STAT 218 – Final Exam

Dr. Allison Theobold

June 4, 2020

Name: _____

Section (circle one): 8:10am 9:10am

Read and Sign the Following Statement:

I understand that give or receiving help on this exam is a violation of academic regulations and is punishable by a grade of **F** in this course. This includes looking at other students' exams and / or allowing other students, actively or passively, to see answers on my exam. This also includes revealing, actively or passively, any information about the exam to any member of Professor Theobold's STAT 218 class who has not yet taken the exam. The use of cell phones is strictly prohibited.

I pledge not to do any of these things.

Signed: _____

Instructions.

- Read and sign the honesty pledge at the top of this page. Your exam will not be graded unless the honesty pledge is signed!
- You may use a calculator. You **may not** use your phone or any device that connects to the internet as a calculator.
- Show all work as clearly as possible. Point totals are shown in brackets next to each part. Formulas without values entered do not count as work.
- All answers should be reported in decimal form, rounded to three decimal places.
- For multiple choice and multi-select problems, completely fill in the provided circle (multiple choice) or square (multi-select) for your desired answer choice(s). If you change an answer, be sure to completely erase your initial selection.
- You have 2 hours and 50 minutes to complete this exam, which is ample time! If you get stuck on a problem, take a deep breath, say something positive about yourself, and write down what you know.

Golden Ticket

Scenario	One Categorical Response	Two Categorical Variables	One Quantitative Response	Two Quantitative Variables	Quant. Response and Categ. Explanatory
Type of plot	Bar plot	Dodged Bar plot, Stacked Bar plot, Filled Bar plot	Dot plot, Histogram, Boxplot	Scatterplot	Faceted Histograms, Side-by-side Boxplots
Summary measure	Proportion	Difference in Proportions	Mean	Slope or Correlation	Difference in Means
Parameter notation	π	$\pi_1 - \pi_2$	μ	Slope: β_1; Correlation: ρ	$\mu_1 - \mu_2$
Statistic notation	$\widehat{p}$	$\widehat{p}_1 - \widehat{p}_2$	$\bar{x}$	Slope: b_1; Correlation: r	$\bar{x}_1 - \bar{x}_2$
Statistical Method(s)	χ^2 Goodness of fit Test	χ^2 Test of Independence, χ^2 Test of Homogeneity, Permutation Test	t-test for One Mean, t-test for Paired Differences, Bootstrap Confidence Interval for One Mean	t-test for β_1, Permutation Test for β_1, Bootstrap Confidence Interval for β_1	t-test for $\mu_1 - \mu_2$, Permutation Test for $\mu_1 - \mu_2$, Bootstrap Confidence Interval for $\mu_1 - \mu_2$

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 $\pm\, t^*_{df} \times SE$

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

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

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

Midterm 1 Question Bank

Q9 [1 point] When you change from a 90% to a 95% confidence interval,which part(s) of the confidence interval change? (Select all that apply)

(a) Statistic (midpoint)

(b) Multiplier

(c) Standard error

Q5 [2 points] Indicate whether each statement about a bootstrap resample is **TRUE** or **FALSE**.

(a) The bootstrap resample and original sample **must** be the same size. _____

(b) The bootstrap resample and original sample are **both** taken from the population. _____

(c) The bootstrap resample can **only** use values that were in the original sample. _____

(d) The bootstrap resample uses **all** of the values that were in the original sample. _____

Q6 [2 points] The purpose of creating a null distribution is to: (Select all that apply)

(a) Discover what statistics might have occurred if the null hypothesis was true.

(b) Approximate the sampling distribution under H_0.

(c) To determine if the null hypothesis is true.

(d) To determine if the observed statistic is unlikely if the null was true.

Q4 [4 points] Researchers are interested in the fish that reside in the Caspian Sea. They have plans to collect many fish and take multiple measurements on each. Match each statistical description on the right with each piece of information given. Put the letter of the statistical description in the blanks on the left.

