Math 58 / 58B - Introduction to (Bio)Statistics

your name here

due Friday, April 17, 2020

Homework 9

Assignment Summary (Goals)

practice creating & interpreting:

- t-tests for the population mean
- t-intervals for the population mean (Conf Int)
- t-intervals for predicting single response (Pred Int)
- t-tests and intervals for two population means

1. Guess the Instructor's Age^1

The file AgeGuesses.txt contains guesses of an instructor's age by her current students (not *your* instructor, but the instructor of the textbook). Let μ represent the average guess of her age by all current students at the university and suppose the sample constitutes a representative sample of all students at this school on this issue.

head(AgeGuesses)

guesses
1 26
2 28
3 300
4 31
5 33
6 35

- (a) Produce numerical and graphical summaries of the distribution and describe what you learn (in context).
- (b) Use a histogram to decide whether the data has strong deviations from the pattern of a normal distribution.
- (c) Use t.test to determine a 90% one-sample t-interval for these data. Include your output and comment on the validity of this procedure. Provide a one-sentence interpretation of this interval.
- (d) Count how many of the class guesses are inside the 90% confidence interval. Compute the percentage of the class guesses that are inside the interval. Is this close to 90%? Should it be?

Recall the code from last week that looked like this:

```
data %>%
filter(variable > low# , varible < high#) %>%
summarise(n())
```

(e) Calculate and interpret a 90% prediction interval. Include the details of your calculation and comment on the validity of this procedure. How does the prediction interval compare (midpoint, length) to the confidence interval? How many of the class guesss are inside the interval?

(note: the interval calculation should be done "by hand" using R as a calculator)

¹From ISCAM, HW 2.5

2. Feeling Motivated?²

A psychology study investigated whether people display more creativity when they are thinking about intrinsic or extrinsic motivations. The subjects were 47 people with extensive experience with creative writing. They were randomly assigned to one of two groups: one group answered a survey about intrinsic motivations for writing (such as the pleasure of self-expression) and the other group answered a survey about extrinsic motivations (such as public recognition). Then all subjects were instructed to write a Haiku poem, and these poems were evaluated for creativity by a panel of judges. The researchers conjectured that subjects who were thinking about intrinsic motivations would display more creativity than subjects who were thinking about extrinsic motivations. The creativity scores from this study are below and also in the file creativity.txt (http://www.rossmanchance.com/iscam2/data/creativity.txt).

- (a) Identify the explanatory and response variables. Also classify each as categorical or quantitative.
- (b) Is this an observational study or a randomized experiment? Explain how you know.
- (c) Create a boxplot of the scores broken down by the two groups (intrinsic and extrinsic motivation). Comment on what they reveal about the researchers' conjecture.

creativity <- read_delim("http://www.rossmanchance.com/iscam2/data/creativity.txt", "\t")</pre>

- (d) Report the mean of the creativity scores for each group. Do these summary values indicate that the intrinsically motivated group did indeed display more creativity than the intrinsically motivated group?
- (e) Use t.test to complete a t-test to test whether there is a difference in the average creativity score for intrinsic vs extrinsic motivation. Provide the hypotheses and p-values.
- (f) Changing the research question slightly, suppose you thought the intrinsic motivation would, on average, add 10 points to the creativity scores. Specify the corresponding null and (two-sided) alternative hypotheses.
- (g) Use R to compute a 95% confidence interval comparing the two groups. Include your output and interpret the interval.
- (h) Using the confidence interval, does 10 appear to be a plausible value for the difference in the underlying treatment means? Explain your reasoning.
- (i) Summarize your conclusion in the context of this study. Include an explanation of the reasoning process behind your conclusion. Be sure to address the issues of causation (i.e., is a cause-and-effect conclusion warranted?) and generalizability (i.e., how broadly can you legitimately generalize your conclusion?), as well as the issue of statistical significance.

²From ISCAM, HW 4.5

The following problems are NOT DUE... you can use them for extra practice!

Low Carb Diet³

A study by Foster el al., reported in The New England Journal of Medicine (May, 2003), investigated the effectiveness of a popular "low-carb" diet. The researchers randomly assigned 63 obese men and women to either a low-carbohydrate, high-protein, high-fat (Atkins) diet or a low-calorie, high-carbohydrate, low-fat (conventional) diet. The mean amount of weight lost, as percent of body weight, after 3 months, 6 months and 12 months are shown in the table below.

(The baseline weight was carried forward in the case of missing values.)

| Time | Diet | n | Mean | SD |
|-----------|--------------------------|-------|---------|----------|
| 3 months | Low-carb Conventional | 33 30 | 6.8 2.7 | 5.0 3.7 |
| 6 months | Low-carb Conventional | 33 30 | 7.0 3.2 | 6.5 5.6 |
| 12 months | Low-carb Conventional | 33 30 | 4.4 2.5 | 6.7 6.3 |

(a) Is this an observational study or an experiment? Explain.

