Correlation Coefficients

Checking assumptions

Jenine Harris
Brown School

Import and explore the data

```
# import the water data
water.educ <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/da

# examine the data
summary(object = water.educ)
```

```
##    country               med.age      perc.1dollar    perc.basic2015sani
##  Length:97          Min.   :15.00   Min.   : 1.00   Min.   :  7.00
##  Class :character   1st Qu.:22.50   1st Qu.: 1.00   1st Qu.: 73.00
##  Mode  :character   Median :29.70   Median : 1.65   Median : 93.00
##                     Mean   :30.33   Mean   :13.63   Mean   : 79.73
##                     3rd Qu.:39.00   3rd Qu.:17.12   3rd Qu.: 99.00
##                     Max.   :45.90   Max.   :83.80   Max.   :100.00
##                                     NA's   :33
##  perc.safe2015sani perc.basic2015water perc.safe2015water perc.in.school
##  Min.   :  9.00    Min.   : 19.00      Min.   : 11.00     Min.   :33.32
##  1st Qu.: 61.25    1st Qu.: 88.75      1st Qu.: 73.75     1st Qu.:83.24
##  Median : 76.50    Median : 97.00      Median : 94.00     Median :92.02
##  Mean   : 71.50    Mean   : 90.16      Mean   : 83.38     Mean   :87.02
##  3rd Qu.: 93.00    3rd Qu.:100.00      3rd Qu.: 98.00     3rd Qu.:95.81
##  Max.   :100.00    Max.   :100.00      Max.   :100.00     Max.   :99.44
##  NA's   :47        NA's   :1           NA's   :45
##  female.in.school male.in.school
##  Min.   :27.86    Min.   :38.66
##  1st Qu.:83.70    1st Qu.:82.68
##  Median :92.72    Median :91.50
##  Mean   :87.06    Mean   :87.00
```

Codebook

Definitions of the variables:

- country: the name of the country
- med.age: the median age of the citizens in the country
- perc.1dollar: percentage of citizens living on $1 per day or less
- perc.basic2015sani: percentage of citizens with basic sanitation access
- perc.safe2015sani: percentage of citizens with safe sanitation access
- perc.basic2015water: percentage of citizens with basic water access
- perc.safe2015water: percentage of citizens with safe water access
- perc.in.school: percentage of school-age people in primary and secondary school
- female.in.school: percentage of female school-age people in primary and secondary school
- male.in.school: percentage of male school-age people in primary and secondary school

The data were all from 2015.

Checking assumptions for Pearson's r correlation analyses

Correlation coefficients rely on four assumptions:

- Both variables are continuous
- Both variables are normally distributed
- The relationship between the two variables is *linear* (linearity)
- The variance is constant with the points distributed equally around the line (homoscedasticity)

Checking the normality assumption

- Started by using histograms to check the normality assumption.

```
# check normality of female.in.school variable
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  ggplot(aes(x = female.in.school)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
  labs(x = "Percent of school-aged females in school",
       y = "Number of countries",
       title = "Distribution of percentage of school-aged females\nin sc
```

Checking normality with a Q-Q plot

- The normality assumption was violated for `female.in.school`, but might be OK for `perc.basic2015water`.