_____ circumference of the fish	(a) quantitative variable
_____ species of the fish	(b) categorical variable
_____ average length of all fish in the area of consideration	(c) parameter: μ
_____ mean internal temperature of the fish collected in the sample	(d) statistic: $\bar{x}$
_____ one of the fish in the area of consideration	(e) observational unit
_____ method of only studying the fish caught in the net 3pm on Wednesday of the research time frame	(f) cluster sampling method
_____ method of selecting 5% of each species, known to be in the area of consideration, for the sample	(g) stratified sampling method
____ method of dividing up the whole location with netting and sampling 10 random netted areas	(h) convenience sampling method

Q2[21 points] I collected data on 512 different fast food items from Mcdonalds, Chick-Fil-A, Sonic, Arby's, Burger King, Dairy Queen, Subway, and Taco Bell. For each restaurant, I sampled 64 items from their menu and recorded the nutritional content of each item (e.g., calories, saturated fat, calcium, protein, etc.).

(a) [2 pts] Describe the sampling method I used to obtain these 512 fastfood items.

(b) [3 pts] I am interested in studying the linear relationship between the total calories of a food item and the amount of saturated fat that item contains.

Write the null hypothesis for my question of interest, using both words and notation.

(c) [1 pts] Is the alternative hypothesis one- or two-sided? Select one.

- One-sided
- Two-sided

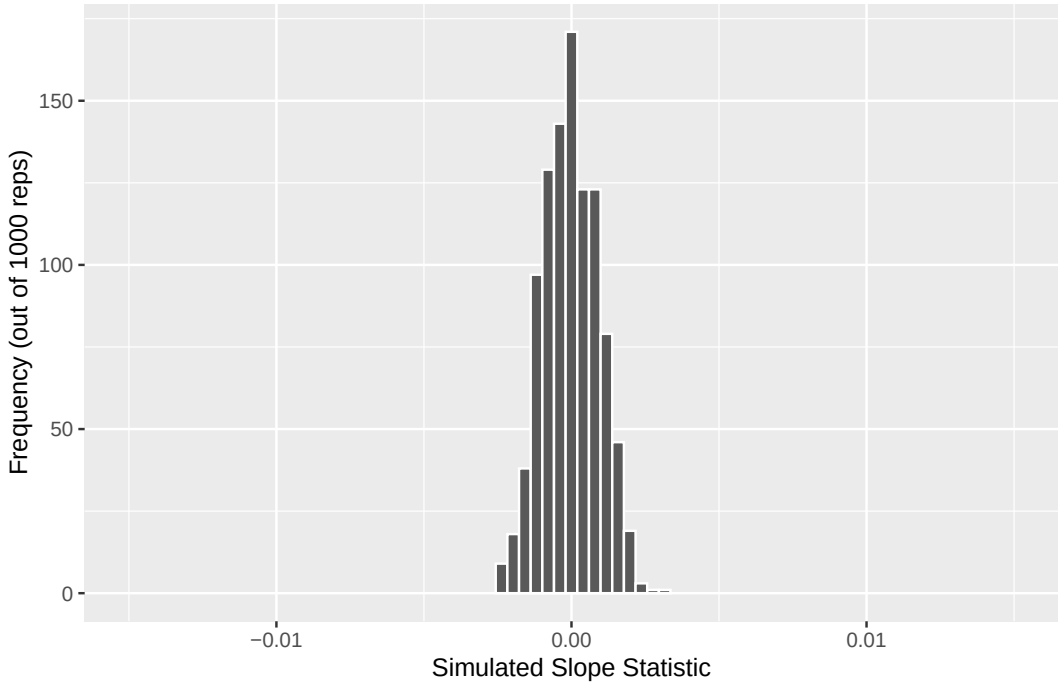
(d) [5 pts] Below is a simulated null distribution from `R`. Fill in the blanks below with one answer in each set of parentheses to correctly explain how one sample on the null distribution was be created. Blanks preceded by (#) should be filled in with a number, all other blanks should be filled in with the context of the study.

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

Assume the null hypothesis is true and _____.

Generate a new sample of 512 ordered pairs by _____.

Calculate and plot the _____ from each simulated sample.



(e) [2 pts] Using the regression output below, draw a vertical line where the observed statistic falls on the null distribution (on the previous page).

term	estimate	std_error
intercept	-0.771	0.406
calories	0.017	0.001

(f) [2 pts] Shade the location of the plot you would use to calculate the p-value.

(g) [1 pts] Estimate the p-value associated with this hypothesis test.

