Review Guide for Midterm 2 Exam

STAT 218: Applied Statistics for Life Science

What to Expect

- You may bring an $8\frac{1}{2} \times 11$ standard sheet of notes (both sides). I will not provide you with formulas, so put what you think you need on here.
- You will turn this in with your exam.
- 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 50 minutes to complete the exam.

> 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 may not be an exhaustive list. You should review all of your notes, assignments, labs, and quizzes from Chapters 4 - 7 (and concepts that are necessary to build off of from the first three chapters).

- Can you describe a distribution for a numerical variable? (shape, center, spread, outliers)
- Identify summary statistics (with symbols) from `favstats` output.
- Identifying pieces of a data set: observation, variable, sample size, parameter, etc.
- Work through hypotheses testing for the following types of scenarios (how does parameter wording change, how does the sampling distribution change?):
  - Two categorical variables (Chi-square Test of Independence)
  - One numerical variable (t-test for a single mean)

  - Comparing a numerical variable across two **independent** groups (two-sample independent t-test)
  - Comparing a numerical variable across two **dependent** groups (paired t-test)

- Understand the concept of a sampling distribution for
  - means
  - differences in means
  - mean of the differences.
- Confidence intervals.
  - How do you find the t-quantile for different confidence levels?
  - How do confidence intervals and levels relate to hypotheses testing?
- Scope of Inference
  - How do you know when you can generalize to the population of interest? (Random Sampling / a Representative Sample from xxx)
  - When can you say *causation* versus *association* only? (Random Assignment of xxx)
- How do you interpret the p-value? Can you do this for various scenarios?

Practice Problems

Suggestion: you can "create" your own questions building off of these scenarios. Ask yourself about causation, generalization, practice calculating confidence intervals, etc. If not given a t-quantile for something like this, 2 is typically a good substitute for a 95% confidence interval multiplier.

1. You are thinking of using a t-test to investigate a research hypothesis about the mean of a population. The distribution of your sample is extremely skewed (i.e., not normally distributed). Which of the following statements is most correct?

    a. You may use the t-test provided your sample size is large (say at least 30 or so).
    b. You should not use the t-test regardless of the sample size because the population does not have a normal distribution, so the normality assumption won't be met.
    c. You may use the t-test provided you use a fairly large confidence level, say 99%.
    d. You may not use the t-test, but a confidence interval constructed using the t-distribution should be valid.

2. A 95% confidence interval for the difference in population means between Groups A and B (µA - µB) is given as follows: -3.9   µA - µB   -1.0. Which of the following statements is most correct?

    a. We have evidence that the mean of Group A is larger than the mean of Group B.
    b. We have evidence that the mean of Group A is smaller than the mean of Group B.
    c. We have evidence that the mean of Group A is equal to the mean of Group B.
    d. We can't tell - we can't draw any conclusions unless we have a p-value.

3. Suppose you are investigating the weight of Snickers bars. Your hypotheses are as follows:

    H_0: The mean weight for all Snickers bars is 48 grams ($\mu = 48$).
    H_A: The mean weight for all Snickers bars is more than 48 grams ($\mu > 48$).

You take a random sample of 40 Snickers bars, and you conclude that the mean weight of all Snickers bars is more than 48 grams with a p-value of 0.02.

i. Which of the following statements best describes what that p-value means?

a. If the mean weight of all Snickers bars is 48 grams, the probability that a random sample of 40 Snickers would have a mean as high or higher than your sample mean is 0.02.
b. If the mean weight of all Snickers bars is more than 48 grams, the probability that a random sample of 40 Snickers would have a mean as high or higher than your sample mean is 0.02.
c. Only 2% of all Snickers bars weigh more than 48 grams.
d. The probability that the true mean weight of all Snickers bars equals 48 grams is 0.02.

ii. Suppose that the standard deviation of your sample was $s = 5g$. The standard error (i.e., standard deviation of the sample mean) is:

a. 8
b. 5
c. 0.125
d. 0.79

iii. Suppose a 95% confidence interval is computed to estimate the mean weight of all Snickers bars. Which of the following values will definitely be within the limits of this confidence interval?

a. The population mean, μ.
b. The sample mean, $\bar{x}$.
c. The p-value.
d. None of the above.