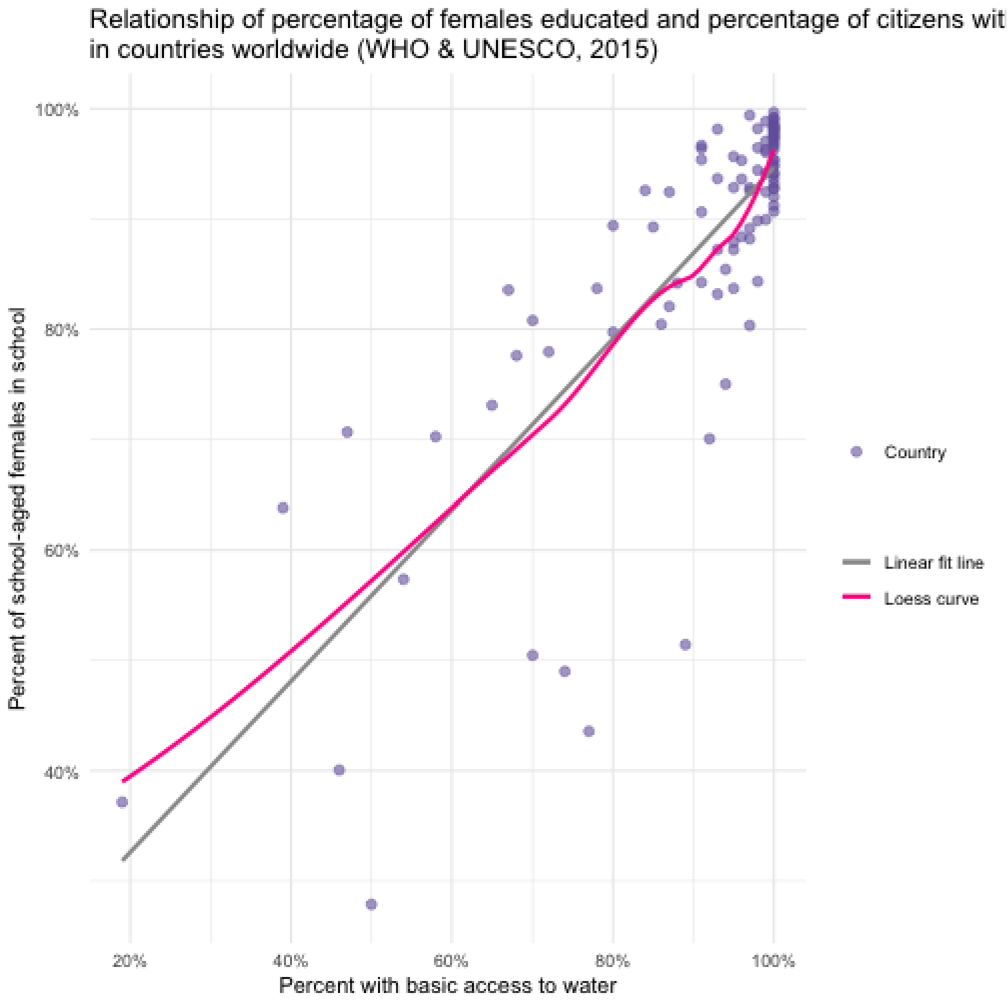
```
# Q-Q plot of water access variable to check normality
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  ggplot(aes(sample = perc.basic2015water)) +
  stat_qq(aes(color = "Country"), alpha = .6) +
  geom_abline(aes(intercept = mean(x = perc.basic2015water),
                  slope = sd(x = perc.basic2015water),
                  linetype = "Normally distributed"),
              color = "gray60", size = 1) +
  theme_minimal() +
  labs(x = "Theoretical normal distribution",
       y = "Observed values of percent of people\nwith basic water acces
       title = "Distribution of percentage of citizens\nwith basic water
  ylim(0,100) +
  scale_linetype_manual(values = 1, name = "") +
  scale_color_manual(values = "#7463AC", name = "")
```

Checking the linearity assumption

- The linearity assumption requires that the relationship between the two variables falls along a line.
- The assumption is met if a scatterplot of the two variables shows that the relationship that falls along a line.
- If it is difficult to tell, a **Loess curve** can be added to confirm linearity.
- A Loess curve shows the actual relationship between the two variables without constraining the line to be straight like the linear model `method = lm` option does.

```
# female education and water graph with linear fit line and Loess curve
water.educ %>%
  ggplot(aes(y = female.in.school/100, x = perc.basic2015water/100)) +
  geom_point(aes(size = "Country"), color = "#7463AC", alpha = .6) +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE)
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  theme_minimal() +
  labs(y = "Percent of school-aged females in school",
       x = "Percent with basic access to water",
       title = "Relationship of percentage of females educated and perce
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent) +
```

The Loess curve



What do non-linear relationships look like?

Checking the homoscedasticity assumption

- The final assumption is the equal distribution of points around the line, which is often called the assumption of homoscedasticity.
- Examine the pattern of data points around the line.
- The funnel shape of the data indicated that the points were not evenly spread around the line from right to left.
- On the left of the graph they were more spread out than on the right, where they were very close to the line.
- This indicates the data do not meet this assumption.

Statistical test of constant variance

- The Breusch-Pagan test can be used to test the null hypothesis that *the variance is constant* around the line.
- The Breusch-Pagan test relies on the chi-squared distribution.
- The `bptest()` function from the lmtest package can be used to test this null hypothesis.

```
# Breusch-Pagan test for equal variance
testVar <- lmtest::bptest(formula = water.educ$female.in.school ~ water.
testVar
```

```
##
##      studentized Breusch-Pagan test
##
## data:  water.educ$female.in.school ~ water.educ$perc.basic2015water
## BP = 12.368, df = 1, p-value = 0.0004368
```

Interpreting the Breusch-Pagan test

- The Breusch-Pagan test statistic has a low p-value ($BP = 12.37$; $p = 0.0004$), indicating that the null hypothesis that the variance is constant would be rejected.
- When the null hypothesis that the variance is constant is rejected, the assumption of constant variance is *not met*.
- This is consistent with the graph given the difference in spread around the line at the lower and higher ends of the graph.

Interpreting the assumption checking results

- In all, the correlation analysis for female education and water access met two of the four assumptions.
- It failed the assumption of normally distributed variables and the assumption of homoscedasticity but met the variable type assumption and the linearity assumption.
- There are a few options for what they could do with these results:
    - (1) report the results and explain that the analysis does not meet assumptions, so that it is unclear if what is happening in the sample is a good reflection of what is happening in the population;
    - (2) transform the two variables to meet the assumptions for Pearson's r and conduct the analysis again; and
    - (3) choose a different type of analysis with assumptions that can be met by these data.