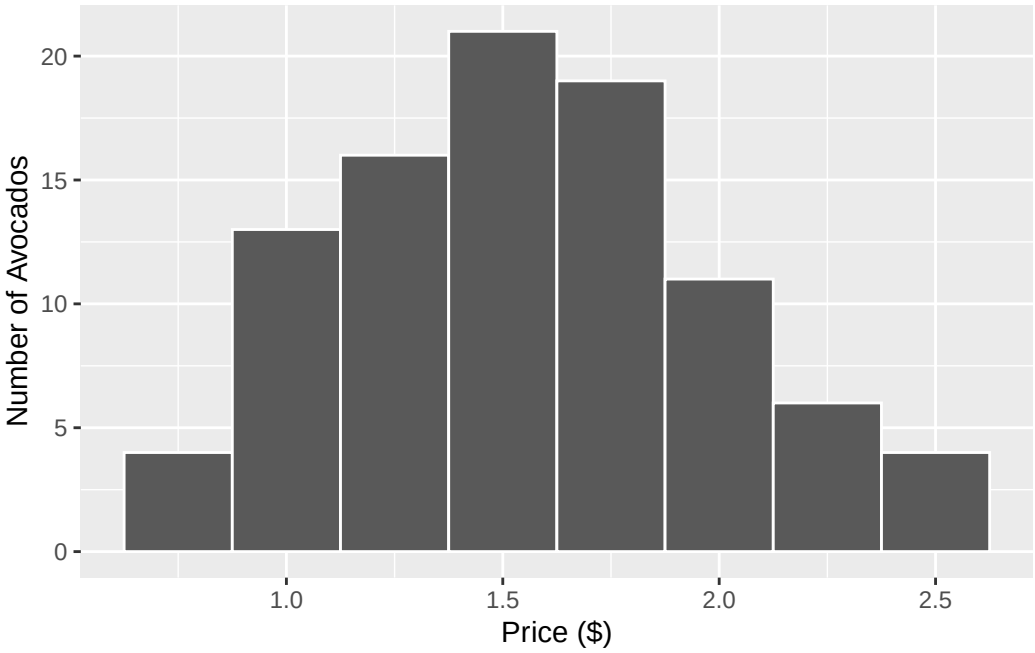
3. Type I and Type II Errors:

- A type I error incorrectly finds evidence to support the alternative when the null is true. TRUE / FALSE
- A type II error incorrectly *does not* find evidence to support the alternative when the null is false. TRUE / FALSE

Avocado Prices

4. It is a well-known fact that Millennials LOVE Avocado Toast. It's also a well-known fact that all Millennials live in their parent's basements. Clearly, they aren't buying home because they are buying too much Avocado Toast! Was the Avocadopocalypse of 2017 real? In the past, an avocado cost \$1.36. To test the avocadopaocalypse, we collected a random sample of 94 avocados from 2017. The average price of an avocado was \$1.44 with a sample standard deviation of 0.41.

Research Question: Is the price of an avocado in 2017 higher than the price of an avocado in the past?



min	Q1	median	Q3	max	mean	sd	n	missing
0.78	1.29	1.55	1.815	2.61	1.560319	0.4283237	94	0

i. State the variable of interest and identify its type.

ii. In words, state the parameter of interest and assign the correct symbol.

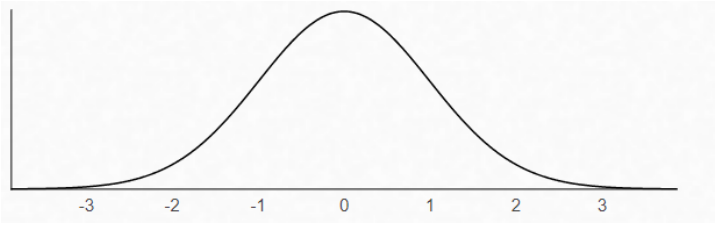
iii. Identify the sample mean, sample standard deviation, and sample size (and assign symbols).

iv. Using symbol notation, state the null and alternative hypothesis.

v. Verify the conditions necessary to use a single t-test to answer the research question.

vi. Calculate the test statistic and assign an appropriate symbol.

vii. On the distribution below, show how you would determine the p-value.



viii. Write a conclusion in context of the research question.

ix. What type of error could have possibly occurred based off of your conclusion?

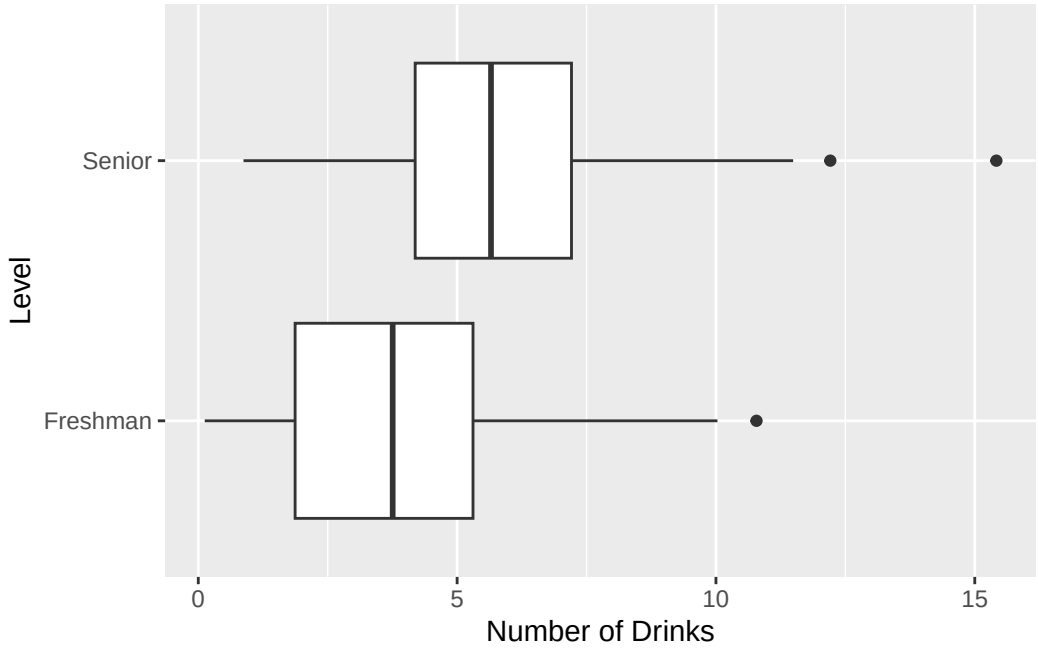
a. Type I: We concluded the price of an avocado in 2017 is greater than in the past when it actually is the same price.

b. Type I: We concluded the price of an avocado in 2017 is not greater than in the past when it actually is the same price.
c. Type II: We concluded the price of an avocado in 2017 is greater than in the past when it actually is the same price.
d. Type II: We concluded the price of an avocado in 2017 is not greater than in the past when it actually is the same price.

Alcoholic Drinks

5. A study was conducted to compare the mean number of alcoholic drinks consumed on a typical night between freshmen versus seniors at Cal Poly. In an anonymous survey, a random sample of 105 freshmen and 31 seniors were asked, *"How many alcoholic drinks do you have on a typical night in which you consume alcohol?"*

Research Question: Is there evidence that the mean number of drinks consumed differs between freshmen and seniors? The R output is shown below.




```
     level       min       Q1   median       Q3      max     mean       sd   n
1 Freshman 0.1249162 1.870711 3.755602 5.307095 10.78342 3.848318 2.442048 105
2   Senior 0.8740698 4.190036 5.654873 7.209817 15.41719 6.123835 3.096234  31
  missing
1       0
2       0
```

i. Write H_0 and H_A based on the research question using appropriate symbols.

ii. The normality assumption for the two-sample t-test is met in this case. Explain how you know this.

iii. The t test-statistic for this problem is given by $t = 3.76$. Show how this t-statistic was computed.

```
# A tibble: 1 x 7
  statistic  t_df  p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
1      3.76  41.6 0.000522 two.sided       2.28     1.05     3.50
```

iv. Write a conclusion addressing the research question in the context of the problem. State the test statistic, degrees of freedom, and p-value as part of your conclusion.

v. Find and interpret the 95% confidence interval for the difference in means, $\mu_{Senior} - \mu_{Freshman}$.

```
qt(0.975, 42)
```

```
[1] 2.018082
```

Wild Mushrooms

6. Wild mushrooms, such as chanterelles or morels, are delicious, but eating wild mushrooms carries the risk of accidental poisoning. Even a single bite of the wrong mushroom can be enough to cause fatal poisoning. An amateur mushroom hunter is interested in finding an easy rule to differentiate poisonous and edible mushrooms. They think that the mushroom's gills (the part which holds and releases spores) might be related to a mushroom's edibility. They used a data set of 8124 mushrooms and their descriptions. For each mushroom, the data set includes whether it is edible or poisonous and the spacing of the gills (Broad or Narrow).

