Review Guide for Midterm 2 Exam – Thursday, November 6, 2025

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.
- I *will* provide you with a table of scenarios (see Canvas).
- 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 3 - 6 (and concepts that are necessary to build off of from the first two chapters).

- 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?):
  - One categorical variable with more than two categories (Chi-square Goodness of Fit Test)

  - Two categorical variables (Chi-square Test of Independence Test)
  - One numerical variable (t-test for a single mean)

- How do you interpret the p-value? Can you do this for various scenarios?
- Descriptive statistics for one numerical variable

  - Can you describe a distribution for a numerical variable? (shape, center, spread, outliers) – histogram, dotplot, boxplot.
  - Identify summary statistics (with symbols) from `favstats` output – mean, sd, Q1, Q3, median, etc.

- Understand the concept of a sampling distribution for

  - chi-square
  - means (CLT)

- Confidence intervals for a mean

  - 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)

Practice Problems

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

1. 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 *distribution of sample means*) 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.

iv. Can you generalize the results of your study to all snickers bars? Explain.

2. 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 835 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
$\pi_{fall} =$	$\pi_{spring} =$	$\pi_{summer} =$	$\pi_{winter} =$

(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( ______, ______, ______, ______)
)
```

(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?

3. 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
```

(iv) 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.

(v) Is this an observational study or designed experiment? Why?

5. 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.

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

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

(iv) Is this an observational study or designed experiment? Why?

6. In 2020, a groundbreaking pesticide known as "EcoShield" emerged within the realm of agricultural and horticulture sciences. This new pesticide sparked an investigation into its potential effects on the average lifespan of fruit flies, a species extensively studied in life sciences for its short life cycle and genetic correlations.

In the past, the average lifespan of typical fruit flies was recorded as 40 days. To investigate this, a study was conducted in 2022[1], where a random sample of typical fruit flies was exposed to the "EcoShield" pesticide.

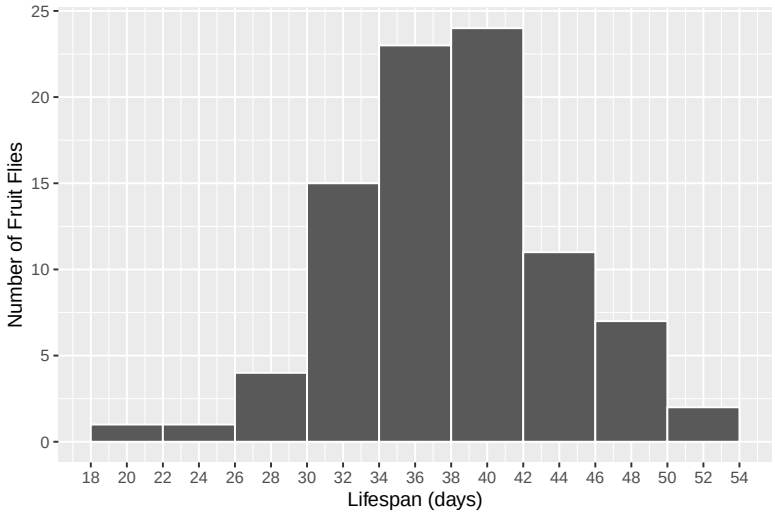
[1]Disclaimer: This is a made up study and data set and does not necessarily reflect the effects of EcoShield.

Research Question: Does exposure to a new pesticide, "EcoShield," reduce the average lifespan of fruit flies compared to their traditional average lifespan?

The observed study data, summary statistics, and histogram are given below:

```
# A tibble: 6 x 4
  fly_id exposure  environment lifespan
   <int> <chr>     <chr>          <dbl>
1      1 EcoShield Plant             39
2      2 EcoShield Water             49
3      3 EcoShield Water             36
4      4 EcoShield Water             42
5      5 EcoShield Food              34
6      6 EcoShield Plant             37
```

```
 min Q1 median Q3 max     mean       sd  n missing
  21 35   38.5 42  52 38.78409 5.825921 88       0
```



i. Identify the variable of interest and its data type:

categorical / numerical

ii. In context of the problem, state the parameter of interest.

iii. Identify the following values from above and state what each symbol represents.

$\bar{x}$ = _____, is the _____.

s = _____, is the _____.

n = _____, is the _____.

iv. Set up the null and alternative hypotheses, **in symbols only**

v. Check the normality assumption, is it reasonable to use the t-test? Explain.

vi. Calculate the t-test statistic. Show your work!

Below, I have provided the correct t-distribution curve to test the hypothesis.

vii. How many degrees of freedom does the t-distribution have?

viii. On the t-distribution curve above, show how you would estimate the p-value to test the hypotheses. *Make sure to show all parts!*

ix. Testing the hypothesis results in a p-value of 0.11. Write a conclusion in context of the study.

x. Calculating the 95% confidence interval, we get (37.92, 40.48). Where does the 40 fall in relation to the 95% confidence interval? Why does this make sense?