Midterm 1 Question Bank

Stat 218

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

Provided Formulas

$$R^2 = r^2 = \frac{s_y^2 - s_{residuals}^2}{s_y^2}$$

$$IQR = Q3 - Q1$$

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

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

Q1 Dr. John Arbuthnot, an 18th century physician, writer, and mathematician is famous for performing the first hypothesis test of significance. Dr. Arbuthnot was interested in the ratio of newborn boys to newborn girls, so he gathered the baptism records for children born in London for every year from 1629 to 1710. Artbuthnot found that in every year, the number of males born in London exceeded the number of females.

(a) [2 pts] Describe the sampling method used by Dr. Arbuthnot.

(b) [2 pts] Describe how this sampling method could be biased.

(c) [3 pts] A preview of the dataset is provided below. Use this preview to address the following questions.

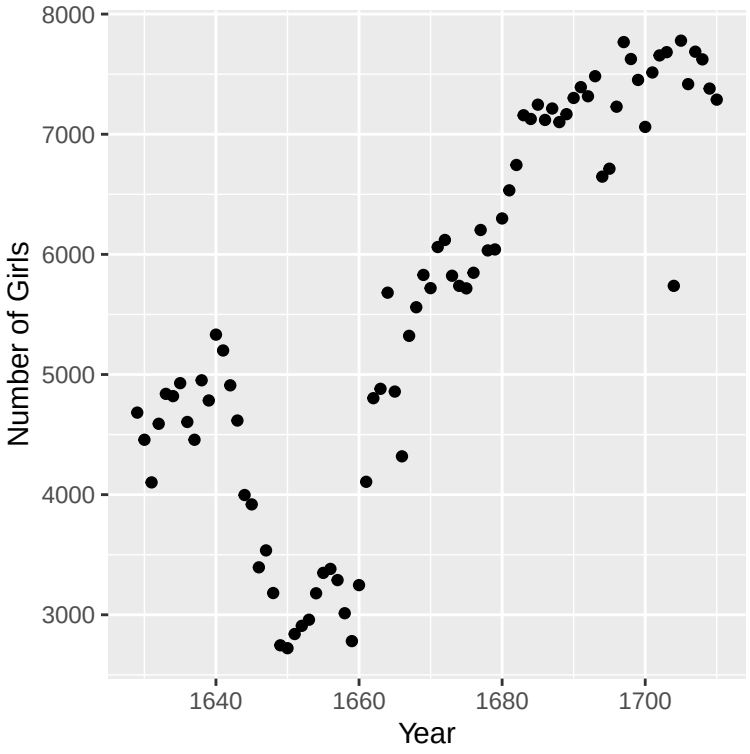
```
## # A tibble: 82 x 3
##     year  boys girls
##    <int> <int> <int>
##  1  1629  5218  4683
##  2  1630  4858  4457
##  3  1631  4422  4102
##  4  1632  4994  4590
##  5  1633  5158  4839
##  6  1634  5035  4820
##  7  1635  5106  4928
##  8  1636  4917  4605
##  9  1637  4703  4457
## 10  1638  5359  4952
## # ... with 72 more rows
```

- Identify the cases in the data set.

- List the variables. Indicate whether each variable is categorical or quantitative. If the variable is quantitative, give the units.

- What would the dimensions of the data set be? (number of rows by number of columns)

(d) [3 pts] A scatterplot displaying the number of girls born over time is displayed below. Describe the relationship you see in the scatterplot. Be sure to address the form, direction, strength, and outliers present in the scatterplot.



(d) [3 pts] Would it be appropriate to model the relationship between the number of girls born and the year with a linear regression? Justify your belief!

Q2 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 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.

(c) [2 pts] Is the alternative hypothesis one- or two-sided? Select one.