(h) [3 pts] Which of the following is a correct interpretation of the p-value obtained? (Circle one)

- In less than 1 out of 1000 simulated samples, we would observe a sample slope of 0.017 or more extreme, if there is no linear relationship between the total calories and the saturated fat of a fast food item.
- If there is a linear relationship between the total calories and the saturated fat of a fast food, we would observe a sample slope of 0.017 or more extreme with a probability of less than 1 out of 1000.
- The probability of seeing a sample slope between the total calories and the saturated fat of a fast food item of 0.017 or more extreme is less than 0.1
- The probability that there is no linear relationship between the total calories and the saturated fat of a fast food item, is less than 0.1

(i) [2 points] Given the p-value for the hypothesis test, would the 95% confidence interval for β_1 contain 0?

Midterm 2 Question Bank

Q1[8 points] Researchers in Southern England collected data on grassland butterflies. They were interested in whether movement patterns varied across species and between male and female butterflies. Researchers observed 164 butterflies over the three-year length of this study, of which 28 were female and 136 were male. These 164 butterflies were considered to be representative of all grassland butterflies. The butterfly movements were observed by measuring how far they flew (in meters) from one landing site to the next, called step distance. This was done by placing a flag at each landing site and measuring the distance between the flags using a mapping software.

The researchers were interested in investigating if there was a difference between how far male and female butterflies travel, on average, between landing sites.

(a)[4 pts] Fill in each blank with one of the options in parentheses to best describe the variables collected.

Step distance is the (explanatory/response) _____ and it is (categorical/quantitative) _____.

Sex is the (explanatory/response) _____ and it is (categorical/quantitative) _____.

(b)[2 pts] Which visualizations would be the **most** appropriate to display the relationship between step distance and sex of the butterfly? Select all that apply.

(i) Segmented Bar plot

(ii) Scatterplot

(iii) Side-by-side Boxplot

(iv) Faceted histograms

(c) [2 pts] Assuming a statistical difference in step distance is found between the male and female butterflies in the sample, what is the scope of inference for this study? Select one.

(i) Sex causes a difference in average step distance for all grassland butterflies.

(ii) Sex is associated with a difference in average step distance for the sample of grassland butterflies.

(iii) Sex causes a difference in average step distance for the sample of grassland butterflies.

(iv) Sex is associated with a difference in average step distance for all grassland butterflies.

Q4[21 points] As you may be aware, many individuals are concerned about the presence of BPA in plastics, especially plastics that make contact with food and drinks. Currently, these is an incomplete understanding of how exposure to BPA affects our ingestion. Last year Dr. Hagobian in the Kinesiology and Public Health carried out a study to investigate the role of Bisphenol A (BPA) in metabolism and endocrine disruption.

Dr. Hahobian recruited 11 subjects, each of whom ate two types of cookies on two separate visits, one visit in December and the second in February. On one visit they ate the BPA-laced cookie and, on a different day, a placebo cookie (with no BPA). Thirty minutes after eating the cookie on each occasion, they were given a glucose tolerance test to measure their glucose metabolism.

A summary of the glucose test results (mmol/L) after eating each type of cookie as well as the difference in glucose results for each subject is shown below.

Cookie	Mean	Standard Deviation	n
Placebo	5.259	0.762	11
BPA	5.355	1.462	11
Difference: Placebo - BPA	-0.095	1.153	11

(a)[4 points] For simplicity, Dr. Hagobian could have given all subjects the BPA cookie on their first visit in December, and the Placebo cookie on the second visit in February, but he didn't. Instead, when a subject came for their first visit, he flipped a coin. If it was heads, they received BPA on that visit (and Placebo on their second visit). If it was tails they received the Placebo cookie first. Why did he add this extra coin flipping step instead of the simpler approach of just giving everyone one type of cookie in December and the other type in February?

(b)[2 points] Dr. Hagobian is interested in testing whether BPA causes a shift in glucose levels. Which analysis would be more appropriate? Circle one.

Difference in Two Independent Means **Mean of the Paired Differences**

(c)[3 points] Based on your answer to (b), write out the null and alternative hypotheses for Dr. Hagobian's test using **notation**. *Be sure to indicate the order of subtraction being used!*

H_0**:**

H_A**:**

(d)[3 points] To perform the analyses you selected in (b), what conditions does Dr. Hagobian need to check before obtaining a p-value? Circle all that apply.

