Midterm 2 Review Solutions

STA 209, Week 11

For extra practice, several review problems are provided on this Review Sheet. However, they are not comprehensive, so do not limit your studying to just these problems. While the problems are intended to match the difficulty of those on the exam, the length of this review set does not represent the actual length of the exam. Solutions to these problems are posted the Exams page of the course website.

Problem 1

A poll asked Grinnell College students in an intro stats class "Suppose you are invited to play a high-stakes casino game. The game costs \$1,000 to play, but reliable sources tell you that there is a 50% chance that you win \$2,000 and a 50% chance that you win nothing. Would you play the game?"

Of the 24 respondents, 17 of them answered "no". Treating this sample as a simple random sample, the 95% confidence interval for the proportion of "no" answers among all Grinnell students is (0.526, 0.890)

a. In your own words, explain what is meant by "95% confidence."

b. How do you imagine the length of the confidence interval (the distance between the upper and lower endpoints) would change if the confidence level were changed to to 99%?

---

a. By 95% confidence, we mean that if we were to use a similar procedure to construct confidence intervals, for 95% of random samples, the resulting interval would contain the true population proportion.

b. If we were to increase the confidence level, without changing the sample and sample size, we would also increase the length of the interval, since the confidence interval would span a larger percentage of the bootstrap distribution.

Problem 2

Annual salaries (in dollars) for individuals with job title "business analyst" are reported on glassdoor.com from major American firms. 16 salaries are given below, arranged from smallest to largest:

51,413	51,875	51,951	52,200	54,600	59,757	60,000	60,000
61,445	61,896	62,610	63,250	67,750	69,629	71,063	94,566

Use the following code to load the data into R.

```
salary_data<-data.frame(salary =  c(51413, 51875, 51951, 52200,
   54600, 59757, 60000, 60000,
   61445, 61896, 62610, 63250,
   67750, 69629, 71063, 94566))
```

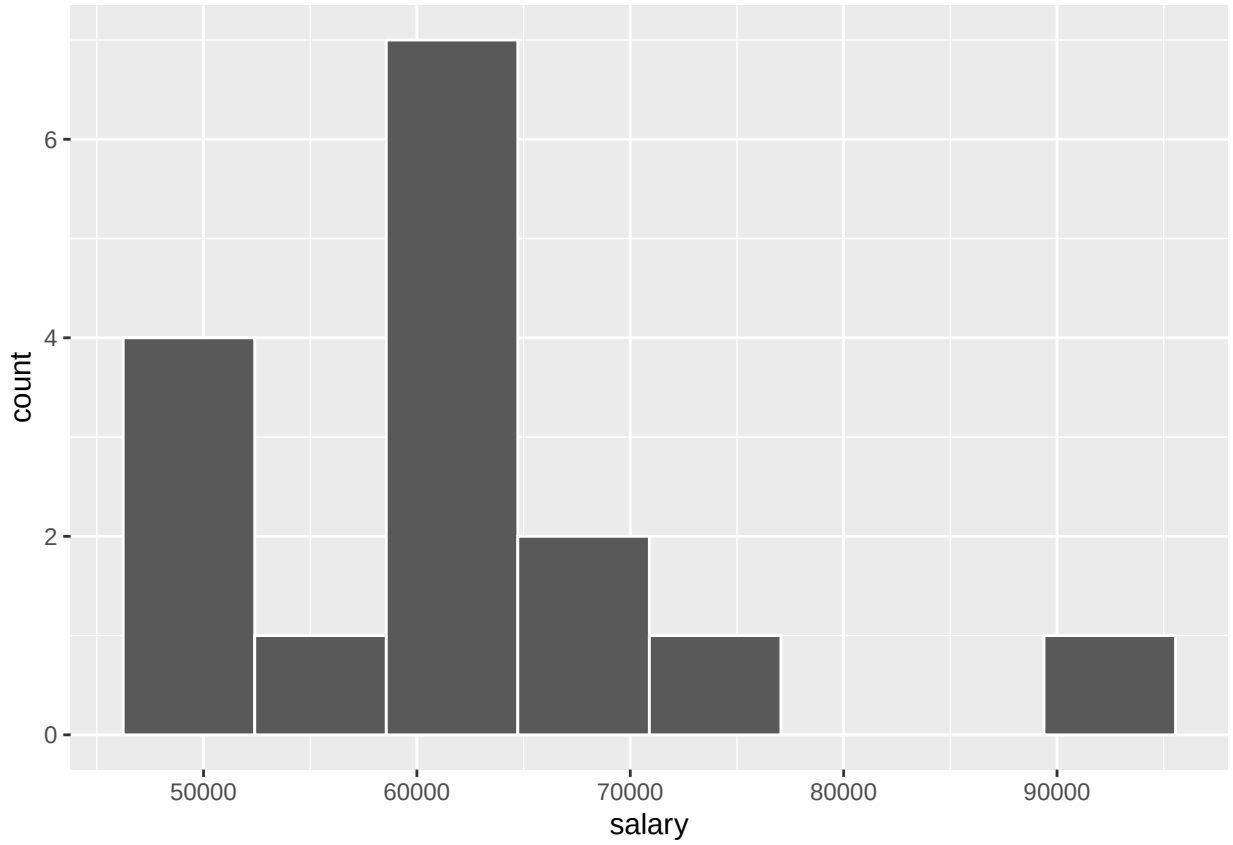
a. Visualize the data and calculate several summary statistics. Describe the center, shape and spread of the data.