This is an experiment, because the researchers randomly assigned subjects to either the low-carb diet or the conventional diet.

(b) Identify the explanatory and response variables.

The explanatory variable is the type of diet (low-carb or traditional) to which the subject was assigned. The response variable is the amount of weight loss as a percentage of body weight.

(c) Report the relevant hypotheses (in symbols) for testing whether the mean weight losses differ significantly between the two diets.

The hypotheses are:

 $\begin{aligned} H_0: \mu_{lowcarb} &= \mu_{conventional} \\ H_a: \mu_{lowcarb} &\neq \mu_{conventional} \end{aligned}$

(d) Calculate the t-test statistic for testing these hypotheses at the 3-month point. Also report the p-value and your test decision at the .05 significance level.

The test statistic is:

$$t = \frac{\overline{x}_{lc} - \overline{x}_{conv}}{\sqrt{\frac{s_{lc}^2}{n_{lc}} + \frac{s_{conv}^2}{n_{conv}}}} = \frac{6.8 - 2.7}{\sqrt{\frac{(5.0)^2}{33} + \frac{(3.7)^2}{30}}} \approx 3.72.$$

The approximate (two-sided) p-value, based on a t-distribution with 29 df, is 2(0.000425) = 0.00085. mosaic::xqt(.975, df= 29, ncp = 0) # two-sided rejection region

 $^{^{3}}$ From ISCAM, HW 4.9



[1] 2.04523





[1] 0.0008508014

At the $\alpha = 0.05$ significance level, we reject H_0 (3.72 > 2.05) and conclude that the experimental data provide strong evidence that the two types of diets produce different weight loss on average after three months, p-value = 0.00085.

(e) Repeat (d) for comparing the weight losses between the two diets at the 6-month point and again at the 12-month point.

One of the ISCAM applets allows for all the calculations automatically: http://www.rossmanchance.com/applets/TBIA.html

For six-month weight loss, the Theory Based Inference applet reports that the (unpooled) test statistic is t = 2.49 and p-value is 0.0187, so we still reject H0 (at the 0.05 level) and conclude that the experimental data provide strong evidence that the two types of diets produce different weight loss on average after six months. The evidence of a difference is less strong after six months than after three months, however.

For twelve-month weight loss, the Theory Based Inference applet reports that the (unpooled) test statistic is t = 1.16 and p-value is 0.2556, so we fail to reject H0 (at the 0.05 level) and conclude that the experimental data provide no/little evidence to believe that the two types of diets produce different weight loss on average after twelve months.

(f) Summarize your conclusions from these three tests. In particular, what do you notice about the trend in the p-value as time passes, and what does that reveal?

One of the ISCAM applets allows for all the calculations automatically: http://www.rossmanchance.com/applets/TBIA.html

The low-carb diet does appear to produce significantly more weight loss on average than the conventional low-fat diet for 3-6 months. However, by 12 months this difference was no longer statistically significant; in fact, there is some evidence that the amount of weight loss decreases over time for both groups. We do not know how these subjects were selected, but we suspect that they are volunteers so it is not clear whether they are representative of all obese men and women. This was a randomized experiment, so we can say that the significant difference after three and six months is caused by the type of diet used. On the whole, there is some evidence that the low-carb diet is beneficial but that the benefit decreases over time and is gone by the time the subjects have been on the diet for a year.

(g) Report the 95% confidence intervals for the difference in mean weight loss between the two diets at each time point. Comment on how these confidence intervals change across the three time points.

The 95% confidence intervals for $\mu_{lc} - \mu_{conv}$ from the Theory Based Inference applet are: After 3 months: (1.85, 6.35) After 6 months: (0.75, 6.85) After 12 months: (-1.38, 5.28)

These intervals shift toward lower values at later time periods. After three months, the CI suggests that the low-carb diet does better (produces more weight loss) by an average of roughly 1.85 to 6.35 percent of body weight. After six months, the CI is a bit wider and extends to smaller values but is still all positive, still suggesting that the low-carb diet does better. But after twelve months, the interval includes zero, suggesting that the mean percentage weight loss might be the same for the two types of diets.

Close Friends⁴

One of the questions asked of a random sample of adult Americans on the 2004 General Social Survey was:

From time to time, most people discuss important matters with other people. Looking back over the last six months - who are the people with whom you discussed matters important to you? Just tell me their first names or initials.

The interviewer then recorded how many names each person gave, with the person's sex.

(a) The relevant parameter for this study can be symbolized as $\mu_{men} - \mu_{women}$. Describe what this parameter means in this context.

This parameter is the difference between the population mean number of close friends for adult American men and the population mean number of close friends for adult American women.

(b) State the appropriate null and alternative hypotheses (in symbols) for testing whether American men and women differ with regard to average number of close friends.

