Correlation Coefficients

Effect size for correlation

Jenine Harris
Brown School

Exploring 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.

Effect size for Pearson's r

- The correlation coefficient is considered its own effect size since it measures the strength of the relationship.
- There is also another value that is easy to calculate and has a more direct interpretation to use as an effect size with r.
- This metric is the **coefficient of determination**, which is the percentage of the variance in one variable that is shared, or explained, by the other variable.
- The notation for the coefficient of determination is r^2

Calculating the coefficient of determination

- There are several ways to compute the coefficient of determination.
- For a Pearson's r correlation coefficient, the coefficient of determination can be computed by squaring the correlation coefficient:
- $r_{xy}^2 = \left(\frac{cov_{xy}}{s_x s_y} \right)^2$

Using R to calculate the coefficient of determination

- The coefficient of determination is often referred to just as r-squared and reported as r^2 or more commonly, R^2.
- There is no specific R command for computing the coefficient of determination directly from the data, but there are many options for computing it from the output of a correlation analysis.
- The most straightforward way might be to use `cor()` and square the result, but it is also possible to use `cor.test()` and square the correlation from the output of this procedure.

```
# conduct the correlation analysis
# assign the results to an object
cor.Fem.Educ.Water <- cor.test(x = water.educ$perc.basic2015water,
                               y = water.educ$female.in.school)
```

Using R to calculate the coefficient of determination

```
# explore the object
str(cor.Fem.Educ.Water)
```

```
## List of 9
##  $ statistic  : Named num 13.3
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named int 94
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 2.21e-23
##  $ estimate   : Named num 0.809
##   ..- attr(*, "names")= chr "cor"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "correlation"
##  $ alternative: chr "two.sided"
##  $ method     : chr "Pearson's product-moment correlation"
##  $ data.name  : chr "water.educ$perc.basic2015water and water.educ$female.ir
##  $ conf.int   : num [1:2] 0.726 0.868
##   ..- attr(*, "conf.level")= num 0.95
##  - attr(*, "class")= chr "htest"
```

```
# square the correlation coefficient
r.squared <- cor.Fem.Educ.Water$estimate^2
r.squared
```

Interpreting R-squared

- The result 0.65 can be multiplied by 100 to find that `female.in.school` and `perc.basic2015water` have 65.39% shared variance.