(i) Independence of the differences

(ii) Independence of the observations within each group

(iii) Independence of the observations between the groups

(iv) Independence of the variables

(v) Equal variance between the groups

(vi) Linear relationship between the variables

(vii) Normality of the differences

(viii) Normality of the observations within each group

(e)[3 points] Using `R`, Dr. Hagobian obtained the following table.

statistic	p_value	estimate	lower_ci	upper_ci
-0.6807	0.5115	-0.2314	-0.8476	0.3848

Which of the following would be the best overall conclusion in the context of Dr. Hagobian's study? Your selection should reflect the hypotheses you wrote in part (c)!

(i) With such a large p-value, we have significant evidence to reject the null hypothesis. We conclude the true mean of the differences in glucose between eating a BPA cracker and a Placebo cracker is not 0.

(ii) With such a large p-value, we have insufficient evidence to reject the null hypothesis. We conclude the true mean of the differences in glucose between eating a BPA cracker and a Placebo cracker is 0.

(iii) With such a large p-value, we have insufficient evidence to reject the null hypothesis. We do not have evidence to suggest the mean of the differences in glucose between eating a BPA cracker and a Placebo cracker is different from 0.

(vi) With such a large p-value, we have significant evidence to reject the null hypothesis. We conclude the true mean glucose after eating a BPA cracker is different from the true mean glucose after eating a Placebo cracker.

(v) With such a large p-value, we have insufficient evidence to reject the null hypothesis. We conclude there is no difference in the true mean glucose after eating a BPA cracker and the true mean glucose after eating a Placebo cracker.

(vi) With such a large p-value, we have insufficient evidence to reject the null hypothesis. We do not have evidence to suggest the true mean glucose after eating a BPA cracker is different from the true mean glucose after eating a Placebo cracker.

(f)[2 points] Based on the decision you reached in (e), what type of error could you have made? Circle one.

Type I Error **Type II Error** **No error was made**

(g)[2 points] If instead Dr. Hagobian had 100 subjects, the chance of the error described in part (f) would

increase **decrease** **stay the same**

(h)[2 points] In a different study, Dr. Hagobian obtained a p-value of 0.0425 and a 95% confidence interval of (-1.129, 0.0437). Which of the following statements about these findings is true? Circle one.

(i) The results of the hypothesis test and the confidence interval tend to agree with each other at the 5% significance level. Four percent of the time we would obtain a statistic like the one we saw somewhere in the interval of -1.129 mmol/L to 0.0437 mmol/L.

(ii) The results of the hypothesis test and the confidence interval are conflicting at the 5% significance level. With a p-value of 0.0425 we have evidence to reject the null hypothesis, which would mean that our confidence interval would not contain 0.

(iii) The results of the hypothesis test and the confidence interval are conflicting at the 5% significance level. There's a 95of 0.0425 would be in the interval (-1.129, 0.0437).

(vi) The results of the hypothesis test and the confidence interval seem to agree with one another at the 5% significance level. With a p-value of 0.0425 we do not have evidence to reject the null hypothesis, thus indicating that 0 should be in our interval.

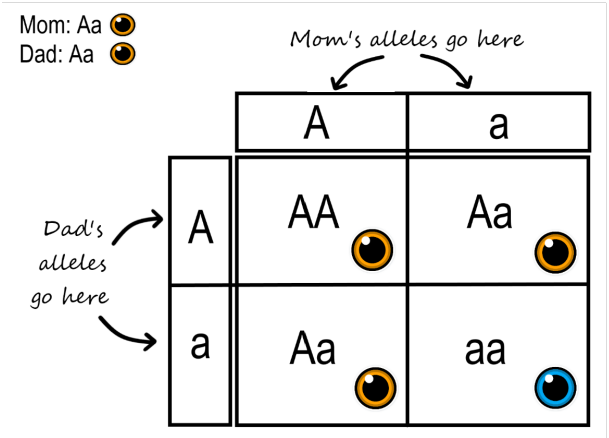
Final Exam Bank

Q2 [18 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 Punnett square 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, displayed 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 plant 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
$p_{\text{TT}} =$	$p_{\text{Tt}} =$	$p_{\text{tt}} =$

(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
$p_{\text{Tall}} =$	$p_{\text{Short}} =$

(b) [3 points] If the table above represents what is assumed to be true about tallness inheritance **if the null hypothesis is true**, 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" your observed number of tall and short plants were from what you would have expected to see 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?