b. Use `infer` and the percentile method to construct a 90% confidence interval for the mean business analyst salary.

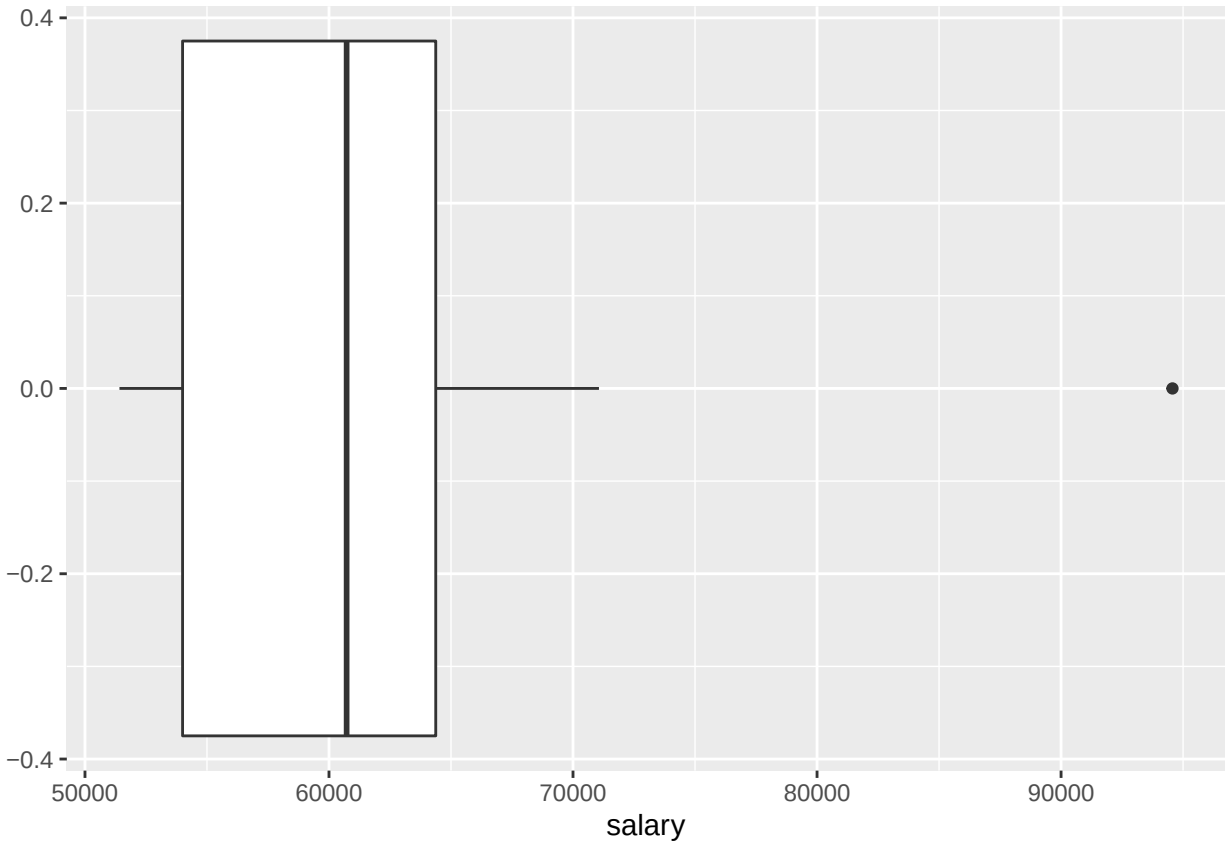
c. Based on your confidence interval, is it plausible that the true mean salary is \$50,000?

