Correlation Coefficients

Partial correlations

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.

Introducing partial correlations

- It is possible that females in school and water access might both be related to poverty and that poverty might be the reason both of these variables increase at the same time.
- Perhaps poverty is the reason for the shared variance between these two variables.
- Perhaps countries with higher poverty have fewer females in school and lower percentages of people with basic water access.
- *Partial correlation* examines how multiple variable share variance with each other.

Partial correlation as shared variance

- A Venn Diagram can help visualize shared variance.
- There are two ways the variables overlap.
    - There are places where *just two* of the variables overlap (X and Y overlap, X and Z overlap, Y and Z overlap) and there is where X and Y and Z all overlap in the center of the diagram.
- The overlap between *just two* colors is the **partial correlation** between the two variables.
- It is the extent to which they vary in the same way after accounting for how they are both related to the third variable involved.

Computing Pearson's r partial correlations

- The ppcor package can be used for partial correlation.
- Using the function for partial correlation (`pcor()`) and for the partial correlation statistical test (`pcor.test()`) requires having a small data frame that consists only of the variables involved in the correlation with no missing data.
- Create this data frame including the females in school, basic water access, and poverty variables.

```
# create a data frame with only female education
# poverty and water access
water.educ.small <- water.educ.new %>%
  select(female.in.school, perc.basic2015water, perc.1dollar) %>%
  drop_na()

# check the new data
summary(water.educ.small)
```

```
##  female.in.school perc.basic2015water  perc.1dollar
##  Min.   :40.05    Min.   : 39.00      Min.   : 1.00
##  1st Qu.:81.74    1st Qu.: 85.75      1st Qu.: 1.00
##  Median :92.45    Median : 95.50      Median : 1.65
##  Mean   :86.52    Mean   : 88.88      Mean   :13.63
##  3rd Qu.:96.28    3rd Qu.:100.00      3rd Qu.:17.12
```

Check the Pearson's r in small data frame

```
# examine the bivariate correlations
water.educ.small %>%
  summarize(corr.fem.water = cor(x = perc.basic2015water, y = female.in.
            corr.fem.pov = cor(x = perc.1dollar, y = female.in.school),
            corr.water.pov = cor(x = perc.basic2015water, y = perc.1doll
```

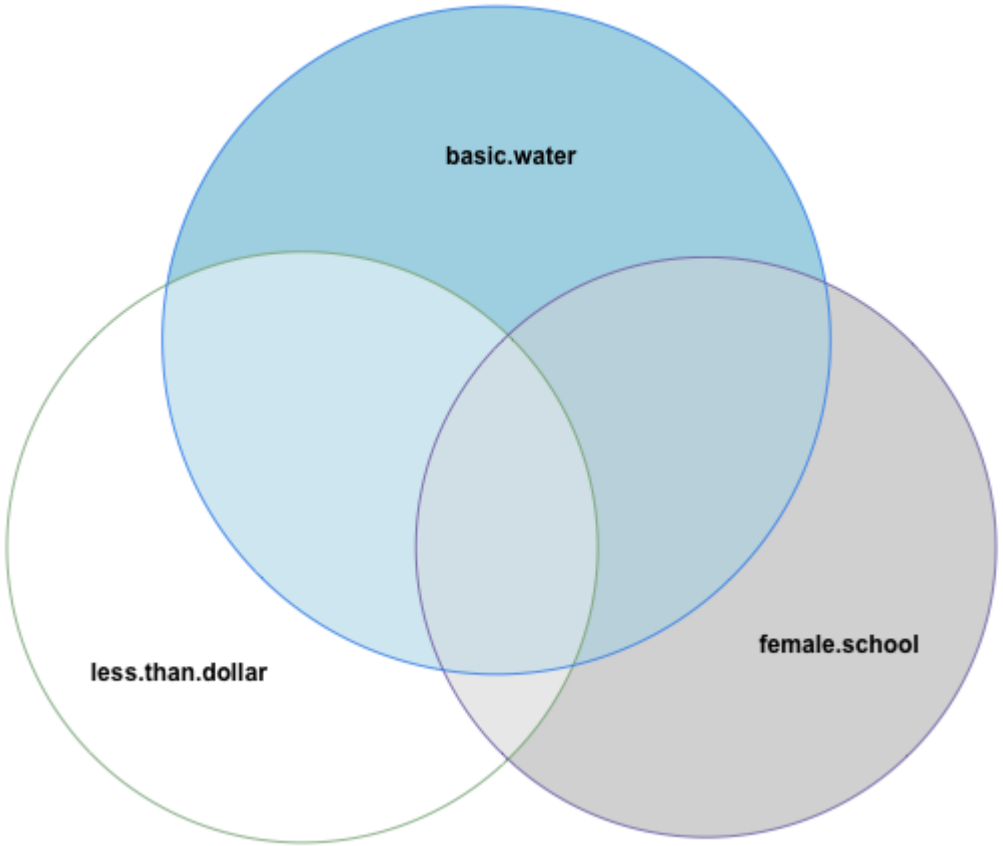
```
##   corr.fem.water corr.fem.pov corr.water.pov
## 1      0.7650656   -0.7144238     -0.8320895
```

