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.

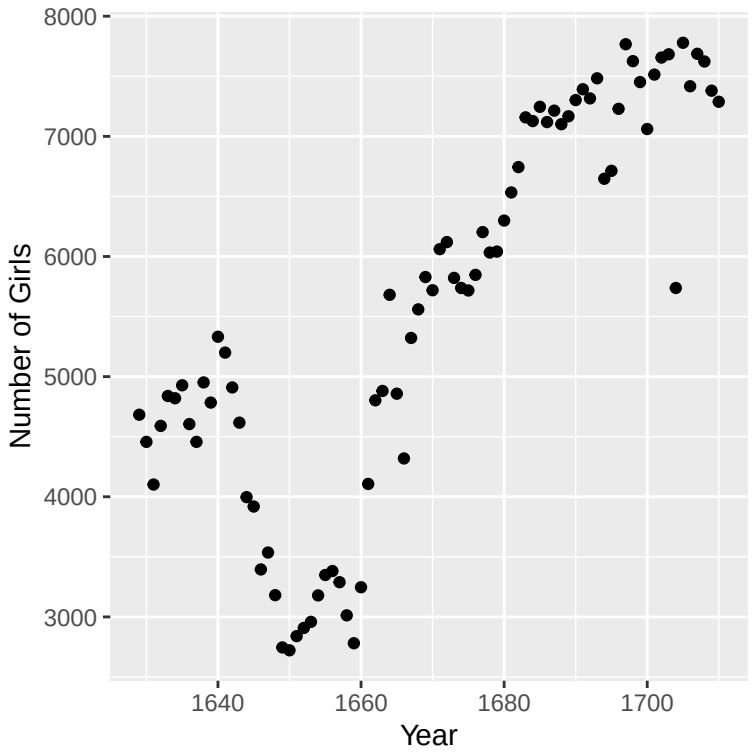
```
## # A tibble: 82 x 3
##     year  boys girls
##    <int> <int> <int>
##  1  1629  5218  4683
##  2  1630  4858  4457
##  3  1631  4422  4102
##  4  1632  4994  4590
##  5  1633  5158  4839
##  6  1634  5035  4820
##  7  1635  5106  4928
##  8  1636  4917  4605
##  9  1637  4703  4457
## 10  1638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

- Identify the cases in the data set.

- List the variables. Indicate whether each variable is categorical or quantitative.

- 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.



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

- One-sided
- Two-sided

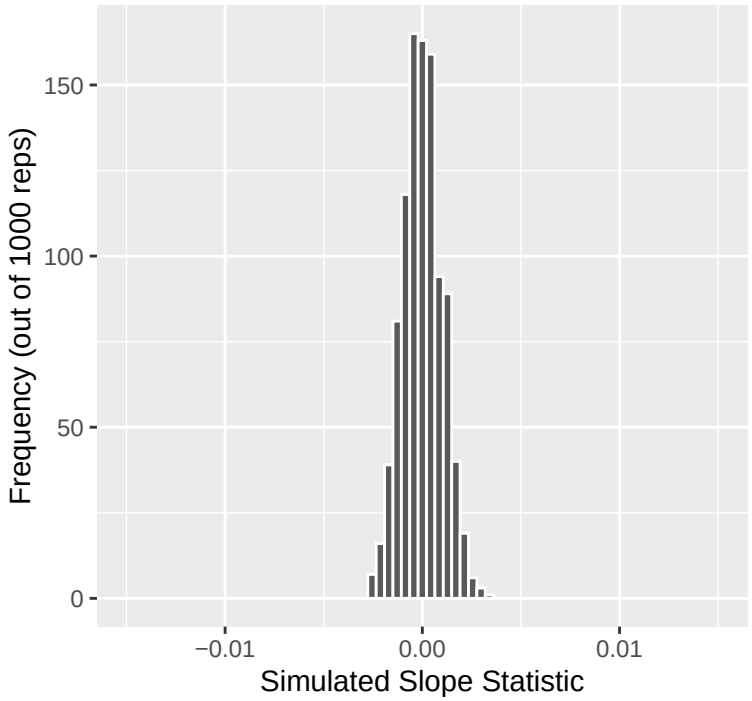
(d) [5 pts] On the following page is the plot of the simulated null distribution from `R`. Fill in the blanks below with one answer in each set of parentheses to correctly explain how one sample on the null distribution would be created. Blanks preceded by (#) should be filled in with a number, all other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

Q3 The Atlantic marsh Fiddler Crab, *Minuca 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 the summer of 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.

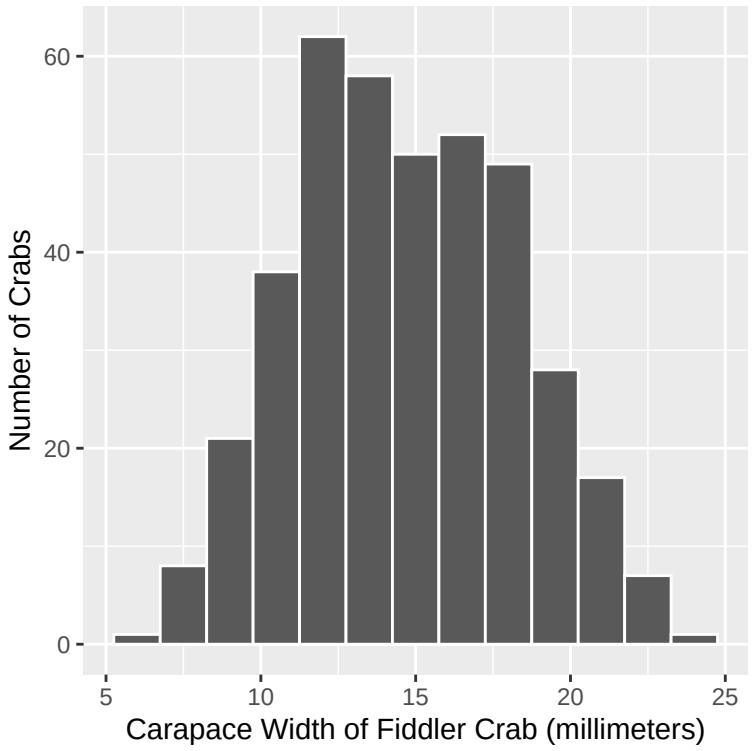
```
## Rows: 392
## Columns: 6
## $ date       <date> 2016-07-24, 2016-07-24, 2016-07-24, 2016-07-24, 2016-07-24~
## $ latitude   <dbl> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,~
## $ site       <chr> "GTM", "GTM", "GTM", "GTM", "GTM", "GTM", "GTM", "GTM", "GT~
## $ size       <dbl> 12.43, 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.