a. Based on the visualizations and summary statistics, the center of the distribution is about 60000 (median), with spread of about 10000 (IQR). Since the mean is larger than the median, the distribution is right-skewed.

```
ggplot(salary_data, aes(x = salary))+geom_histogram(bins = 8, color = "white")
```




```
ggplot(salary_data, aes(x = salary))+geom_boxplot()
```




```
salary_data %>%
  summarize(
    mean = mean(salary),
    sd = sd(salary),
    min = min(salary),
    q1 = quantile(salary,.25),
    median = median(salary),
    q3 = quantile(salary, .75),
    max = max(salary)
  )
```

```
##       mean       sd   min    q1  median    q3   max
## 1 62125.31 10693.81 51413 54000 60722.5 64375 94566
```

b.

```
library(infer)
set.seed(100)
salary_data %>%
  specify(response = salary) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.90, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   58342.   66659.
```

The 90% confidence interval for mean salary is (58342, 66659).

c. Since 50000 is not contained within the 90% confidence interval, this value is not plausible for the mean salary.

Problem 3

A company is interested in testing whether one of their energy drink products has an effect on physical endurance. A sample of 100 high school students is recruited. Subjects are randomly divided into a treatment and a control group. Subjects in the treatment group are asked to run on a treadmill after consuming 4 ounces of Minotaur Energy Drink, while subjects in the control group are asked run on a treadmill without consuming any beverage. An experimenter records the amount of time a subject spends on the treadmill before reporting exhaustion.

a. What are the explanatory and response variables in this experiment?

b. State the implied null and alternative hypotheses for this experiment.

c. Describe how you could use 100 slips of paper, along with the data from the experiment, to generate a new simulated sample assuming the null hypothesis were true.

d. Suppose the researchers collected data and found that only 4% of statistics in the null distribution were **larger** than the observed statistic from the experiment. What conclusions should the researchers make about their hypotheses? (Use an $\alpha = 0.05$ significance level).

e. Suppose that the observed average time until exhaustion among the control group was 10 minutes and 12 seconds, and that the average time until exhaustion among the treatment group was 10 minutes and 20 seconds. Regardless of your answer to the previous part, do you feel that the observed effect size is of practical significance? Justify your answer.

---

a. In this investigation, the explanatory variable is group (treatment vs. control), while the response variable is time until exhaustion.

b. The researchers are investigating whether the energy drink has an effect on physical endurance. Let mu_t be the *average* time until exhaustion among the treatment population, and let mu_c be the *average* time until exhaustion among the control population. Our hypotheses are

H_0: mu_t - mu_c = 0

H_a: mu_t - mu_c =/= 0

c. If the null hypothesis were true, we assume that exhaustion time and group are independent. To simulate one new data set under this hypothesis, we write the exhaustion times of all 100 participants on sheets of paper. Then we shuffle them together and randomly choose 50 to represent a new control group and the remaining 50 to represent the new treatment group. We calculate the mean exhaustion times among the two groups.

d. Since we are using a two-sided alternative hypothesis, the p-value is *double* the proportion of simulated statistics greater than the observed statistic; in this case, the p-value is 0.08. At the 0.05 level, we would not reject the null hypothesis.

e. The observed difference in average exhaustion times between the two groups was 8 seconds. Since both average exhaustion times are approximately 10 minutes, this difference represents approximately a 1.5% increase in exhaustion time. This does not seem to be a practically significant difference.

