Correlation Coefficients

Spearman’s rho

Jenine Harris
Brown School

foundations in public health
biostatistics

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.

Spearman's rho when Pearson's r assumptions are not met

- There are other correlation statistics that do not have the same strict assumptions.
- The most commonly used alternative to the Pearson's r correlation coefficient is the Spearman's rho rank correlation coefficient.
- Using Spearman's rho is just using another transformation, but instead of computing the arcsine or raising the variables to a power, the values of the variables are transformed into ranks, like with some of the alternatives to the t-tests.
- The values of the variables are ranked from lowest to highest and the calculations for correlation are conducted using the ranks instead of the raw values for the variables.

Computing Spearman's rho correlation coefficient

- Spearman's rho is computed by ranking each value for each variable from lowest to highest and then computing the extent to which the two variable ranks are the same.
- Leslie remembered that most of the time the Greek letters like rho are used to represent the population and there is some other way to represent the sample.
- rho is usually denoted "rho" or " r_s "
- r_s for females in school and water access would be computed by first ranking the values of percentage of females in school from lowest to highest and then ranking the values of water access from lowest to highest.
- Then, once the ranks were assigned, the r_s correlation coefficient is computed:
    - $\rho = \frac{6\sum d^2}{n(n^2-1)}$
- Where d is the difference between the ranks of the two variables and n is the number of observations

NHST Step 1: Write the null and alternate hypotheses

H0: There is no correlation between the percentage of females in school and the percentage of citizens with basic water access ($\rho = 0$)

HA: There is a correlation between the percentage of females in school and the percentage of citizens with basic water access ($\rho \neq 0$)

NHST Step 2: Compute the test statistic

- Using the `cor.test()` function, test the null hypothesis of no correlation between females in school and basic water access by adding `method = spearman` as one of the options.

```
# spearman correlation female education and water access
spear.fem.water <- cor.test(x = water.educ$perc.basic2015water,
                            y = water.educ$female.in.school,
                            method = "spearman")
spear.fem.water
```

```
##
##      Spearman's rank correlation rho
##
## data:  water.educ$perc.basic2015water and water.educ$female.in.school
## S = 34050, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7690601
```

- While Pearson's r between female education and basic water access was 0.81, r_s was slightly lower at 0.77.

rho test statistic

- Instead of a t-statistic, the output for r_s reports the S test statistic.
  - $S = (n^3 - n)\frac{1-r_p}{6}$
- Where r_p is the Pearson correlation coefficient and n is the sample size.
- The p-value in the output of the `cor.test()` function is not from the S test statistic.
  - Instead it is determined by computing an approximation of the t-statistic and degrees of freedom.
  - This special approximation of the t-statistic is computed $t_s = r\sqrt{\frac{n-2}{1-r^2}}$

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

- In this case, t is 13.33 with 94 degrees of freedom (n = 96).
- A plot of the t-distribution with 94 degrees of freedom revealed that the probability of a t-statistic this big or bigger would be very tiny if the null hypothesis were true.
- The p-value in the output for the Spearman analysis is very tiny makes sense given this distribution.

Assumption checking for Spearman's rho

The assumptions for r_s are:

- The variables must be at least ordinal or even closer to continuous
- the relationship between the two variables must be **monotonic**

The first assumption is met.

Checking the monotonic assumption

- A **monotonic** relationship is a relationship that only goes in a single direction.
- The relationship does not have to follow a straight line, it can curve as long as it is always heading in the same direction.

Examining the relationship graphically

```
water.educ %>%
  ggplot(aes(y = female.in.school, x = perc.basic2015water)) +
  geom_smooth(aes(color = "linear fit line"), method = "lm", se = FALSE)
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  geom_point(aes(size = "Country"), color = "#7463AC", alpha = .6) +
  theme_minimal(base_size = 14) +
  labs(y = "Percent of females in school",
       x = "Percent with basic water access",
       title = "Females in school and water\naccess (WHO & UNESCO, 2015)
  scale_color_manual(name="Type of fit line", values=c("gray60", "deeppi
  scale_size_manual(values = 2)
```

Interpreting the results

- For the female education and water access analysis, the Loess curve only goes up, which demonstrates that the relationship between females in school and water access meets the monotonic assumption.

- The values of females in school consistently goes up while the values of access to water go up.

- The relationship does not change direction.

- The best option for this analysis was to report and interpret r_s since the assumptions are met, while the assumptions failed for Pearson's r with the original data and with the transformed variables.

    - Interpretation: There is a statistically significant positive correlation between basic access to drinking water and female education ($r_s = 0.77$; $p < .001$). As the percentage of the population with basic access to water increases, so does the percentage of females in school. The data meet the monotonic relationship and variable type assumptions.