Conduct the partial correlation analysis

```
# conduct partial Pearson correlation
educ.water.poverty <- ppcor::pcor(x = water.educ.small, method = "pearso
educ.water.poverty
```

```
## $estimate
##                     female.in.school perc.basic2015water perc.1dollar
## female.in.school           1.0000000           0.4395917   -0.2178859
## perc.basic2015water        0.4395917           1.0000000   -0.6336436
## perc.1dollar              -0.2178859          -0.6336436    1.0000000
##
## $p.value
##                     female.in.school perc.basic2015water perc.1dollar
## female.in.school       0.0000000000       3.125684e-04 8.626064e-02
## perc.basic2015water    0.0003125684       0.000000e+00 2.490386e-08
## perc.1dollar           0.0862606413       2.490386e-08 0.000000e+00
##
## $statistic
##                     female.in.school perc.basic2015water perc.1dollar
## female.in.school           0.000000           3.822455   -1.743636
## perc.basic2015water        3.822455           0.000000   -6.397046
## perc.1dollar              -1.743636          -6.397046    0.000000
##
## $n
## [1] 64
```

Visualizing the partial correlation



Computing Spearman's rho partial correlations

- The data do not meet the assumptions for the Pearson's r correlation.
- The assumptions that applied to the two variables for a Pearson's r correlation would apply to all three variables for a partial Pearson's r correlation.
- Each variable would be continuous and normally distributed, each pair of variables would demonstrate linearity, and each pair would have to have constant variances (homoscedasticity).
- Since several assumptions were not met, compute the Spearman correlation, which is more appropriate in this case.
- The Spearman assumption of monotonic relationship would apply to each pair of variables.

```
# conduct partial correlation with Spearman
educ.water.poverty.spear <- ppcor::pcor(x = water.educ.small, method = "
educ.water.poverty.spear
```

```
## $estimate
##                     female.in.school perc.basic2015water perc.1dollar
## female.in.school           1.0000000           0.4305931   -0.2841782
## perc.basic2015water        0.4305931           1.0000000   -0.5977239
## perc.1dollar              -0.2841782          -0.5977239    1.0000000
##
```

Significance testing for partial correlations

- The original r_s between female education and water access was 0.77 but the partial Spearman's r_s correlation between females in school and water access after accounting for poverty was .43.
    - Including poverty reduced the magnitude of the correlation by nearly half.
- Like the r and r_s correlations, the partial correlations can be tested for statistical significance using a t-test.
- The t-statistic for each partial correlation is shown in the output from `pcor()`.
- The second chunk of numbers are the p-values and the third chunk of numbers are the t-test test statistics.
- Interpretation: The partial correlation between percentage of females in school and the percentage of citizens who have basic water access was moderate, positive, and statistically significant ($rho_{partial}$ = 0.43; t = 3.73; p < .05). Even after poverty is accounted for, increased basic water access was moderately, positively, and significantly associated with an increased percentage of females in school.

Checking assumptions for partial correlations

- The variables all meet the assumption of being at least ordinal we have already checked for a monotonic relationship between females in school and percentage with basic water.
- Checked the monotonic assumption for the females in school with poverty and for percentage with basic water and poverty.

Interpreting results when assumptions are not met

- When assumptions are not met, there are a few possible strategies, including:
  - (1) interpreting the results for the sample only, and
  - (2) recoding one of the variables to be categorical and using a different type of analysis.
- Interpretation: The partial correlation between percentage of females in school and the percentage of citizens who have basic water access was moderate, positive, and statistically significant ($rho_{partial}$ = 0.43; t = 3.73; p < .05). Even after poverty is accounted for, increased basic water access was moderately and positively associated with an increased percentage of females in school. The assumptions were not met, so it is not clear that the partial correlation from the sample of countries can be generalized to the population of all countries.