Problem 4

Use the following code to load the `gss` data set from the `infer` package.

```
library(infer)
data(gss)
```

We are interested in investigating whether the proportion of individuals with a college degree differs between the two primary political parties (Democrats and Republicans). A respondent's political party affiliation is recorded in the `partyid` variable, and whether or not they have a college degree is recorded in the `college` variable.

a. Write the null and alternative hypothesis, both in words and in symbols.

b. Use `dplyr` verbs to create a new data set just consisting of individuals whose political party affiliation is either Republican or Democrat.

c. Calculate the difference in proportion of college degree holders for the two political parties.

d. Use `infer` to simulate the distribution of statistics under the null hypothesis, and calculate the p-value of the observed statistic.

e. Make a conclusion about the null hypothesis at the $\alpha = 0.10$ significance level.

---

a. Let p_r and p_d represent the proportion of republicans and democrats with college degree, respectively. Our null and alternative hypotheses are:

H_0: p_r - p_d = 0

H_a: p_r - p_d =/= 0

b.

```
two_parties <- gss %>%
  filter(partyid == "dem" | partyid == "rep")
```

c.

We can calculate the observed difference using `infer`:

```
diff_prop_college <- two_parties %>%
  specify(response = college, explanatory = partyid, success = "degree") %>%
  calculate(stat = "diff in props", order = c("dem", "rep"))
```

```
## Dropping unused factor levels c("ind", "other", "DK") from the supplied explanatory variable 'partyi
```

```
diff_prop_college
```

```
## Response: college (factor)
## Explanatory: partyid (factor)
## # A tibble: 1 x 1
##      stat
##     <dbl>
## 1 -0.0951
```

Or alternatively, using `dplyr`

```
two_parties %>%
  group_by(partyid, college) %>%
  summarize(count = n()) %>%
  mutate(prop = count/sum(count))
```

```
## `summarise()` has grouped output by 'partyid'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 x 4
## # Groups:   partyid [2]
##   partyid college   count  prop
##   <fct>   <fct>     <int> <dbl>
## 1 dem     no degree   119 0.672
## 2 dem     degree       58 0.328
## 3 rep     no degree    71 0.577
## 4 rep     degree       52 0.423
```

```
diff_prop_college_dply <- 0.3276836  - 0.4227642
```

d. Based on a simulation of 5000 null statistics, the approximate p-value is 0.1212.

```
set.seed(1002)
null_dist_diff_prop_college <- two_parties %>%
  specify(response = college, explanatory = partyid, success = "degree") %>%
  hypothesize(null= "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("dem", "rep"))
```

```
## Dropping unused factor levels c("ind", "other", "DK") from the supplied explanatory variable 'partyi
```

```
null_dist_diff_prop_college %>%
  get_p_value(obs_stat = diff_prop_college, direction = "both" )
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.121
```

