Final Exam Question Bank

Stat 218

Your exam questions will be randomly selected from this bank of questions. There **will not** be a solution key posted. It is your responsibility to discuss your ideas with your group members and / or with Dr. Theobold during office hours prior to the exam.

Golden Ticket

Scenario	One Categorical Response	Two Categorical Variables	One Quantitative Response
Type of plot	Bar plot	Side-by-Side Bar plot, Stacked Bar plot, Mosaic plot	Dot plot, Histogram, Boxplot
Summary measure	Proportion	Difference in Proportions	Mean
Parameter notation	π	$\pi_1 - \pi_2$	μ
Statistic notation	$\widehat{p}$	$\widehat{p}_1 - \widehat{p}_2$	$\bar{x}$

Provided Formulas

$$IQR = Q3 - Q1$$

1.5 IQR Rule: above $Q3 + (1.5 \times IQR)$ or below $Q1 - (1.5 \times IQR)$

$$\hat{y} = b_0 + b_1 \times x$$

$$Residual = y - \hat{y}$$

t-based confidence interval: point estimate $\pm\ t^*_{df} \times SE$

$$SE(\mu) = \frac{\sigma}{\sqrt{n}}$$

$$SE(\mu_1 - \mu_2) = \sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}$$

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Final Exam Question Bank

Wild Mushrooms

Q1[17 points] 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

(a)[4 pts] 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) _____.

(b)[3 points] Calculate the proportion of mushrooms with a broad gill size that are poisonous. *Give appropriate notation with informative subscripts if necessary.*

Work:

Answer: _____

Notation: _____

(c)[3 points] Calculate the proportion of mushrooms with a narrow gill size that are poisonous. *Give appropriate notation with informative subscripts if necessary.*

Work:

Answer: _____

Notation: _____

(d)[2 points] Using your answers to (b) and (c), fill in the correct names next to each color, to label the bar chart showing the relationship between gill size (broad or narrow) and whether the mushroom is edible.
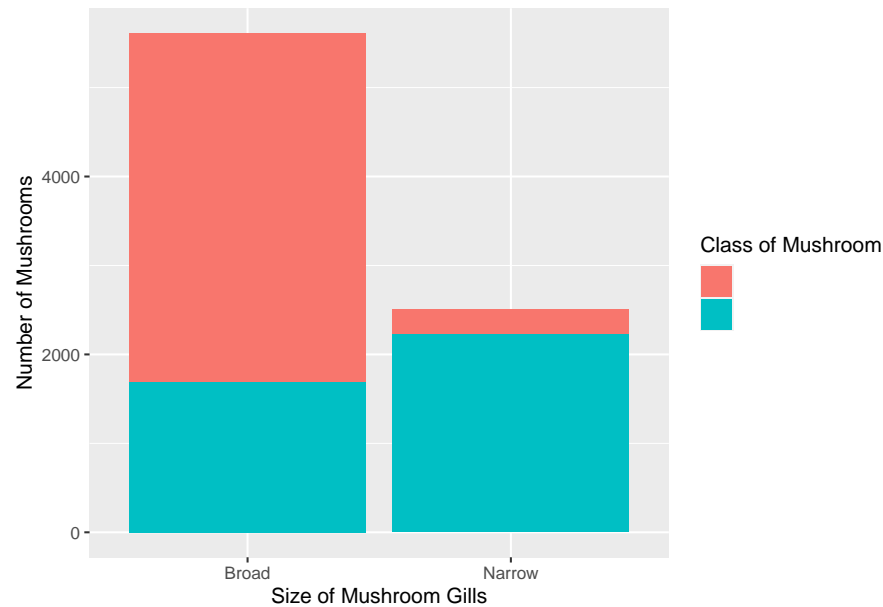
(d)[3 points] Based on the plot, describe the relationship between a mushrooms gill size and whether it is edible or not.

(e)[2 points] Suppose the Chi-Squared test resulted in a "significant" p-value. Which of the following would be the correct scope of inference for this study?

