MA256 Lesson 16 - 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).

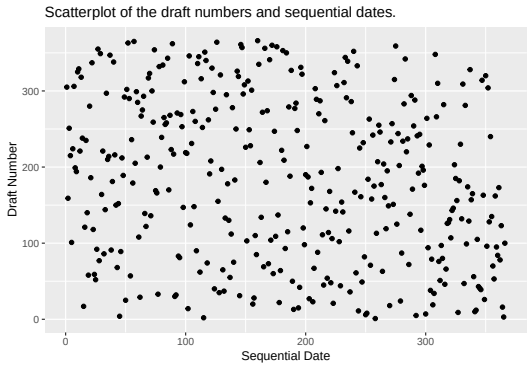
A1) In a perfectly fair, random lottery, what should be the value of the correlation coefficient between draft number and sequential date of birthday?

0 because we would want a completely fair random process with no correlation between sequential birth date and draft number.

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

```
p.draft <- draft %>% ggplot(aes(x = sequential_date, y = draft_number)) +
  geom_point() +
  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?

No, but there are too many observations to tell with the naked eye.

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

0 or very close to it.

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

Yes, it seems to be a fair random process.

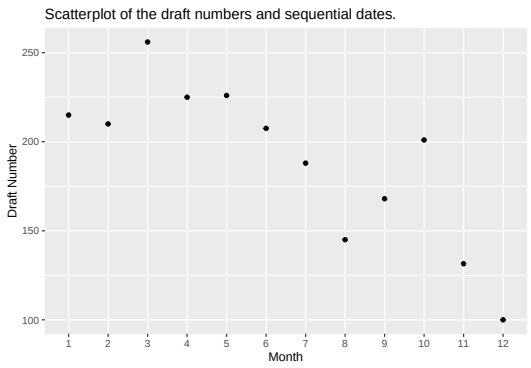
It's difficult to see much of a pattern or association in the scatterplot, so it seems reasonable to conclude that this was a fair, random lottery with a correlation coefficient near zero. But let's dig a little deeper:

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(month) %>% summarise(med = median(draft_number))
draft.tab
```

```
## # A tibble: 12 x 2
##    month   med
##    <dbl> <dbl>
##  1     1  215
##  2     2  210
##  3     3  256
##  4     4  225
##  5     5  226
##  6     6  208.
##  7     7  188
##  8     8  145
##  9     9  168
## 10    10  201
## 11    11  132.
## 12    12  100
```

```
draft.tab %>% ggplot(aes(x=month, y=med)) +
  geom_point() +
  scale_x_discrete(limits = factor(1:12)) +
  labs(x = "Month",
       y= "Draft Number",
       title = "Scatterplot of the draft numbers and sequential dates.")
```



The first 6 months' median draft numbers are much higher (in the 200's) compared to the last 6 months (mostly in 100's) so it seems that people born in the last 6 months will get drafted before people born in the first 6 months.

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(draft$sequential_date, draft$draft_number)
```

```
## [1] -0.2260414
```

The number reveals a negative correlation so as sequence birth dates increase, draft numbers decrease. The scatter plot made it seem like there was no correlation, but I see now that the top right and bottom left corners have fewer dots than the other 2 corners.

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.

One possible explanation will ALWAYS be that whatever was observed occurred randomly by-chance. The other possible explanation is that whatever was observed is due to an association (between sequence birth numbers and draft numbers).

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?

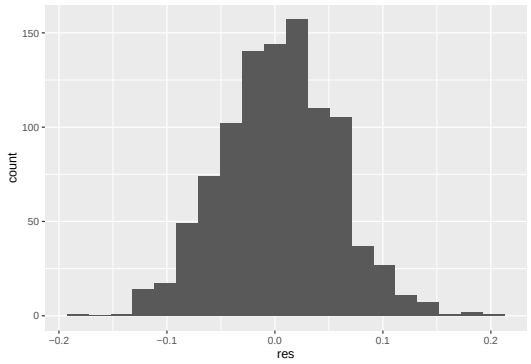
Simulate! The statistic is the correlation coefficient r = -0.226 (given in A7). Shuffle and randomly assign draft numbers to sequential dates. Calculate r for each of the 1000 simulations to create a dotplot of possible "random by-chance" values. Count the number of random by-chance simulations that produced an r value equal to or more extreme as -0.226. Then divide by 1000 to get the probability of seeing this r value assuming the null hypothesis is true.

A10) Simulate with 1000 replications. How many observations are more extreme than your statistic? What is your estimated p-value?

```
set.seed(256)
M <- 1000
RES <- data.frame(res = rep(NA, M))
mystat <- -0.2260414

for (i in 1:M){
  draft.shuff <- sample(draft$draft_number)
  RES$res[i] <- cor(draft$sequential_date, draft.shuff)
}

RES %>% ggplot(aes(x = res)) + geom_histogram(bins = 20)
```




```
#estimate p-value
sum(abs(RES$res) > abs(mystat)) / M
```

```
## [1] 0
```

0 obs. est. p-value = 0

Interpret the p-value: This is the probability of what, assuming what?

The probability of seeing $r = -0.226$ or more extreme assuming the null hypothesis is true.

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.

There is very strong evidence against the null hypothesis, meaning that the 1970 draft lottery was not fair given the correlation observed has an extremely low probability of occurring randomly.

Once they saw these results, statisticians were quick to point out that something fishy happened with the 1970 draft lottery. Th e 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 fi rst, 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.