- 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 for μ, how many degrees of freedom should be used?

(d) [4 pts] A 95% confidence interval for the mean carapace width of Fiddler Crabs 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: _____

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

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

Q5 [4 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 [3 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.

Q7 [2 points] An article in the San Luis Tribune claims that the average age for people who receive food stamps in SLO is 40 years. A Cal Poly student believes the average age is less than that. The student obtains a random sample of 100 people in SLO who receive food stamps, and finds their average age to be 39.2 years. Performing a hypothesis test, the student finds their sample mean to be statistically significantly lower than the age of 40 stated in the article (p-value < 0.05). Indicate fore 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.

- Valid
- Invalid

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

- Valid
- Invalid

Q8 Kinesiology professor Suzanne Phelan has been studying effective tools to help women return to a healthy weight after pregnancy. Dr. Phelan designed a study with 488 new moms. Each of the new mothers was assigned to either standard care postpartum, or to an intervention group that included being part of online social support group moderated by a dietitian and lifestyle coach. Over the course of the study many measurements were taken including the weights of moms at the time their baby was born and again 6 months postpartum.

(a) [2 points] Using information above to briefly support your answer, was this an experiment or an observational study?

(b) [2 points] Dr. Phelan could have asked the new moms to choose whether they wanted to part of the online social group or not rather than assign the moms to the standard care and intervention groups. Why would this have been an inferior strategy for conducting this study? Briefly explain.