Please Note: According to The Audubon Society Field Guide to North American Mushrooms, there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, leave them be" for Poisonous Oak and Ivy.

Class	Broad	Narrow	Total
Edible	3920	288	4208
Poisonous	1692	2224	3916
Total	5612	2512	8124

i. Fill in each blank with one of the options in parentheses to best describe the variables collected.

Whether the mushroom is edible or poisonous is the (explanatory / response) and it is (categorical / quantitative).

Gill size (Broad or Narrow) is the (explanatory / response) and it is (categorical / quantitative).

ii. Calculate the proportion of mushrooms with a broad gill size that are poisonous. *Leave your value in* ***unreduced*** *fraction form.*

$$\underbrace{\qquad\qquad}_{\text{(notation)}} = \underbrace{\qquad\qquad}_{\text{(value)}}$$

iii. Calculate the proportion of mushrooms with a narrow gill size that are poisonous. *Leave your value in* ***unreduced*** *fraction form.*

$$\underbrace{\qquad\qquad}_{\text{(notation)}} = \underbrace{\qquad\qquad}_{\text{(value)}}$$

iv. Using your answers to (ii) and (iii), fill in the correct names next to each color, to label the stacked bar chart showing the relationship between gill size (broad or narrow) and whether the mushroom is edible.

v. Based on the plot, describe the relationship between a mushrooms gill size and whether it is edible or not.

vi. Suppose the Chi-Squared test resulted in a p-value of 0.023. Which of the following would be the correct scope of inference for this study?

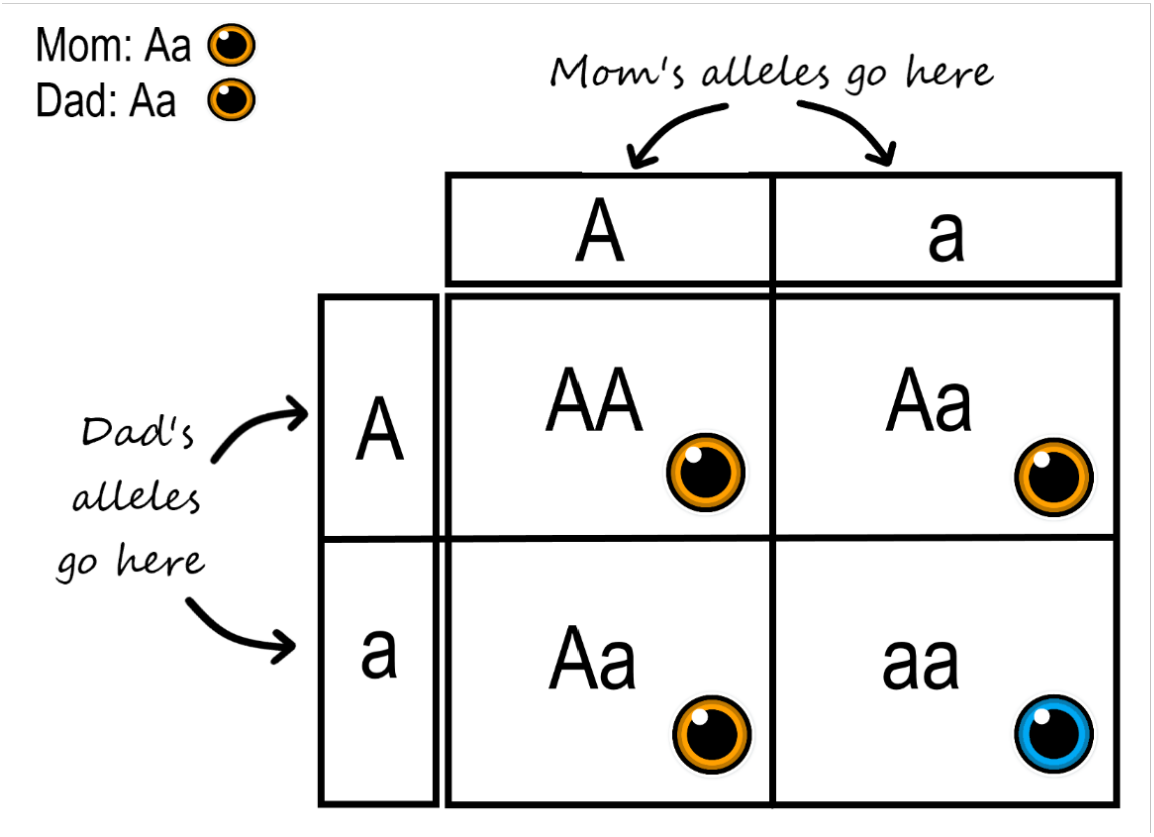
a. It can be inferred for all mushrooms that gill size causes a mushroom to be poisonous.
b. It can be inferred for all mushrooms that gill size is associated with whether a mushroom is poisonous.
c. It can be inferred for this sample of mushrooms that gill size causes a mushroom to be poisonous.
d. It can be inferred for this sample of mushrooms that gill size is associated with whether a mushroom is poisonous.

Mendelian Genetics

7. 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.

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.

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.

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

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

v. After you cross your plants, you measure the characteristics of the 400 offspring. You note that there are 305 tall pea plants and 95 short pea plants.

Fill in the table summarizing these observed counts.

_____	_____	Total
		400

vi. Fill in the table below, summarizing the expected counts for these 400 plants.

_____	_____	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.

Tall:

Short:

X^2 statistic:

viii. The p-value associated with your X^2 statistic is 0.5645424. 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?

Soil Samples

8. 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 content of lettuce. In their study, the researchers randomly assigned 34 lettuce seedlings to one of three growing conditions: soil, hydroponic A, or hydroponic B. At the end of the growing period (60 days), nitrate measurements of the lettuce were taken (mg / kg).

Results from the study are presented in the table below.

```
# A tibble: 3 x 4
  `Treatment Group` `Mean Growth` `Standard Deviation of Growth` `Sample Size`
  <fct>                     <dbl>                          <dbl>         <int>
1 Soil                      3800.                           149.             9
2 Hydroponic A              4725.                           116.            12
3 Hydroponic B              3915.                           109.            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.*

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 ___   5771863. 2885932. ___       <0.0001
2 Residuals      31     468314.   15107. <NA>      <NA>
```

ii. In the context of the research and in plain English, what are the null and alternative hypotheses investigated in the ANOVA analysis above?

H_0:

H_A:

iii. Rewrite the null hypothesis above to use **symbols** for the parameters that are being tested.

H_0:

H_A:

iv. The alternative hypothesis investigated in the ANOVA table above is

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

Circle one.

True **False**

v. What are the degrees of freedom associated with `Growing Method`?

vi. What is the value of the F-statistic?

vii. The value of the F-statistic would be larger if the nitrate standard deviations were smaller for each group. Circle one.

True **False**

viii. The value of the F-statistic would be larger if the nitrate means were more different across the groups. Circle one.

True **False**

ix. 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

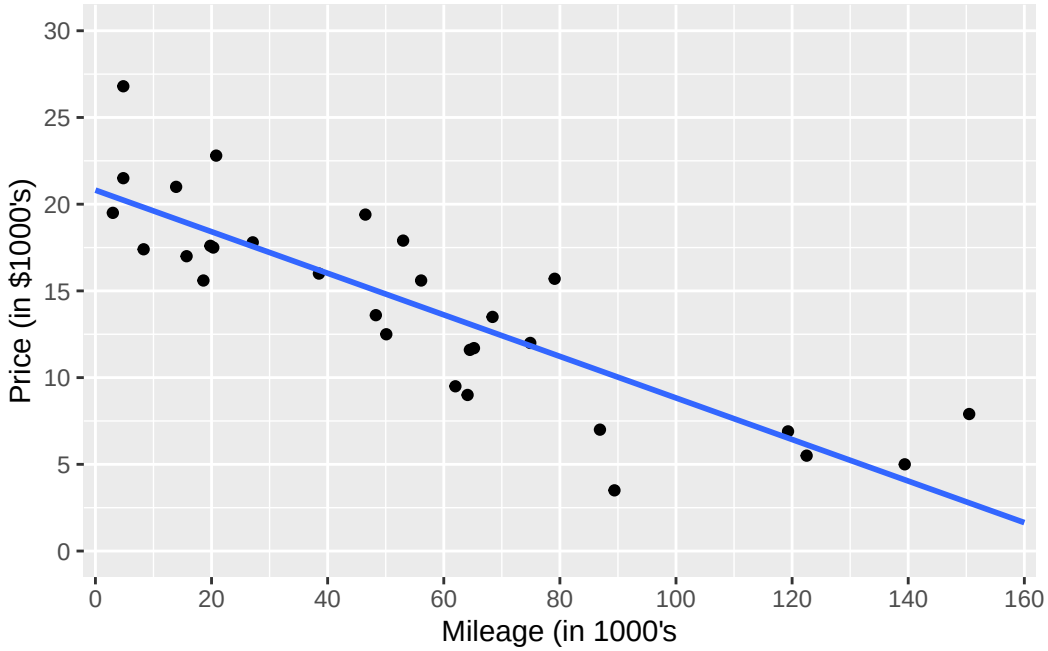
x. Citing values from the ANOVA table to support your answer, what conclusions could be drawn regarding the hypotheses stated above?

xi. 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%?

```
# 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 0.0389
```

Honda Accord

9. A scatterplot of Price (asking price in dollars) versus Mileage (previous miles driven) for a sample of used Honda Accords is shown below. A simple linear regression line was fit to the data using R, and the results are as follows.




```
# 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)   140.       10.9       12.9 <0.001    118.      163.
2 Price          -6.02      0.71      -8.5 <0.001     -7.46     -4.57
```

i. Which variable was used as the response variable in this analysis: Price or Mileage?

ii. Find the slope of the fitted regression line and interpret this slope in the context of the problem (i.e., use the slope to provide a detailed explanation of the relationship between Price and Mileage).

iii. What makes this the "best line"?

iv. Interpret the slope in context of the problem.

v. Can we interpret the y-intercept? If yes, interpret in context, if not, explain why not.

vi. The t-statistic for testing whether the slope differs from zero is given by t = -10.40. Show how this was computed.

vii. Is there evidence that Mileage is an effective predictor of Price? State yes or no and provide a p-value to justify your answer.

viii. Calculate the residuals for the provided two cars.

```
  Age Price Mileage
4   7  12.5    50.1
8   2  22.8    20.8
```

Scope of Inference

10. Researchers randomly selected participants and then randomly assigned them to either walk for half an hour three times a week or to sit quietly reading a book for half an hour three times a week. At the end of a year, the change in the participants' blood pressure over the year was measured, and the change was compared for the two groups.

i. This is a randomized controlled experiment (i.e., a designed experiment) because

a. `Blood pressure was measured at the beginning and end of the study.`

b. The two groups were compared at the end of the study.

c. The participants were randomly assigned to either walk or read, rather than choosing their own activity.

d. A random sample of participants was used.

ii. If a statistically significant difference in blood pressure change between the two activities is found, then which of the following conclusions is most correct?

a. It cannot be concluded that the difference in activity caused a difference in the change in blood pressure because in the course of a year there are lots of possible confounding variables.

b. It cannot be concluded that the difference in activity caused a difference in the change in blood pressure because it might be the opposite; people with high blood pressure might be more likely to read a book than to walk.

c. It can be concluded that the difference in activity caused a difference in the change in blood pressure because of the way the study was done; the randomization of subjects to treatment groups should balance out the effects of any confounding variables.

11. To assess the effects of tanning beds on the incidence of skin cancer, a researcher collected and compared mole biopsies from individuals who reported that they used tanning beds regularly and compared them to the mole biopsies from individuals who reported never using a tanning bed. The researcher analyzed the moles for cancerous cells and was unaware of which samples were from the regular tanning group and the group that had never used a tanning bed.

i. Is this an observational study or a designed experiment?

a. Observational study because the people themselves chose whether to use tanning beds.
b. Designed experiment (i.e., randomized controlled experiment) because one group used tanning beds and the other group did not.
c. There is not enough information provided to decide whether this was an observational study or a designed experiment.

ii. Suppose there were a significantly higher number of cancerous moles for the tanning bed users than for the non-tanning bed users. Can you conclude based on this study that using tanning beds leads to cancerous moles?

a. Yes - the researcher was blind to which samples belonged to those who used tanning beds and those who didn't.
b. No - there may be other lifestyle differences between people who choose to tan and people who do not (e.g., sunscreen habits, differences in diet, etc.)
c. Yes - there were a significantly higher number of cancerous moles in the individuals who used tanning beds.