4. An article in the San Luis Tribune claims that the average age for people who receive food stamps in SLO is 40 years. A Cal Poly student believes the average age is less than that. The student obtains a random sample of 100 people in SLO who receive food stamps, and finds their average age to be 39.2 years. Performing a hypothesis test, the student finds their sample mean to be statistically significantly lower than the age of 40 stated in the article (p-value < 0.05). Indicate fore each of the following interpretations whether they are valid or invalid.

i. The statistically significant result indicates that the majority of people who receive food stamps are younger than 40.

- valid
- invalid

ii. An error must have been made. This difference in means (39.2 vs. 40 years) is too small to be statistically significant.

- valid
- invalid

5. Can Tai Chi help improve postural stability in Parkinson's patients? To investigate, researchers randomly assigned 65 Parkinson's patients to practice Tai Chi daily for 24 weeks, and another 65 Parkinson's patients to practice stretching daily for 24 weeks. The outcome we will focus on is functional reach (measured in cm.).

```
    Exercise      min       Q1   median       Q3      max     mean       sd  n
1    Tai Chi 17.21915 27.23141 29.81011 33.23330 42.60890 30.14409 4.834595 65
2 Stretching 11.82380 21.03960 24.46298 29.33581 40.86007 25.48402 6.203606 65
  missing
1       0
2       0
```

i. Identify the response and explanatory variables (and their types).

ii. Describe a tactile simulation strategy (using either coins, or spinners, or cards) to calculate a p-value to investigate whether Tai Chi can help improve functional reach in Parkinson's patients. Be sure to provide enough details so that anyone could carry out the simulation by reading just your description here. (*Hint: Think back to the yawn experiment.*)

iii. A 95% confidence interval comparing Tai Chi and Stretching was calculated from the above data: (2.73, 6.59). Interpret this interval in the context of this study.

iv. A 99% confidence interval based on the above data would be the 95% confidence interval. Circle one.

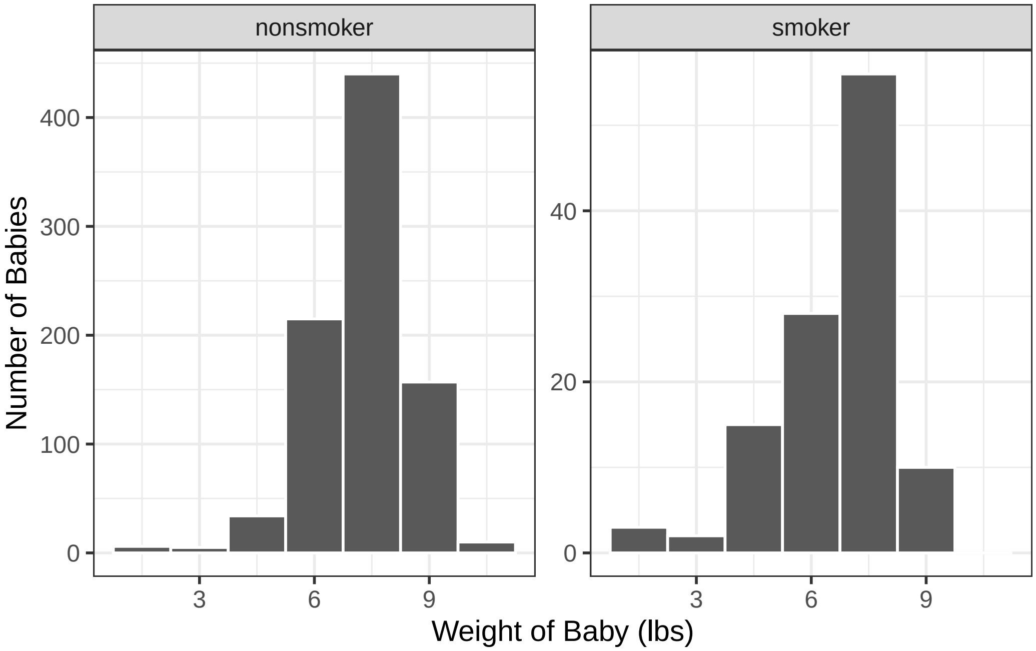
- narrower than 95% confidence interval
- the same as the 95% confidence interval
- wider than 95% confidence interval