- One-sided
- Two-sided

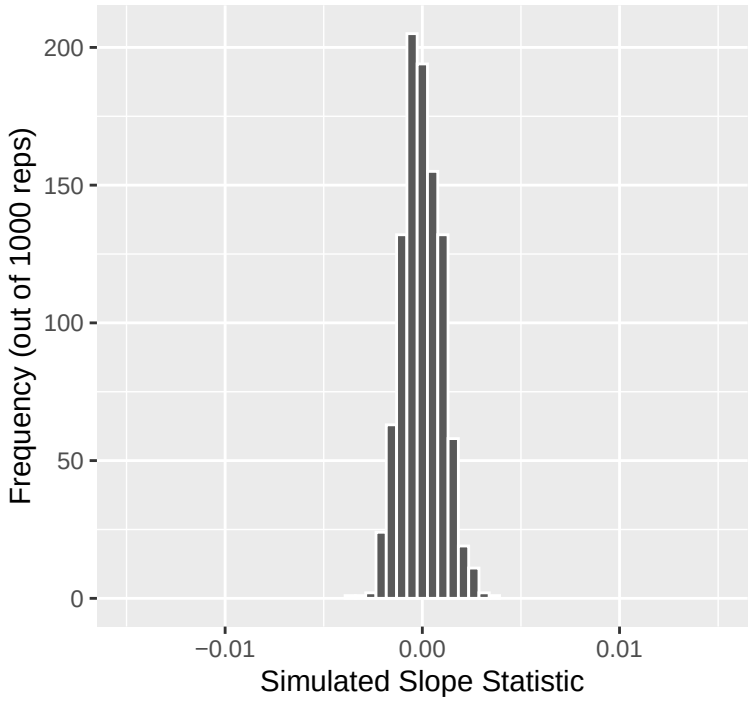
(d) [5 pts] On the following page is the plot of the simulated null distribution from `R`. Fill in the blanks below with one answer in each set of parentheses to correctly explain how one sample on the null distribution would be created. Blanks preceded by (#) should be filled in with a number. All other 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 (slope/correlation) _____ 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.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-0.771	0.406	-1.9	0.058	-1.569	0.026
calories	0.017	0.001	24.89	0	0.015	0.018

(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] Based on your estimated p-value, what would you conclude for this hypothesis test? (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

Q3 The Atlantic marsh fiddler crab, Minuca pugnax (formerly Uca pugnax), lives in salt marshes throughout the eastern coast of the United States. Historically, M. pugnax were distributed from northern Florida to Cape Cod, Massachusetts, but like other species have expanded their range northward due to ocean warming.

The Plum Island Ecosystem Long Term Ecological Research site collected data on fiddler crabs from 13 marshes on the Atlantic coast of the United States in summer 2016. The marshes spanned from northeast Florida to northeast Massachusetts. Researchers were able to collect between 25 and 37 adult male fiddler crabs at each marsh.

(a) [3 pts] A preview of the dataset is provided below. Use this preview to address the following questions.

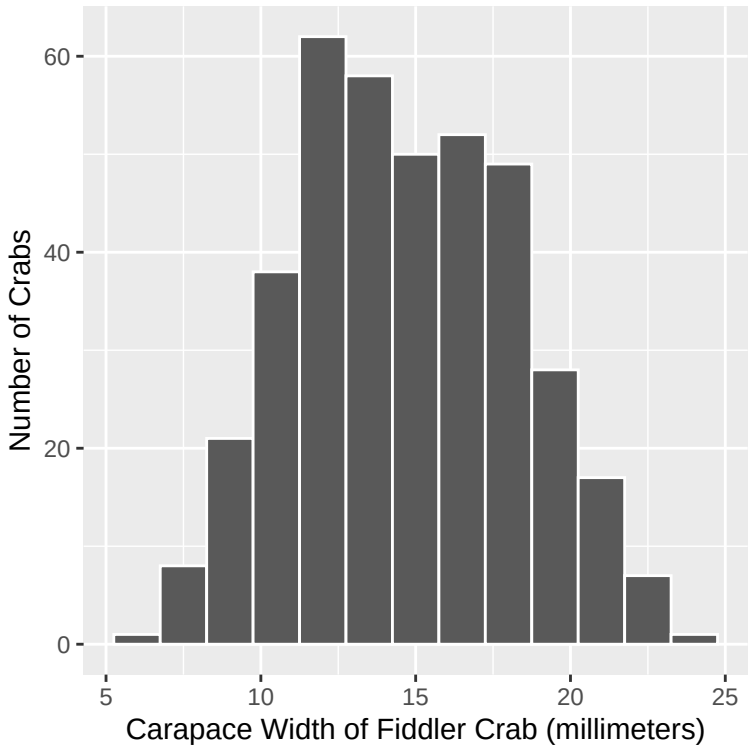
```
## Rows: 392
## Columns: 6
## $ date       <date> 2016-07-24, 2016-07-24, 2016-07-24, 2016-07-24, 2016-07-24~
## $ latitude   <dbl> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,~
## $ site       <chr> "GTM", "GTM", "GTM", "GTM", "GTM", "GTM", "GTM", "GTM", "GT~
## $ size       <dbl> 12.43, 14.18, 14.52, 12.94, 12.45, 12.99, 10.32, 11.19, 12.~
## $ air_temp   <dbl> 21.792, 21.792, 21.792, 21.792, 21.792, 21.792, 21.792, 21.~
## $ water_temp <dbl> 24.502, 24.502, 24.502, 24.502, 24.502, 24.502, 24.502, 24.~
```