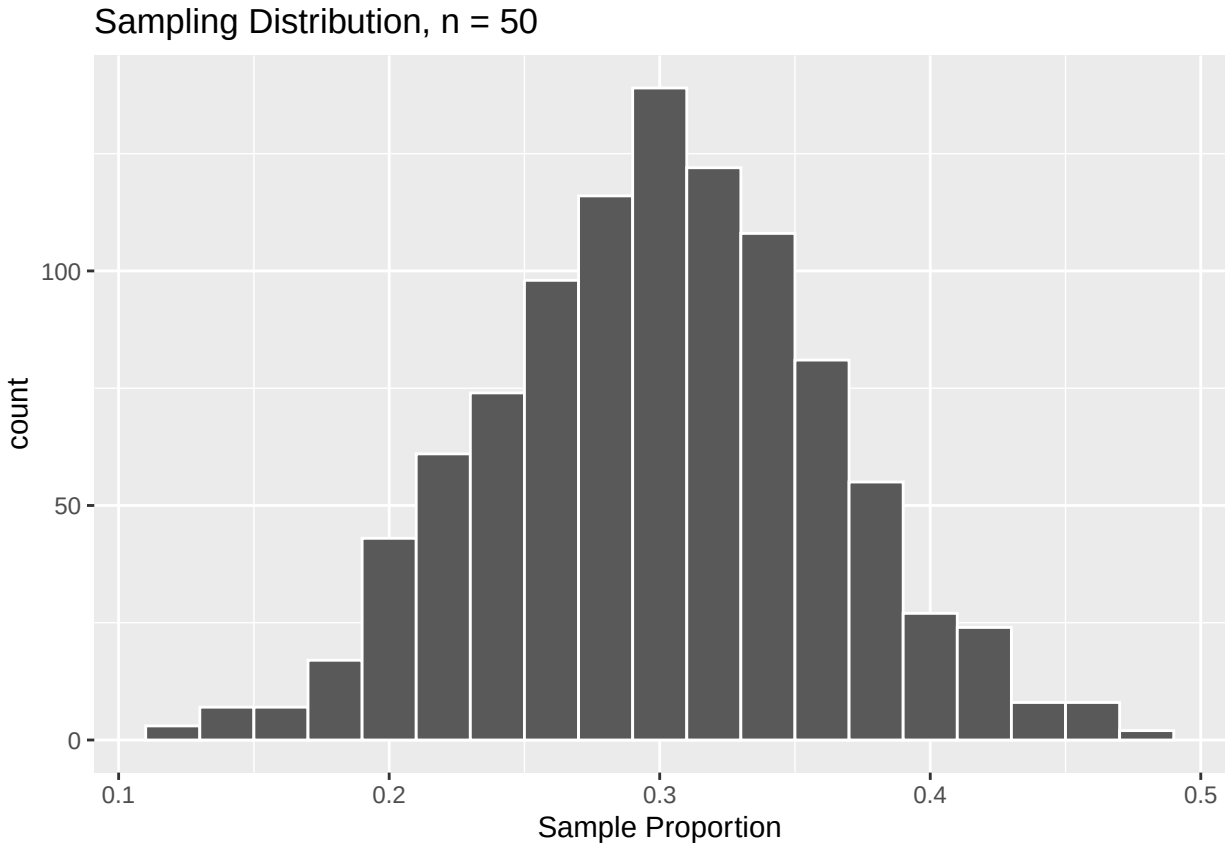
e. Since the p-value is greater than 0.10, we would not reject the null hypothesis at the 0.1 significance level. This sample does not give sufficient evidence that there is a difference in proportion of college degree holders between democrats and republicans.

Problem 5

The dataset `HollywoodMovies` contains information on movies to come out of Hollywood between 2012 and 2018. Of the 1295 movies in the data set, 386 of them are dramas. If we take 1000 samples of size n = 50 from the population of movies in this period and record the proportion of movies in the sample that are dramas, we get the following sampling distribution:

```
library(Lock5Data)
data(HollywoodMovies)

set.seed(121)
rep_sample_n(HollywoodMovies, reps = 1000, size = 50) %>% group_by(replicate) %>%
  filter(Genre == "Drama") %>%
  summarize(prop = n()/50) %>%
  ggplot(aes(x = prop))+geom_histogram(binwidth = 0.02, color = "white")+
  labs(x = "Sample Proportion", title = "Sampling Distribution, n = 50")
```



a. Based on the histogram, estimate the approximate value of the standard error of the sampling distribution, as well as the mean of the sampling distribution.

b. If we were to create a new sampling distribution using sample sizes of $n = 100$, would we expect the *center* of the new distribution to be smaller than, about the same as, or larger than the center of the distribution above?

c. If we were to create a new sampling distribution using sample sizes of $n = 100$, would we expect the *standard error* of the new distribution to be smaller than, about the same as, or larger than the standard error of the distribution above?

d. If we created a new sampling distribution using 10000 samples of size n = 50, would we expect the *center* of the new distribution to be smaller than, about the same as, or larger than the center of the distribution above?

e. If we created a new sampling distribution using 10000 samples of size n = 50, would we expect the *standard error* of the new distribution to be smaller than, about the same as, or larger than the standard error of the distribution above?

---

a. The standard error is the standard deviation of the sampling distribution. Since this distribution is approximately bell-shaped, we know that 95% of data are within 2 standard deviations of the mean. Based on the histogram, the mean seems to be about .3, and it appears that 95% of the data are between .18 and .42, suggesting that the standard deviation is about (.42 - .3)/2 = 0.06.

b. The mean of every sampling distribution will be the true population parameter. Changing the sample size from n = 50 to n = 100 will not effect the center of the sampling distribution.

c. On the other hand, increasing the sample size will cause the sampling distribution to narrow, resulting in a decrease in the standard error.

d. The number of samples used to estimate the sampling distribution will change the consistency of the appearance of the histogram (more samples used will result in more consistency in the shape). However, changing the number of samples will not have large effect on the mean and standard error of the sampling distribution.

e. As indicated above, changing the number of samples will not have large effect on the mean and standard error of the sampling distribution.

