Correlation Coefficients

Exploratory data analysis

Jenine Harris
Brown School

Exploring the data

- Importing the data using the `here()` function

```
# 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
```

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.

Understanding the data

- The data frame in the Environment pane shows 97 countries and 10 variables.
- Except for `country`, all of the variables appeared to be numeric.
- Compute the mean and standard deviation for the two variables of interest `female.in.school` and `perc.basic2015water`.

```
# open the tidyverse
library(package = "tidyverse")

# descriptive statistics for females in school and water access
water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  summarize(m.f.educ = mean(x = female.in.school),
            sd.f.educ = sd(x = female.in.school),
            m.bas.water = mean(x = perc.basic2015water),
            sd.bas.water = sd(x = perc.basic2015water))
```

```
##   m.f.educ sd.f.educ m.bas.water sd.bas.water
## 1 87.01123   15.1695    90.15625     15.81693
```

Interpret the descriptive stats

The mean percent of school-aged females in school was 87.06 (sd = 15.1) and the mean percent of citizens who had basic access to water was 90.16 (sd = 15.82).

Plotting the variables

```
# plot females in school and water access
plot.fem.sch <- water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  ggplot(aes(x = female.in.school)) +
  geom_histogram()

plot.water <- water.educ %>%
  drop_na(female.in.school) %>%
  drop_na(perc.basic2015water) %>%
  ggplot(aes(x = perc.basic2015water)) +
  geom_histogram()

gridExtra::grid.arrange(plot.fem.sch, plot.water, ncol = 2)
```

Make a scatterplot to examine the relationship

- Create scatterplots with new layers for `ggplot()` to change the scale on the x and y axes
- The `scale_x_continuous()` and `scale_y_continuous()` layers with the `label =` argument can be used to change the scale on the x-axis and y-axis so that it shows percentages.
- To use these scales, divide the percent variables by 100 in the `aes()` in order to get a decimal version of the percentages for use with the `labels = scales::percent` option.

```
# explore plot of female education and water access
water.educ %>%
  ggplot(aes(y = female.in.school/100, x = perc.basic2015water/100)) +
  geom_point(aes(color = "Country"), size = 2, alpha = .6) +
  theme_minimal() +
  labs(y = "Percent of school-aged females in school",
       x = "Percent with basic water access",
       title = "Relationship of percentage of females in school\nand per
  scale_color_manual(values = "#7463AC", name = "") +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent)
```

Make a scatterplot to examine the relationship

- Create scatterplots with new layers for `ggplot()` to change the scale on the x and y axes
- The `scale_x_continuous()` and `scale_y_continuous()` layers with the `label =` argument can be used to change the scale on the x-axis and y-axis so that it shows percentages.
- To use these scales, divide the percent variables by 100 in the `aes()` in order to get a decimal version of the percentages for use with the `labels = scales::percent` option.