- Identify the cases in the data set.

- List the variables. Indicate whether each variable is categorical or quantitative. If the variable is quantitative, give the units.

- What would the dimensions of the data set be? (number of rows by number of columns)

(b) [3 pts] A histogram displaying the size of the sample of fiddler crabs is displayed below. Describe the shape of the distribution. Be sure to address the center, spread, shape, and outliers.



(c) [2 pts] When using a t-distribution to find a 95% confidence interval, how many degrees of freedom should be used?

(d) [4 pts] A 95% confidence interval for the mean carapace width of a Fiddler Crab was found to be (14.31, 15.01). Below is the researchers' interpretation of this confidence interval:

There is a 95% chance that that every sample of Fiddler Crabs will have a mean carapace width between 14.31 and 15.01 millimeters.

Identify **two** mistakes committed and fix them. Be brief but clear in your description.

Mistake 1: _____

Fix: _____

Mistake 2: _____

Fix: _____

(f) [2 points] Can the researchers use their interval to make inferences about **all** fiddler Crabs in the United States? Justify your answer!

Q4 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

Q5 [4 points] Indicate whether each statement about a bootstrap sample is true or false.

- The resample and original sample MUST be the same size.
- The resample and original sample BOTH are taken from the population.
- The resample can ONLY use values that were in the original sample.
- The resample uses ALL values that were in the original sample.

Q6 The purpose of creating a bootstrap distribution is to:

1. Estimate the mean of the unknown sampling distribution.
2. Estimate the spread of the unknown sampling distribution.
3. Both A and B.
4. None of the above.

Q7 A newspaper article claims that the average age for people who receive food stamps in a community is 40 years. A local researcher believes that the average age is less than that. The researcher takes a random sample of 100 people in the community who receive food stamps, and finds their average age to be 39.2 years, which is statistically significantly lower than the age of 40 stated in the article (p-value $< .05$). Indicate for each of the following interpretations whether they are valid or invalid.

(a) The statistically significant result indicates that the majority of people who receive food stamps is younger than 40.

(a) Valid

(b) Invalid

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

(a) Valid

(b) Invalid

Q6 Normal human body temperature is said to be 98.6 degrees Fahrenheit. A researcher studying body temperatures took a random sample of 50 people and recorded the body temperatures for each. The average body temperature of the 50 people was 98.26. Using these data, estimate the true average body temperature of people.

(a) Define the parameter of interest, using proper notation.

(b) What type of inference should be done to answer the research question? (circle one)

- Confidence Interval
- Hypothesis Test
- No inference is needed. The true average body temperature is 98.6.
- No inference is needed. The true average body temperature is 98.26.

(c) The bootstrap distribution with 1001 bootstrap resamples is below. What does one dot on the plot represent in the context of this problem?

(d) Explain how you could use cards to create **one** bootstrap resample.

(e) The percentile method is used in the picture to give a confidence interval of (98.081, 98.512). What level of confidence is used for this interval?

(f) How many bootstrap resamples would be cut-off in the left-tail of the distribution to create a 95% confidence interval? (circle one)

- 50
- 25
- 5
- 10

(g) Use the plot and t* = 1.677 to create a 90% confidence interval using the t*SE method. Show your work.

(h) Interpret your 90% confidence interval from e.

(i) Based off your interval in (h), can we be 90% sure that the true average body temperature is lower than 98.6 degrees Fahrenheit? Explain.