```
sum(abs(RES$res) > abs(0.014)) / M
```

```
## [1] 0.805
```

p=.805 so there is no evidence against the null hypothesis. This draft was conducted fairly with an extremely high probability. The null distribution for correlation would be similar regardless of year.

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?

$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$
y is the dependent variable, x_i is the independent variable.
β_1 is the slope/slope coefficient
β_0 is the y-intercept coefficient
ϵ_i is residual for the ith observation.

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(draft_number ~ sequential_date,
     data = .)
summary(draft.lm)
```

```
##
## Call:
## lm(formula = draft_number ~ sequential_date, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -210.837  -85.629   -0.519   84.612  196.157
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     225.00922   10.81197  20.811  < 2e-16 ***
## sequential_date  -0.22606    0.05106  -4.427 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.2 on 364 degrees of freedom
## Multiple R-squared:  0.05109,    Adjusted R-squared:  0.04849
## F-statistic:  19.6 on 1 and 364 DF,  p-value: 1.264e-05
```

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(seqdate_i) = 225 - 0.22606 \cdot (seqdate_i)$
This model tells us that for every unit increase in the sequential date, we expect a decrease of 0.226 in the draft number.

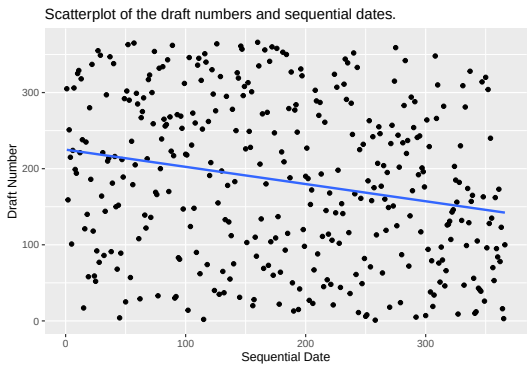
B3) Notice that the results also provdie the t-value and the p-value. When have we used the t-statistic before? What is the null/alternative hypotheses we are testing here? What is our conclusion?

We used the t-statistic when we had a single quantitative response variable and was comparing it to a (null) value/mean.
The null hypothesis is that the explanatory variable IS NOT associated with the response variable ($H_0 : \beta_1 = 0$);
The alternative hypothesis is that the explanatory variable IS associated with the response variable ($H_0 : \beta_1 \neq 0$)
We evidence to reject the idea that the draft number is not associated witht the sequential date.

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

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

```
## `geom_smooth()` using formula = 'y ~ x'
```



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

1 - Scatter plot shows linear trend

2 - Approximately Equal # Data Points Above & Below regression line (Balanced)

3 - Data points have approximately the same variance for all values of the explanatory variable (Symmetric).

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 <- mean(draft$draft_number)
SSEybar <- sum((draft$draft_number - ybar)^2)
SSEreg <- sum((draft.lm$fitted.values - draft$draft_number)^2)
r.sq <- (SSEybar - SSEreg) /  SSEybar
r.sq
```

```
## [1] 0.05109473
```

```
# OR
mystat^2
```

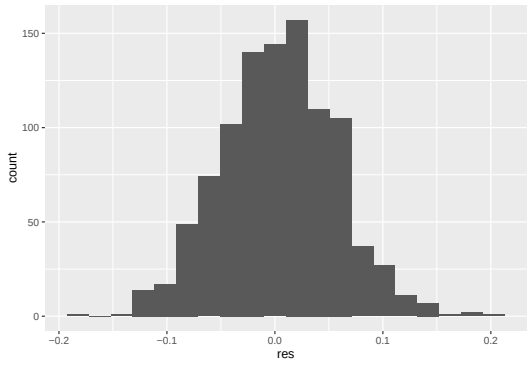
```
## [1] 0.05109471
```

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 <- -0.2260414  # beta_1

for (i in 1:M){
  draft$shuff <- sample(draft$draft_number)
  draft.lm.shuff <- draft %>% lm(shuff ~ sequential_date, data = .)
  RES2$res[i] <- summary(draft.lm.shuff)$coefficients["sequential_date", "Estimate"]
  xx <- summary(draft.lm)
}

RES2 %>% ggplot(aes(x = res)) + geom_histogram(bins = 20)
```




```
#estimate p-value
sum(abs(RES2$res) > abs(mystat)) / M
```

```
## [1] 0
```

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.

In the population of all Honda Civics, the slope tells how much the price changes on average for each year older the car is; β is the symbol for the population slope.

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

$H_0 : \beta = 0; H_a : \beta \neq 0$

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

Yes, the three validity conditions (linear trend, similar distributions, and equal spread around the line) are all reasonably well met for this dataset.

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

```
honda.lm <- hondas %>% lm(Price ~ age, data = .)
summary(honda.lm)
```

```
##
## Call:
## lm(formula = Price ~ age, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4200.5 -1418.2   -92.1   936.8  6362.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18785.3      431.0   43.59   <2e-16 ***
## age          -1397.5      103.9  -13.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2295 on 88 degrees of freedom
## Multiple R-squared:  0.6728, Adjusted R-squared:  0.6691
## F-statistic: 180.9 on 1 and 88 DF,  p-value: < 2.2e-16
```

$p - value = 0$

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

With a p-value of 0 we have very strong evidence against the null and in favor of the alternative that the population slope describing the association between the number of years before 2006 a Honda Civic was manufactured and the sale price is different from zero.