(i) It can be inferred for all mushrooms that gill size causes a mushroom to be poisonous.

(ii) It can be inferred for all mushrooms that gill size is associated with whether a mushroom is poisonous.

(iii) It can be inferred for this sample of mushrooms that gill size causes a mushroom to be poisonous.

(iv) It can be inferred for this sample of mushrooms that gill size is associated with whether a mushroom is poisonous.

Mendelian Genetics

Q2 [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 easily 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.
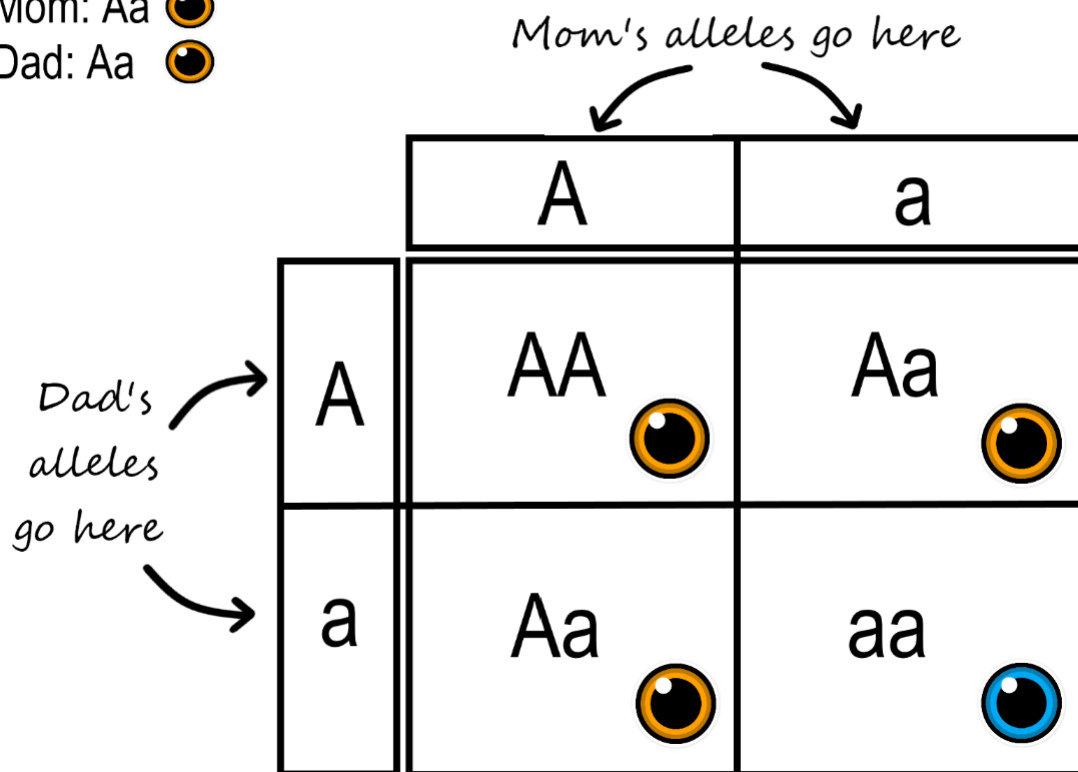
(a) [2 points] When Mendel crossed his pea plants, he learned that tall (T) was dominant to short (t). Suppose in your Biology course you carried out an experiment to test if the plot offspring would follow Mendelian inheritance.

Fill in the table defining what the expected tallness inheritance for your plants should be.

	T	t
T		
t		

(b) [3 points] 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
$p_{TT} =$	$p_{Tt} =$	$p_{tt} =$

(c) [2 points] 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
$p_{Tall} =$	$p_{Short} =$

(b) [3 points] If the table above represents what is assumed to be true about tallness inheritance under H_0, state the alternative hypothesis using either words or notation.

(c) [1 point] 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.

Create a table summarizing these observations.

(d) [4 points] 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**:

(e) [3 points] 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?