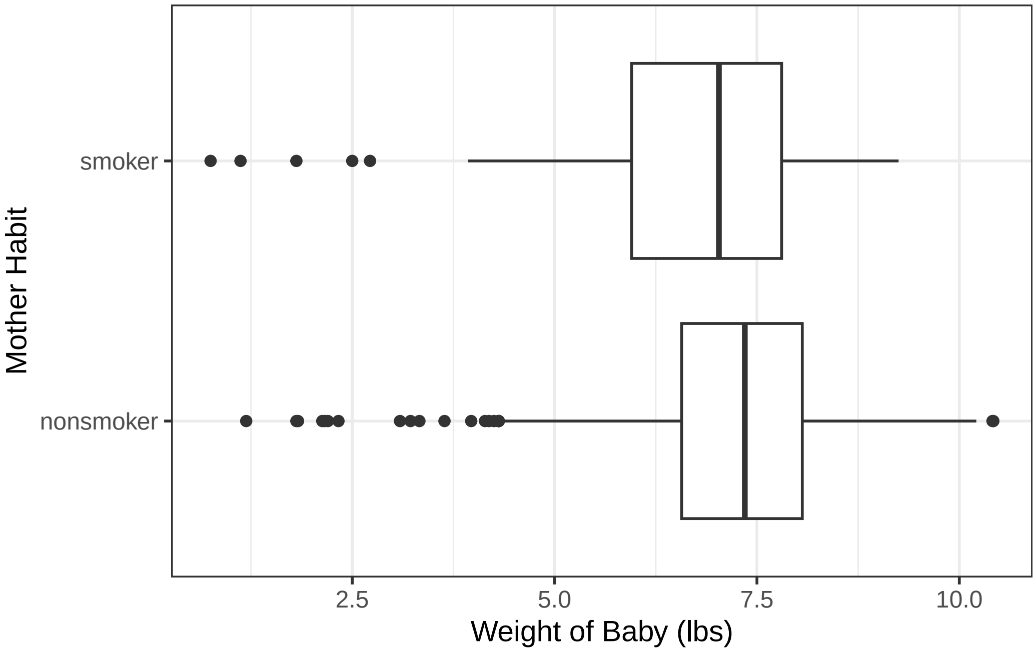
v. Suppose that the observed standard deviations had turned out to be: $s_{\text{Tai Chi}} = 5.8$ and $s_{\text{Stretching}} = 7.2$. What would be the primary change to the 95% confidence interval be compared to the original of (2.73, 6.59)?

- The middle of the interval would be a smaller number
- The middle of the interval would be a bigger number
- The width of the interval would be a bigger number
- The width of the interval would be a smaller number
- The interval would remain unchanged

6. In 2004, the state of North Carolina released to the public a large dataset containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This analysis will focus on a random sample of 1,000 observations from the published dataset.

```
      habit  min     Q1 median    Q3   max     mean       sd   n missing
1 nonsmoker 1.19 6.5700   7.35 8.060 10.42 7.269873 1.232846 867       0
2    smoker 0.75 5.9525   7.03 7.805  9.25 6.677193 1.596645 114       0
```





i. Hospital administration at Duke University Hospital are interested in the difference in the mean baby birth weight between mothers who do not smoke and mothers who do smoke. Using the summary statistics above, report the observed statistic for this comparison. Indicate in your answer what notation should be used for this statistic.

ii. These hospital administrators are interested in estimating the true difference in mean birth weight between mothers who do not smoke and mothers who do smoke. The administrators learned in their Statistics class how to obtain a confidence interval for a difference in means using a *t*-distribution. Using the plots above and your knowledge of how the data were collected evaluate whether it would be appropriate for the administrators to use a *t*-distribution to obtain a confidence interval for the true difference in means.

iii. Select the correct t-quantile from below and construct a 99% confidence interval.

`qt(0.95, 959)` = 1.6464441

`qt(0.975, 959)` = 1.9624407

`qt(0.99, 959)` = 2.3302426

`qt(0.995, 959)` = 2.5809657

iv. Interpret the 99% confidence interval in the context of this study.

v. Based on your confidence interval above, which of the following is the most likely p-value for a two-sided hypothesis test? Circle one.

