Review Problems for Exam 1

These review problems are not meant to be comprehensive. Make sure to review the lessons and homework too!

Problem 1. Suppose we want to estimate the average number of hours of TV watched (per day) for all college students in the United States. We take a random sample of 50 college students, and we find that the sample mean is 2.5 hours with a sample variance of 1.5 hours.

a. State the general formula for a confidence interval about the mean for this scenario.

b. To form a 98% confidence interval for the population mean using the formula in part a, what R code would you use to find the critical value?

c. Use that code to calculate the appropriate critical value and the lower endpoint of the 98% CI.

d. What does "98% confident" mean in the context of this problem?

Problem 2.

a. Which are valid hypotheses when conducting a hypothesis test?

(i) $H_0 : \bar{x} = 5$ vs. $H_A : \bar{x} \neq 5$
(ii) $H_0 : \bar{x} \neq 5$ vs. $H_A : \bar{x} = 5$
(iii) $H_0 : \mu = 5$ vs. $H_A : \mu \neq 5$
(iv) $H_0 : \mu \neq 5$ vs. $H_A : \mu = 5$

b. Using the hypotheses above, if the sample mean is 4, the sample size is 30, and the sample standard deviation is 2, what is the test statistic?

c. State the `R` code you would use to calculate the corresponding p-value.

d. If H_0 is really false, but we fail to reject it, which kind of error is that?

Problem 3. The dataset `TeenPregnancy` in `Stat2Data` contains information about pregnancy rate for each of the 50 states in 2010 (from survey data). Suppose you want to predict state *Teen* pregnancy rate (number of pregnancies per 1000 teenage girls) from *Church* attendance (percentage who attended church in previous week, from state survey).

a. In order to choose a simple linear regression model, we need both variables to be quantitative. What is the other thing we need?

b. Make a scatterplot of the explanatory variable (x-axis) versus the response variable (y-axis). What is the approximate range of church attendance?

c. Fit a simple linear regression model predicting *Teen* pregnancy rate from *Church* attendance. State the fitted model.

d. Interpret the slope in context.

e. Comment on whether or not each condition of linear regression is met, and how you made that determination.

f. Based on our rule of thumb from class, which state(s) would be called very unusual outliers?

g. Based on our rule of thumb from class, which state(s) have very unusual leverage?

h. Based on our rule of thumb from class, which state(s) have very unusual Cook's distance?

i. Predict the teen pregnancy rate of a state whose survey reported 35% of female teens attended church in the previous week.

j. Predict the teen pregnancy rate of a state whose survey revealed 55% of female teens attended church in the previous week. *Hint.* This is a trick question...

k. Is the slope of the least squares regression line for predicting *Teen* from *Church* signficantly different from zero? Show details to support your answer. Use a signficance level of 0.05.

l. Construct and intepret a 90% confidence interval for the slope coefficient in your model.

m. What percentage of the variation in *Teen* is explained your model?

Solutions to Problem 1.

a. $\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$

b. `qt(1 - 0.02 / 2, 50 - 1)`

c. $2.5 - 2.405\sqrt{\frac{1.5}{50}} = 2.083$

d. In this setting, "98% confident" means that if we were to repeatedly take random samples of 50 college students in the US and construct the corresponding confidence intervals, 98% of the intervals would contain the true mean hours of TV watched per day for all college students in the US.

Solutions to Problem 2.

a. (iii) $H_0 : \mu = 5$ vs. $H_A : \mu \neq 5$

b. $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4 - 5}{2/\sqrt{30}} = -2.739$

c. `2 * pt(-2.739, 30 - 1)`

d. Type II error

Solutions to Problem 3.

Load the data with

```
library(Stat2Data)
data(TeenPregnancy)
```

The solutions below contain the R code without the output. Run the code in JupyterLab to see the corresponding output.

a. A consistent linear trend.

b.

```
plot(TeenPregnancy$Church, TeenPregnancy$Teen)
```

Range of the sample data variable *Church* is 17% to 51%.

c.

```
fit <- lm(Teen~Church, data=TeenPregnancy)
summary(fit)
```

Fitted model: $\widehat{Teen} = 28.8613 + 0.7921 Church$

d. Every one percent increase in church attendance is associated with an increase of 0.7921 teen pregnancies per 1000 teenage girls, on average.

e.

```
plot(fit, which = 1)  # residuals vs. fitted values
plot(fit, which = 2)  # Normal Q-Q plot of residuals
```

(a) Zero-mean: automatically MET by using least squares regression.

(b) Independence: no indication that a state's responses would depend on another state's responses, so probably okay; MET.

(c) Randomness: a random sample of teenage girls was chosen from each state; MET.

(d) Linearity: residuals vs. fitted plot looks evenly spread above and below 0 with no obvious pattern, although there are a few unusual points; MET.

(e) Constant variance: residuals vs. fitted plot looks like a consistent band, although there are a few unusual points; MET.

(f) Normality: normal Q-Q plot shows an approximately straight line; MET.

f.

```
plot(fit, which = 5)  # standardized residuals, leverage, Cook's distance
```

Utah (standardized residual > 3)

g. Utah ($h_{44} > 6/50 = 0.12$)

h. None (all $D_i < 1$)

i. $28.8613 + 0.7921(35) = 56.585$ pregnancies per 1000 teen girls

j. We should not use this model to make a prediction based on a church attendance percentage outside the range of the sample data.

k. The p-value of the t-test for SLR slope is 0.000159 is less than the given significance level of 0.05. Therefore, we reject H_0 for the t-test for SLR slope. We conclude that we have signfiicant evidence that the slope is different from zero. In other words, there is a statistically significant relationship between *Teen* and *Church*.

l.

```
confint(fit, level=0.90)
```

$(0.403545, 1.180677)$

m. $r^2 = 0.2593$. 25.9% of the variation in *Teen* is explained by this model.