Seasonal Colds

Q3 [22 points] 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 is included below:

Season	Number of Patients
Fall	165
Spring	292
Summer	169
Winter	374

(a) [4 points] 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
$p_{\text{fall}} =$	$p_{\text{spring}} =$	$p_{\text{summer}} =$	$p_{\text{winter}} =$

7

(b) [3 points] If the table above represents what is assumed to be true under H_0, state the alternative hypothesis using either words or notation.

(c) [4 points] Compute the table of expected counts.

Fall	Spring	Summer	Winter

(d) [2 points] What is the summer cold cell's contribution to the X^2 statistic?

(e) [2 points] Evaluate whether the conditions required to use the χ^2 distribution to obtain a p-value are violated.

(f) [3 points] An X^2 statistic of 124 was obtained for these data. Fill in the R code below to find the p-value for this statistic. *Hint:* The `lower.tail` input takes one of two options `TRUE` or `FALSE`.

```
pchisq(_____, df = _____, lower.tail = _____)
```

(g) [4 points] Using the code you input above, a p-value less than 0.000000123 was obtained. Based on this p-value what would you conclude about the Doctor's hypothesis regarding the distribution of colds throughout the year?

Professor Attractiveness

Q4 [21 points] Data were scraped from *ratemyprofessors.com*, a website which allows college and university students to assign ratings to professors and campuses of American, Canadian, and United Kingdom institutions. In order for the rating to be posted, a rater must rate the professor in the following categories: `overall quality`, `level of difficulty`, and their `hotness`.

The Dean of the College of Science and Math would like to know if the difficulty rating of female professor's at Cal Poly is related to how attractive they are perceived to be.

A table of observed counts in shown below.

Difficulty	Attractive	Unattractive	Total
Easy	76	54	130
Hard	45	170	215
Medium	51	90	141
Total	172	314	486

(a)[3 points] What type of test should be performed with these data to address the research question?

(b) [4 points] Write out the null and alternative hypotheses for the test you stated in **(a)**.

H_0 :

H_A :

(c) [3 points] When analyzing the data, the administrators chose to use a χ^2 distribution to obtain their p-value. Was this an appropriate choice? Why or why not?

(d) [2 points] What distribution should be used to find the p-value for the hypothesis test above?

(e) [4 points] A p-value less than 0.000000142 (X^2 statistic $= 50$) was found using the distribution stated in **(d)**. Based on the p-value obtained, report your conclusion to the hypothesis test **in the context of the problem.**

(f) [2 points] Compute the relative risk of being labeled "Hard" for "attractive" female professors compared to "unattractive" female professors.

(g) [3 points] Interpret the relative risk you obtained in **(f)** in the context of these data. Be sure to indicate how the comparison was made and what the sign indicates!

Guilty Decisions

Q5[11 points] The effect of guilt on how a decision maker focuses on a problem was investigated in the *Journal of Behavioral Decision Making* (January 2007). A total of 155 volunteer students participated in the study, where each was randomly assigned to one of three emotional states (guilt, anger, or neutral) through a reading / writing task. Immediately after the task, the students were presented with a decision problem (e.g., whether or not to spend money on repairing a very old car). The researchers found that a higher proportion of students in the guilty-state group chose not to repair the car than those in the neutral-state and anger-state groups.

(a)[3 pts] What is the study design? Explain your reasoning. Select one.

  (i) Observational study. The researchers did not take a random sample of students.

  (ii) Observational study. There is no random assignment of students to emotional state.

  (iii) Experiment. The students are a representative sample of all students.

  (iv) Experiment. The students were randomly assigned to emotional state.

(b)[3 pts] Which types of sampling bias may be present in this study? Select all that are present, or if you believe there is no bias present, select option 4) No bias.

  (i) Selection bias

  (ii) Non-response bias

  (iii) Response bias

  (iv) No bias

(c)[3 pts] The researchers found a difference in the proportion of students in the guilty-state group chose not to repair the car than those in the neutral-state and anger-state groups. Can we conclude that the emotional state group caused a higher proportion of students to not repair the car? Select one.

