Correlation Coefficients

NHST for correlation

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.

NHST Step 1: writing the null and alternate hypotheses

H0: There is no relationship between the two variables ($r = 0$)

HA: There is a relationship between the two variables ($r \neq 0$)

NHST Step 2: Computing the test statistic

- The null hypothesis is tested using a t-statistic comparing the correlation coefficient of r to a hypothesized value of zero, like the one-sample t-test.
- The one sample t-test t-statistic formula where m_x is the mean of x and se_{m_x} is the standard error of the mean of x:

$$t = \frac{m_x - 0}{se_{m_x}}$$

- Since r is not the same as a mean, here is the revised equation:

$$t = \frac{r_{xy}}{se_{r_{xy}}}$$

Substituting the se into the equation for t

Equation to use to compute the t-statistic for the significance test of r (replaced se with formula for se):

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

- Use of this formula requires r_{xy} and n.
- The correlation between water access and female education is 0.81, but it is unclear what the value of n is for this correlation.
- While the overall data frame has 97 observations, some of these have missing values. To find the n for the correlation between `perc.1dollar` and `female.in.school`, use `drop_na()` and adding `n()` to `summarize()` to count the number of cases after dropping the missing `NA` values.

```
# correlation between water access and female education
water.educ %>%
  drop_na(perc.basic2015water) %>%
  drop_na(female.in.school) %>%
  summarize(cor.females.water = cor(x = perc.basic2015water,
                                    y = female.in.school),
```

Computing t by hand & with R

$$ \begin{equation} t=\frac{r*{xy}\sqrt{n-2}}{\sqrt{1-r}{xy}^2}}=\frac{.8086651\sqrt{96-2}}{\sqrt{1-(.8086651)^2}} = 13.33 \end{equation} $$

- The t-statistic was 13.33.
- R code:

```
# test for correlation coefficient
cor.test(x = water.educ$perc.basic2015water,
         y = water.educ$female.in.school)
```

```
##
##     Pearson's product-moment correlation
##
## data:  water.educ$perc.basic2015water and water.educ$female.in.school
## t = 13.328, df = 94, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7258599 0.8683663
## sample estimates:
##       cor
## 0.8086651
```

NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)

- Although `cor.test()` prints out a p-value, examine the probability distribution used to convert the test statistic into a p-value.
- This t-statistic was for a situation where there were *two* variables involved even though the r is a single statistic; with two variables involved, two is subtracted from the sample size for a d.f. of 94.

NHST Steps 4 & 5: Reject or retain the null hypothesis based on the p-value

- The p-value was very tiny, well under .05.
- This p-value is the probability that the very strong positive relationship (r = .81) observed between percentage of females in school and percentage with basic water access would have happened if the null were true.
- It is extremely unlikely that this correlation would happen in the sample if there were not a very strong positive correlation between females in school and access to water in the population that this sample came from.
- The 95% confidence interval is the confidence interval around r, so the value of r in the sample is .81 and the likely value of r in the population that this sample came from is somewhere between .7258599 and .8683663.
- Interpretation: The percentage of people basic access to water is statistically significantly, positively, and very strongly correlated with the percentage of primary and secondary age females in school in a country ($r = .81$; $t(94) = 13.33$; $p < .05$). As the percentage of people living with basic access to water goes up, the percentage of females with education also goes up. While