Review Guide for Midterm 1 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.
- The exam is mostly multiple choice, but will have a couple of short answer questions mixed in.

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 1 - 4.

- Identifying pieces of a data set: observation, variable, sample size, etc.
- Parameter vs Statistic
- What impacts strength of evidence? How does this relate to the size of the p-value?
- How do you interpret the p-value? Can you do this for various scenarios?
- Confidence intervals for a single proportion.
  - All confidence levels.
  - How does this relate to hypotheses testing?
- Work through hypotheses testing for the following types of scenarios:
  - One categorical variable with two outcomes (single proportion)
  - One categorical variable with multiple categories (chi-square goodness of fit)
  - Two categorical variables with multiple categories (chi-square test)

Practice Problems

1. Suppose a governor is concerned about their "negatives" (i.e., the percentage of state residents who express disapproval with their job performance). Their campaign pays for a series of television ads, hoping that they can keep the negatives below 30%. They use follow-up polling to assess the ads' effectiveness. Their hypotheses are as follows:

H_0 : The ads are not effective

H_A : The ads are effective (i.e., the negatives are below 30%)

The negatives come in at 22%, and the p-value obtained is 0.08.

(i) Which of the following is the most correctly written conclusion given a significance level of 0.05?

a. We have evidence that the ads are effective.
b. We have evidence that the ads are not effective.
c. There is not enough evidence to conclude that the ads are effective.
d. We have evidence that 8% of the state residents disapprove of the governor.

(ii) Which of the following is the most correctly written conclusion given a significance level of 0.10?

a. We have evidence that the ads are effective.
b. We have evidence that the ads are not effective.
c. There is not enough evidence to conclude that the ads are effective.
d. We have evidence that 8% of the state residents disapprove of the governor.

(iii) Recall that the p-value was 0.08. Which of the following is the most correct interpretation of this p-value?

a. There is an 8% chance that the ads were effective.
b. There is an 8% chance that the ads were not effective.
c. There is an 8% chance that a state resident disapproves of the governor.
d. There is an 8% chance that the poll could have resulted in 28% or fewer negatives even if the ads were not effective.

(iv) Suppose instead, our alternative is to test whether the negatives are different from 30%. What would you expect your p-value to be?

a. 0.08
b. 0.16
c. 0.04
d. 0.22

2. 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.

ii. State the Null and Alternative Hypotheses in symbols.

iii. Carry out a simulation study to test the hypotheses. What are the inputs for the single proportion applet?

iv. *Interpret* the p-value. Note this is not the conclusion or asking you to make a decision, but to interpret.

v. State a conclusion in context of the research question.

3. A random sample of 200 students is surveyed on a university campus. They are asked if they use a laptop in class to take notes. Suppose that based on the survey, 70 of the 200 students responded "yes." Your goal is to estimate the true proportion of all students on this campus who use a laptop to take notes.

(i) Find the observed statistic (also called the point estimate).

(ii) Find the margin of error associated with the 95% confidence interval.

(iii) Construct a 95% confidence interval for the true proportion of all students on this campus who use a laptop to take notes.

(iv) Suppose, instead, the survey results found that 175 out of 500 students responded "yes." How will your confidence interval change?

a. The center of the confidence interval will increase.
b. The center of the confidence interval will decrease.
c. The confidence interval will be wider.
d. The confidence interval will be narrower.

3. Suppose that 500 Cal Poly students were surveyed to find a 95% confidence interval for the proportion of all Cal Poly students who have cheated on a test. Answer the following questions by circling your response.

(i) TRUE / FALSE: The population proportion (i.e., the proportion of all Cal Poly students who have cheated on a test) will definitely be contained in the confidence interval.

(ii) TRUE / FALSE: The sample proportion (i.e., the sample statistic) will definitely be contained in this confidence interval.

(iii) TRUE / FALSE: This confidence interval can be used only to describe the 500 students who were surveyed; it cannot be used to make a statement about all Cal Poly students.

(iv) TRUE / FALSE: If a 90% confidence interval were constructed instead of a 95% confidence interval, the margin of error would be smaller and the interval would be narrower.

4. Suppose a 95% confidence interval constructed from a sample mean is (5.5, 10.5). Circle ALL true statement(s).

   a. A 99% confidence interval constructed from the same sample mean will contain 8.
   b. A 99% confidence interval constructed from the same sample mean will NOT contain 8.
   c. A 90% confidence interval constructed from the same sample mean will NOT contain 8.
   d. A 90% confidence interval constructed from the same sample mean will contain 8.

5. Suppose that two different surveys were conducted to investigate whether the majority of Minnesotans feel there is a shortage of affordable home options. In Survey A, 344 of 625 (55%) felt there was a shortage of affordable home options. In Survey B, 331 of 625 (53%) said they felt this way. Which survey (A or B) would result in a smaller p-value when testing whether the majority of Minnesotans feel there is a shortage of affordable home options? (3 pts)

   a. The p-values would be the same since both surveys sampled the same number of subjects.
   b. There is not enough information given to determine anything about the p-values.
   c. Survey A
   d. Survey B

6. A local doctor suspects that there is a seasonal trend in the occurrence of the common cold. She estimates that 40% of the cases each year occur in the winter, 40% in the spring, 10% in the summer and 10% in the fall. A random sample of 1000 patient cases was collected, and the number of cold cases for each season was recorded.

A summary table of the observed counts is included below:

Fall	Spring	Summer	Winter	Total
165	292	169	374	835

(i) If the doctor's suspicion was correct, what *proportions* would you expect for each cell? Insert the corresponding values in each cell.

Fall	Spring	Summer	Winter
$\pi_{fall} =$	$\pi_{spring} =$	π_{summer}	$\pi_{winter} =$

(ii) If the table above represents what is assumed to be true under H_0, state the alternative hypothesis using words (what would the null be in words?).

(iii) Compute the table of expected counts.

Fall	Spring	Summer	Winter

(iv) What is the summer cold cell's contribution to the X^2 test statistic?

(v) Evaluate whether the conditions required to use the chi-square distribution to obtain a p-value are violated.

(vi) A X^2 statistic of 124 was obtained for these data. Fill in the R code below to conduct the chi-square goodness of fit test.

```
chisq_test(x = cold_data,
           response = _________,
           p = c("fall" = ______,
                 "spring" = ______,
                 "summer" = ______,
                 "winter" = ______
           )
)
```

(vii) A p-value of <0.001 was obtained using the code you input above. Based on this p-value what would you conclude about the Doctor's hypothesis regarding the distribution of colds throughout the year?

- How would you interpret this p-value?
- What are the test statistics and associated degrees of freedom?

7. 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.

8. 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.