  (i) Yes, because these data are from a representative sample.

  (ii) No, because the students are volunteers.

  (iii) Yes, because the researchers evened out confounding variables across emotional state group.

  (iv) No, because the sample size is not large enough.

(d) [2 pts] Which type of plot would be the **most** appropriate to display the relationship between each level of emotional state and whether students choose to repair the car? Select all that apply.

  (i) Scatterplot

  (ii) Filled bar plot

  (iii) Pie chart

  (iv) Side-by-side boxplot

Distribution of Fish on Blackfoot River

Q6 [19 points] We have data on fish caught in the Blackfoot River near Helena, Montana by Montana Fish, Wildlife, & Parks personnel over a number of years. They used electrofishing equipment to attract the fish to the boat, then dipped them out of the water with nets, measured length in cm and weight in grams. They are often working in cold conditions in late autumn or early spring, so some measurement error is expected.

These data are not from a random sample. The goal is to catch all fish within a section of the Blackfoot River every few years to assess the health of the population. Montana Fish, Wildlife, & Parks is interested in analyzing if the prevalence of the different species of fish has stayed relative stable over the years.

```
## # A tibble: 10 x 5
##    length weight year  section      species
##     <dbl>  <dbl> <fct> <chr>        <chr>
## 1     540   1200 2004  ScottyBrown  Brown Trout
## 2     335    420 2002  ScottyBrown  Brown Trout
## 3     232    125 2002  ScottyBrown  Brown Trout
## 4     208     90 2000  Johnsrud     Rainbow Trout
## 5     215    115 2000  Johnsrud     Rainbow Trout
## 6     270    210 2000  ScottyBrown  Brown Trout
## 7     170     50 2006  Johnsrud     Rainbow Trout
## 8     230    120 2000  Johnsrud     Rainbow Trout
## 9     352    460 2000  Johnsrud     Whitefish
## 10    146     40 2006  Johnsrud     Rainbow Trout
```

(a) [2 points] Based on the output above, what is the observational unit for this study?

(b) [3 points] Based on the output above, what type of variable is `year`? Given the stated analysis, is this the correct data type for this variable?

(c) [3 points] Based on the bar plot below, describe the relationship between the sampling year and the species of captured fish. Make direct reference to characteristics of the plot!
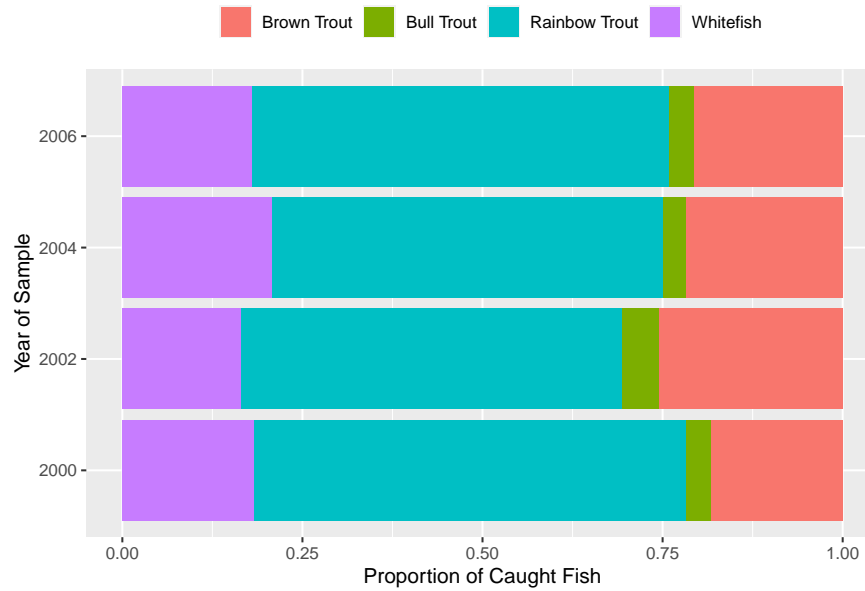

(d) [5 points] The statistician at Montana Fish, Wildlife, & Parks prefers using simulation-based methods rather than theory-based methods. They used computer simulation to obtain the null distribution shown on the next page.

