MA256 Lesson 17 - Two Quantitative Variables (10.1, 10.3-10.5)

Example: Draft Lottery

In 1970, the United States Selective Service conducted a lottery to decide which young men would be drafted into the armed forces (Fienberg, 1971). Each of the 366 birthdays in a year (including February 29) was assigned a draft number. Young men born on days assigned low draft numbers were drafted.

We will regard the 366 dates of the year as observational units. We will consider two variables recorded on each date: draft number assigned to the date and sequential date in the year (so January 31 is sequential date 31, February 1 is sequential date 32, and so on).

A1) In a perfectly fair, random lottery, what should be the value of the correlation coefficient between draft number and sequential date of birthday?

A2) Use `ggplot()` to create a scatter plot.

```
# p.draft <- draft %>% XXXXX +
#   labs(x = "Sequential Date",
#        y= "Draft Number",
#        title = "Scatterplot of the draft numbers and sequential dates.")
#
# p.draft
```

A3) Does the scatterplot reveal much of an association between draft number and sequential date?

A4) Based on the scatterplot, guess the value of the correlation coefficient.

A5) Does it appear that this was a fair, random lottery?

It's difficult to see much of a pattern or association in the scatterplot, so it seems reasonable to conclude that this was a fair, random lottery with a correlation coefficient near zero. But let's dig a little deeper:

A6) Calculate the median draft number for all 12 months. Do you notice any pattern or trend in the median draft numbers over the course of the year? (Hint: Draw a picture to help visualize the data.)

```
# draft.tab <- draft %>% group_by(XXXXX) %>% summarise(med = XXXXX)
# draft.tab
#
# draft.tab %>% XXXXX +
#   scale_x_discrete(limits = factor(1:12)) +
#   labs(x = "Month",
#        y= "Draft Number",
#        title = "Scatterplot of the draft numbers and sequential dates.")
```

A7) Use the R function `cor()` to calculate the correlation coefficient. What does this number reveal? Is it consistent with the scatterplot? Or what you expected?

```
# cor(XXXXX, XXXXX)
```

A8) Let's think about how we would complete a test of significance for correlation. Suggest two possible explanations (hypotheses) which could have generated the value of the nonzero correlation coefficient.

A9) How could we go about determining whether random chance is a plausible explanation for the observed correlation value between sequential date and draft number? What is the statistic? How would you do this? How would you evaluate the strength of evidence?

A10) Simulate with 1000 replications. How many observations are more extreme than your statistic? What is your estimated p-value? Interpret the p-value: This is the probability of what, assuming what?

```
# set.seed(256)
# M <- 1000
# RES <- data.frame(res = rep(NA, M))
# mystat <- XXXXX
#
# for (i in 1:M){
#   draft.shuff <- sample(XXXXX)
#   RES$res[i] <- cor(XXXXX, XXXXX)
# }
#
# RES %>% ggplot(aes(x = XXXXX)) + geom_XXXXX(bins = 20)
#
# #estimate p-value
# sum(XXXXX) / M
```

A11) What conclusion would you draw from this p-value? Do you have strong evidence that the 1970 draft lottery was not conducted with a fair, random process? Explain the reasoning behind your conclusion.

Once they saw these results, statisticians were quick to point out that something fishy happened with the 1970 draft lottery. The irregularity can be attributed to improper mixing of the balls used in the lottery drawing process. (Balls with birthdays early in the year were placed in the bin first, and balls with birthdays late in the year were placed in the bin last. Without thorough mixing, balls with birthdays late in the year settled near the top of the bin and so tended to be selected earlier.) The mixing process was changed for the 1971 draft lottery (e.g., two bins, one for the draft numbers and one for the birthdays), for which the correlation coefficient turned out to be $r = 0.014$.

A12) Use your simulation results to approximate the p-value for the 1971 draft lottery. Is there any reason to suspect that this 1971 draft lottery was not conducted with a fair, random process? Explain the reasoning behind your conclusion. Also explain why you don't need to paste in the data from the 1971 lottery first.

```
# (put something useful here)
```

Changing gears... let try to model the question about the draft using a linear model.

B1) First, let's define a linear model. Where have you seen something like this before?

B2) Estimate the coefficients in R. To do this, use the `lm()` function and show the summary of the linear model using `summary()`. Use the coefficient estimates to write out the linear model. How do you interpret this model?

```
# draft.lm <- draft %>% lm(XXXXX)
# summary(draft.lm)
```

B3) Notice that the results also provdie the t-value and the p-value. When have we used the t-statistic before? What is the null/alternative hypotheses we are testing here? What is our conclusion?

B4) Plot the linear model on top of the figure object `p.draft` that you had saved above.

```
# p.draft + geom_XXXXX(method = "lm", se = FALSE)
```

B5) There are four validity conditions. Write them down. Does our data set meet these conditions?

1 -

2 -

3 -

4 -

B6) The coefficient of determination (R^2) is the *% of total variation in the response variable that is explained by the linear relationship with the explanatory variable. The equation for R^2 is given as:

$$R^2 = \frac{SSE(\bar{y}) - SSE(regressionline)}{SSE(\bar{y})}$$

Where $SSE(\bar{y})$ is the sum of the squared residuals from the horizontal line at the average value of the response variable ($\sum_{i=1}^{n}(\bar{y} - y_i)$) and $SSE(regressionline)$ is the sum of the squared residuals ($\sum_{i=1}^{n}(\hat{y}_i - y_i)$). OR you can simply square the correlation coefficient.

```
# ybar <- XXXXX
# SSEybar <- XXXXX
# SSEreg <- XXXXX
# r.sq <- XXXXX
# r.sq
#
# # OR
# XXXXX
```

B7) Use simulation to see how extreme our slope coefficient is with the actual data.

```
# set.seed(256)
# M <- 1000
# RES2 <- data.frame(res = rep(NA, M))
# mystat <- XXXXX  # beta_1
#
# for (i in 1:M){
#   draft$shuff <- sample(XXXXX)
#   draft.lm.shuff <- draft %>% lm(XXXXX)
#   RES2$res[i] <- summary(draft.lm.shuff)$coefficients["sequential_date", "Estimate"]
#   xx <- summary(draft.lm)
# }
#
# RES2 %>% ggplot(aes(x = res)) + geom_XXXX(bins = 20)
#
# #estimate p-value
# sum(abs(RES2$res) > abs(mystat)) / M
```

Used Hondas

The data in the file `UsedHondas` come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded are the car's age (calculated as 2006 minus year of manufacture) and price. Consider conducting an analysis to test whether the sample data provide strong evidence of an association between a car's price and age in the population in terms of the population slope.

```
# hondas <- read.table("https://www.isi-stats.com/isi/data/chap10/HondaAgePrice.txt", header = TRUE)
```

C1) Describe the population slope in the context of the study and assign a symbol to it.

C2) State the appropriate null and alternative hypotheses in terms of the population slope using the symbol used in part (a).

C3) Is a theory-based test appropriate? Why or why not?

C4) Use a theory-based approach to find a p-value to test the hypotheses stated in part (b).

```
# honda.lm <- hondas %>% XXXXX
# summary(honda.lm)
```

C5) Summarize your conclusion based on the p-value reported in part (d).