Q6 [19 points] Montana Fish, Wildlife, & Parks personnel have collected data on fish caught on the Blackfoot River (outside Helena, Montana) for the last 25 years. To capture the fish, fisheries biologists use electrofishing equipment to attract the fish to the boat, then dip them out of the water with nets. Each fish's length (in cm) and weight (in grams) is then measured. Once the measurements are taken, the fish is tossed back into the river. Biologists are often working in cold conditions in late autumn or early spring, so some measurement error and missing data are expected.

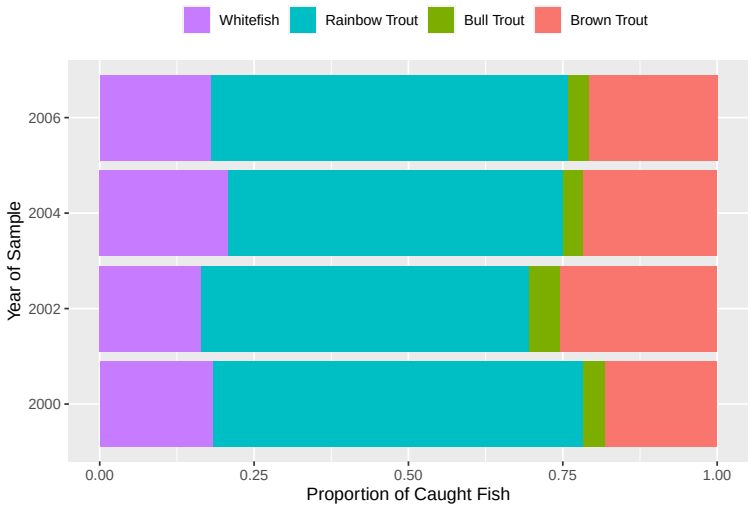
These data are not from a random sample. The goal is to catch all fish within a section of the Blackfoot River every few years to assess the health of the population. Montana Fish, Wildlife, & Parks is interested in analyzing if the prevalence of the different species of fish has stayed relative stable over the years. The dataset consists of 7729 observations, recorded over 4 years.

```
## # A tibble: 7,729 x 5
##    length weight year  section  species
##     <dbl>  <dbl> <fct> <chr>    <chr>
##  1    358    400 2000  Johnsrud Rainbow Trout
##  2    309    290 2000  Johnsrud Rainbow Trout
##  3    302    250 2000  Johnsrud Rainbow Trout
##  4    272    210 2000  Johnsrud Rainbow Trout
##  5    284    230 2000  Johnsrud Rainbow Trout
##  6    268    180 2000  Johnsrud Rainbow Trout
##  7    280    245 2000  Johnsrud Rainbow Trout
##  8    340    380 2000  Johnsrud Rainbow Trout
##  9    267    205 2000  Johnsrud Rainbow Trout
## 10    188     80 2000  Johnsrud Rainbow Trout
## # ... with 7,719 more rows
```

(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!



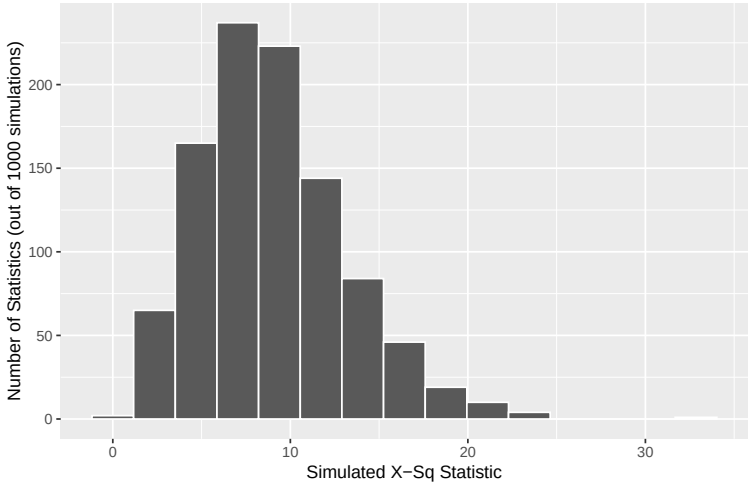
(d) [5 points] The statistician at Montana Fish, Wildlife, & Parks prefers using simulation-based methods rather than theory-based methods. They used computer simulation to obtain the null distribution shown below.

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.

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.

_____

(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 or not they would date someone who is taller than them.

_____

You are phenomenal! Congratulations on completing STAT 218!

Have a great summer!