Fill in the steps necessary to obtain out **one** statistic on the null distribution.

On (#) _____ cards, write _____ on the cards.

Generate a new sample that could have happened if the null hypothesis was true by:

Calculate and plot the _____ from each permutation / computer simulation.
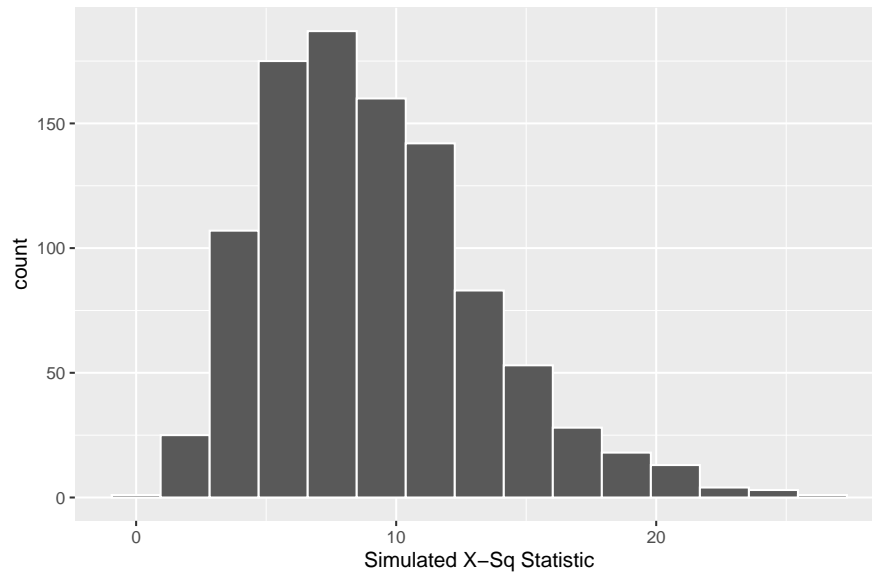
(e) [2 points] The observed X^2 statistic was 57.4155167. Use this statistic to draw a line and shade the direction that should be used when calculating the p-value.

(f) [1 point] Approximate the p-value for this hypothesis test.

(g) [3 points] Using the conditions of the simulation-based method used, evaluate if you believe the p-value you obtained in **(f)** is accurate.

General Concepts

Q7[2 points] Suppose you reject the null hypothesis at the 0.05 level of significance. A colleague had planned to use a 0.01 level of significance instead. Will your colleague also reject the null hypothesis?

(a) Yes

(b) No

(c) Maybe

(d) Changing a level of significance cannot affect decisions.

Q8[2 points] In hypothesis testing,

(a) a type II error occurs when you fail to reject a false null hypothesis.

(b) you do not need to decide on a level of significance before you find a p-value; you can adjust based on the p-value you observe.

(c) one of the main sources of type I errors is that the sample size was not large enough.

(d) p-values are the probability that the null hypothesis is false.

Q9 For each of the following, select the single most appropriate analysis for the situation described. You may use an analysis for more than one situation. (2 pts each)

Chi-Square Test of Independence   One-Way ANOVA
Simple Linear Regression   Chi-Square Test of Homogeneity
Chi-Squared Goodness-of-Fit test   Paired t-test
Confidence interval for μ Hypothesis test for μ
Hypothesis test for $\mu_1 - \mu_2$

(a) 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.

_____

(b) A study of honeybees looked at whether the proportions of different honeybee species varies 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.

_____

(c) 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.

_____

(d) Researchers are interested in determining if the yield of a tomato plant differs among three tomato varieties.

_____

(e) 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. You, a logical person, know that someone's height drastically affects how long it takes them to walk places. What is the best **method** to answer his research question? Circle one.

_____

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

_____