Problem 6

Given a specific sample to estimate a specific parameter from a population, what are the expected similarities and differences in the corresponding sampling distribution (using the given sample size) and bootstrap distribution (using the given sample)? In particular, for each aspect of a distribution listed below, indicate whether the values for the two distributions (sampling distribution and bootstrap distribution) are expected to be approximately the same or different. If they are different, explain how.

a. The shape of the distribution

b. The center of the distribution

c. The spread of the distribution

d. What one value (or dot) in the distribution represents

e. The information needed in order to create the distribution

---

a. The shape of the distributions will be approximately the same. Moreover, both will tend to be approximately bell-shaped and symmetric.

b. The centers of the two distributions will be different. The sampling distribution will be centered at the true value of the population parameter, while the bootstrap distribution will be centered at the value of the observed sample statistic.

c. The spread of both distributions will be approximately the same. The standard deviation of the bootstrap distribution is a good approximation of the standard error.

d. One value in the bootstrap distribution represents the statistic from a bootstrap sample (a sample obtained by sampling with replacement from the original). On the other hand, one value in the sample distribution represents the statistic from a sample obtained from the population.

e. In order to create the sampling distribution, we need to be able to obtain many samples from the original population (usually an impossible task). On the other hand, to create the bootstrap distribution, we just need to resample from the original sample, which is feasible especially with computer assistance.

Problem 7

Resveratrol, an ingredient in red wine and grapes, has been shown to promote weight loss in rodents. One study investigates whether the same phenomenon holds true in primates. The grey mouse lemur, a primate, demonstrates seasonal spontaneous obesity in preparation for winter, doubling its body mass. A sample of six lemurs had their resting metabolic rate, body mass gain, food intake, and locomotor activity measured for one week prior to resveratrol supplementation (to serve as a baseline) and then the four indicators were measured again after treatment with a resveratrol supplement for four weeks. Some p-values for tests comparing the mean differences in these variables (before vs after treatment) are given below. In parts (a) to (d), state the conclusion of the test using a 5% significance level, and interpret the conclusion in context.

a. In a test to see if mean resting metabolic rate is higher after treatment, $p = 0.013$.

b. In a test to see if mean body mass gain is lower after treatment, p = 0.007.

c. In a test to see if mean food intake is affected by the treatment, p = 0.035.

d. In a test to see if mean locomotor activity is affected by the treatment, p = 0.980.

e. In which test is the strongest evidence found? The weakest?

f. How do your answers to parts (a) to (d) change if the researchers make their conclusions using a stricter 1% significance level?

g. For each p-value, give an informal conclusion in the context of the problem describing the level of evidence for the result.

h. The sample only included six lemurs. Do you think that we can generalize to the population of all lemurs that body mass gain is lower on average after four weeks of a resveratrol supplement? Why or why not?

---

a. Reject H_0. The mean resting metabolic rate appears to be higher in lemurs after treatment.

b. Reject H_0. The mean body mass gain appears to be lower in lemurs after treatment.

c. Reject H_0. The mean food intake appears to change after treatment.

d. Do not reject H_0. This sample does not give sufficient evidence to suggest a change in mean locmotor activity after treatment.

e. The strongest evidence corresponds to the lowest p-value, which occurs for the test of mean body mass gain. The weakest evidence corresponds to the highest p-value, which occurs for the test for locomotor activity.

f. Answers to parts (b) and (d) remain the same with a 1% significance level, but parts (a) and (c) would change to "do not reject H_0"

g. We have strong evidence that the mean metabolic rate is lower for lemurs after treatment, very strong evidence that the mean body mass gain is lower, strong evidence that the mean food intake is difference, but no evidence that mean locomotor activity is different.

h. Assuming the sample represents a random sample from the population of all lemurs, we can generalize the findings from this investigation; despite the small sample size, the obesrved effects were larger that we would anticipate due to chance alone, if the null hypothesis were true.