- 0.20
- 0.10
- 0.05
- 0.005

7. Which of the following analyses is most appropriate for the following situation? A teacher compares the pre-test and post-test scores of a group of students to test whether their scores have increased throughout the semester.

   a. A paired t-test for dependent samples
   b. A two-sample t-test for independent samples

8. Which of the following analyses is appropriate for the following situation? A random sample of 100 male students and a random sample of 1000 female students are asked how many Cal Poly Soccer games they attended this past season. The goal is to test whether male Cal Poly students attend more football games, on average, than the female Cal Poly students.

   a. A paired t-test for dependent samples
   b. A two-sample t-test for independent samples

9. Which of the following analyses is appropriate for the following situation? Ten diamond rings are shown to two jewelers, and each jeweler is asked to determine the appraisal value for each ring. The goal is to test whether the two jewelers give different appraisal values, on average.

   a. A paired t-test for dependent samples
   b. A two-sample t-test for independent samples

10. Suppose that in order to investigate the effectiveness of a new blood pressure drug, subjects are randomly assigned to either the placebo group or the treatment group. All subjects have their blood pressure measured at baseline. Those in the placebo group then receive an inactive pill for 2 months, and those in the treatment group receive the new drug for 2 months. At the end of the two months, researchers once again measure each subject's blood pressure and compute the change in blood pressure for each subject.

Research Question Does the new drug improsve the blood pressure for all adults?

i. Sketch out what this data might look like (*Hint: Look back at Example 7.3*).

ii. Write out the parameter of interest for the study.

iii. In symbols, set up the null and alternative hypothesis.

iv. Check the conditions necessary to conduct an analysis using the t-distribution.

v. Suppose we obtain a t-statistic of 0.92 and a p-value of 0.17. What would your conclusion be in context of the research question?

vi. What is the interpretation of the p-value above in context of the problem?

vii. What do you want to know about the subjects in order to generalize your results to all adults?

viii. Suppose we had found *evidence* in support of the research question. Could we say that the new drug *causes* an improvement in blood pressure?

11. In a recent study conducted in the United Kingdom, researchers gathered data on 6,705 children to investigate the potential relationship between a mother's exposure to cats during pregnancy and the occurrence of psychotic episodes in their children. The study included two groups: one consisting of 4,746 children whose mothers did not have cats while pregnant and another group of 1,959 children whose mothers did have cats during pregnancy.

Among the group of 4,746 children with no maternal cat exposure, 536 children experienced one or more psychotic episodes during the course of the study. In contrast, among the 1,959 children whose mothers had cats during pregnancy, 240 children had one or more psychotic episodes.

Research Question: Does the psychotic episode rate differ between children who's moms did have cats while pregnant and those who's moms did not?

(i) Create a contingency table of counts based on the data obtained in this study.

(ii) Find the observed proportion of children who's moms did not have cats while pregnant that had one or more psychotic episode.

(iii) Find the observed proportion of children who's moms did have cats while pregnant that had one or more psychotic episode.

The following output was obtained from a chi-square test to investigate this question.

```
library(infer)
chisq_test(x = cats,
           response = psychotic,
           explanatory = cats)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      1.15        1   0.284
```

(v) Write a solution and make sure to include the chi-square test statistic, degrees of freedom, the p-value, and a conclusion written in everyday language.

12. Ten mice (6–8 weeks old) were randomly assigned to one of two groups; five were exposed to simulated environmental tobacco smoke for 6 h/day, 5 days/week for 5 months. The other 5 mice were kept in clean air during this time period. Then, all of the mice were allowed to recover for a further 4 months in filtered air before being killed for analysis of lung tumor incidence. The results are shown below.

	Tumor	No Tumor	Total
Treated	4	1	5
Control	2	3	5
Total	6	4	10

Research question: Does the proportion of mice that develop a lung tumor differ between those exposed to tobacco smoke and the control group?

(i) Convert the Research Question into H_0 and H_A.

(iii). What proportions would you compare to answer the research question?

(ii). Would it be appropriate to use the chi-square distribution to test the hypotheses? Explain.