The survey responses are summarized in the following table (and in the data file http://www.rossmanchance.com/iscam2/data/CloseFriends.txt):

| Number of close friends | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|--|------------|--|--------------|---|----------|--|----------|------------|
| Number of men responses Number of women responses | 196 201 | $\begin{array}{c} 135\\ 146 \end{array}$ | $108 \\ 155$ | $\begin{array}{c} 100 \\ 132 \end{array}$ | 42 86 | $\begin{array}{c} 40\\ 56 \end{array}$ | 33 37 | 654 813 |

The null hypothesis is that the population mean number of close friends for adult American men is the same as for women $(\mu_{men} - \mu_{women} = 0)$. The alternative is that the population mean number of close friends is not the same for men as for women $(\mu_{men} - \mu_{women} \neq 0)$.

(c) Use technology to produce graphs for comparing the distribution of number of close friends between men and women. Comment on what the histograms reveal about the shapes of the distributions.

The distribution of number of close friends is sharply skewed to the right for both men and women, as shown in the histograms below:

⁴From ISCAM, HW 4.42



(d) Use technology to determine the sample mean and sample standard deviation of the number of close friends for each sex. Report these with appropriate symbols. Also show how to calculate the sample means by hand from the table above.

The sample mean number of close friends for women is x^{-} women = 2.0886, with a standard deviation of swomen = 1.7601 friends. For men these statistics are x^{-} men = 1.8609 and smen = 1.7771 friends.

```
friends %>%
group_by(gender) %>%
summarize(means = mean(NumberCloseFriends), sds = sd(NumberCloseFriends), ns = n())
```

```
## # A tibble: 2 x 4
##
     gender means
                    sds
                            ns
##
     <chr>
            <dbl> <dbl> <int>
## 1 female
            2.09
                   1.76
                           813
## 2 male
             1.86
                   1.78
                           654
```

(e) Conduct a two-sample t-test of the hypotheses from (b). Report the test statistic and p-value. State your test decision at the 0.05 significance level, and summarize your conclusion.

Because this p-value (0.1442) is less than 0.05, we reject the null hypothesis. The sample data provide very strong evidence that American men and women do not have the same number of close friends, on average.

```
t.test(NumberCloseFriends ~ gender, data=friends)
```

```
##
##
   Welch Two Sample t-test
##
## data: NumberCloseFriends by gender
## t = 2.4497, df = 1392.8, p-value = 0.01442
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
   0.04536669 0.41004255
##
## sample estimates:
## mean in group female
                          mean in group male
##
               2.088561
                                    1.860856
```

(f) Produce a 95% confidence interval for the difference in population means (for the number of close friends) between men and women. Also write a sentence or two interpreting what the interval reveals.

This interval is entirely negative, consistent with our conclusion that men and women differ with regard to average number of close friends. We are 95% confident that American men have between 0.045 and 0.410 fewer close friends, on average, than American women do.

```
t.test(NumberCloseFriends ~ gender, data=friends, conf.level=0.95)
##
##
   Welch Two Sample t-test
##
## data: NumberCloseFriends by gender
## t = 2.4497, df = 1392.8, p-value = 0.01442
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.04536669 0.41004255
## sample estimates:
## mean in group female
                          mean in group male
##
               2.088561
                                    1.860856
```

(g) Are the technical conditions for the two-sample t-test satisfied here? Explain.

Yes. Even though the distribution of number of close friends is clearly skewed, the sample sizes (654 and 813) are quite large. Also, these data were collected from a random sample of adult Americans. Using the two-sample t-test is valid.

(h) Now conduct a Z-test of whether these sample data suggest that the proportion of Americans who say they have zero close friends differs between men and women. Report the hypotheses, test statistic, and p-value. State your test decision at the 0.05 significance level, and summarize your conclusion.

```
friends %>%
  group_by(gender) %>%
  summarize(count0 = count(NumberCloseFriends == 0), n())
## # A tibble: 2 x 3
##
     gender count0 `n()`
##
     <chr>
             <int> <int>
## 1 female
               201
                     813
## 2 male
               196
                     654
pOmen = 196/654
p0wo = 201/813
pOhat = (196 + 201) / (654 + 813)
(zscore = (p0wo - p0men) / sqrt(p0hat*(1-p0hat)*(1/813 + 1/654)) )
## [1] -2.248043
2* mosaic::xpnorm(zscore)
```



[1] 0.02457346

With a p-value of 0.0245, we have convincing evidence that the population proportion of American women with zero close friends is not the same as the population proportion of American males .

(i) Produce a 95% confidence interval for the difference in population proportions (who have zero close friends) between men and women. Also write a sentence or two interpreting what the interval reveals.



(zstar = mosaic::xqnorm(.975, 0, 1))

```
## [1] -0.00650361
```

The R output above reveals that a 95% confidence interval for $(p_{women} - p_{men})$ is: (-0.099, -0.007). This interval is entirely negative, which suggests that a smaller proportion of women than men have no close friends. We can be 95% confident that the difference in these population proportions is between 0.007 and 0.099.