Type I and Type II Errors

12. 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. Let π represent the proportion of all students interested in a meal plan.

i. State the Null and Alternative Hypotheses in symbols.

ii. What would be the consequence of a Type II error in this case?

   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.

13. A large nationwide poll recently showed an unemployment rate of 9% in the US. The mayor of a local town wonders if this national results holds true for her town, so she plans on taking a sample of her residents to see if the unemployment rate is significantly different than 9% in her town. Let π represent the long run proportion of unemployment in her town

i. State the correct Null and Alternative Hypotheses in symbols.

ii. Under which of the following conditions would the mayor commit a Type I error?

   a. She concludes the town's unemployment rate is not 9% when it actually is.
   b. She concludes the town's unemployment rate is not 9% when it actually is not.
   c. She concludes the town's unemployment rate is 9% when it actually is.
   d. She concludes the town's unemployment rate is 9% when it actually is not.

3. Donated blood is tested for infectious diseases and other contaminants. Since most donated blood is safe, it saves time and money to test batches of donated blood rather than test individual samples. A certain test is performed to see if a certain toxin is present, and the entire batch is discarded if the toxin is detected. This is similar to using a null and an alternative hypothesis to determine whether to discard or keep the batch. The hypotheses being tested could be stated as:

i. Circle the correct answer to complete the hypotheses.

   H_O : The batch DOES / DOES NOT contain the toxin. H_A : The batch DOES / DOES NOT contain the toxin.

ii. Suppose a researcher carries out this study and finds a p-value of 0.031. Which of the following errors could the researcher have made? Briefly explain your decision.

iii. Describe the consequence of the error above in this context.

Determining the Appropriate Statistical Analysis

For each scenario, determine the following:

i. State the appropriate test needed to address the research questions

   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

ii. State the parameter of interest in words and symbols.
iii. Write the hypotheses in symbols.

14. The average GPA at University of Alabama is historically known to be 3.05. The registrar plans to look at a random sample of records for students graduating in 2017 to see if average GPA has increased.

15. The General Social Survey asked 1381 randomly selected respondents to give "the gender of your best friend." Researchers are interested in whether the respondent's best friend was the same gender as the respondent.

16. The 2009 Sleep in America poll took a random sample of 100 adults to see if people tend to sleep more on weekends. They asked them to report the typical amount of sleep (in hours) they get on weekdays and on weekends.

17. The Women's Health Initiative randomized post-menopausal women to one of two treatments: estrogen plus progestin hormone therapy or a placebo control. They were interested in whether or not the women developed cancer over the next five years.

18. Thirty pilots performed tasks at a simulated altitude of 25,000 feet. The pilots performed the tasks both completely sober and after drinking alcohol (the order was randomized, and there were 3 days between the tests). The response variable was time (in minutes) of useful performance of the tasks. The longer a pilot spends on useful performance, the better.

19. Do women tend to spend more time on housework than men? The National Survey of Families and Households took a random sample of 972 U.S. adults and asked them how much time they spend on housework per week.

20. Is arthroscopic surgery better than placebo for alleviating 121 subjects were randomly assigned to either undergo arthroscopic surgery or a placebo surgery. After two years, they were asked their knee pain score on a pain scale of 0-100.

21. A random sample of 15 high school sophomores take the SAT both before and after undergoing an intensive training course designed to improve such test scores.

22. Suppose you are investigating the following research question: "Does GPA differ, on average, between those who work during school and those who don't?" Which of the following analyses is most appropriate for investigating this research question?

23. A national survey of college students indicated that 90% of college students text while driving. Suppose you are using data from the random sample of Cal Poly students to investigate the following research question: "Is the proportion of Cal Poly students who text while driving less than the national average, 90%?" Which of the following analyses is most appropriate for investigating this research question?

24. Suppose you are using data from the random sample of Cal Poly students to investigate the following question: "Does the proportion who drink alcohol differ between student athletes and non-student athletes?" Which of the following analyses is most appropriate for investigating this research question?

25. A researcher is interested in determining if one could predict the score on a statistics exam from the amount of time spent studying for the exam. She randomly selected 31 statistics students and gathered their exam score as well as the length of time (in minutes) that they studied.

26. Matchmaking data scientists are always investigating what characteristics of a person can produce better matches. Data scientists at Hinge are interested in looking into the relationship between someone's sexual orientation and whether they would date someone who is taller than them.

27. 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?

28. Researchers are interested in determining if the yield differs among seven grape varieties.