Q9 [2 points] 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

Q10 When using a t-distribution to obtain a confidence interval, how does the multiplier change from the 95% to the 90% confidence interval? (Circle the correct answer)

(a) Multiplier is larger

(b) Multiplier is smaller

(c) Multiplier stays the same

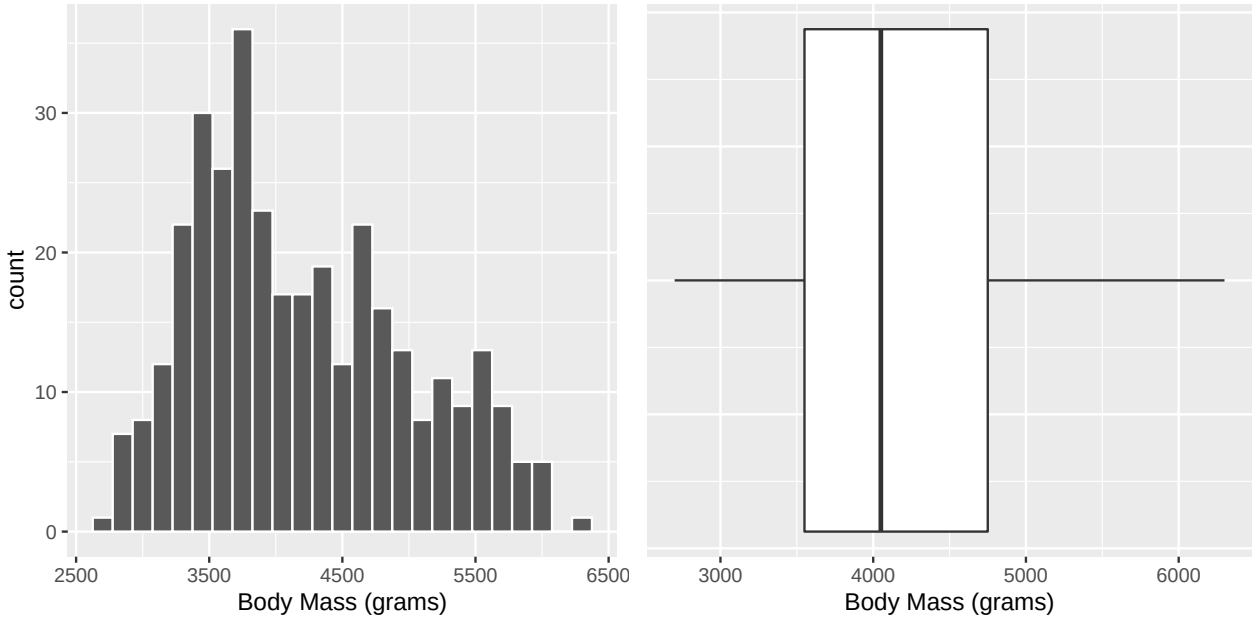
Q11 Based on the plots below, which is the better measure of center for this variable? (Circle one.)

(a) The median, as the shape of the distribution is strongly skewed.

(b) The mean, as the shape of the distribution is symmetric.

(c) The mean, as the shape of the distribution is strongly skewed.

(d) The median, as the shape of the distribution is symmetric.



Q12 The Konza Prairie Long-Term Ecological Research has collected data on bison on the Konza prairie since 1994, making it the longest continuous record of wild ungulate weight gain anywhere in the world. Researchers conduct a round-up once a year at the end of the grazing season wherein each bison is weighed, calves are vaccinated and receive unique IDs, and excess individuals are culled.

For this investigation, we are interested in assessing if, despite the effect of climate change on their habitat, the weight of yearling, male bison is what is described as "healthy" — a weight of approximately 750 pounds.

Below are summary statistics for the 48 of the yearling, male bison captured in 2020.

min	Q1	median	Q3	max	mean	sd	n	missing
490	595	620	662.5	770	629.5	63.17	47	1

(a) [3 points] Define the parameter of interest, using proper notation.

(b) [2 points] 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 weight of a yearling male bison is 629.5 pounds.
- No inference is needed. The true average weight of a yearling male bison is 750 pounds.

(c) [2 points] If we decided to use inferential methods to assess if the true mean weight of yearling bison on the Konza Prairie was healthy, we must verify two conditions. What are these two conditions?

Condition 1:

Condition 2:

(d) [2 points] Using the plot below and the description of the data collection procedure, justify if you believe whether each condition is violated or not.

Condition 1:

Condition 2: