# Correlation Coefficients

## Transforming variables for unmet assumptions

**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.

# Checking assumptions for Pearson's r correlation analyses

Correlation coefficients rely on four assumptions:

- Both variables are continuous
- Both variables are normally distributed
- The relationship between the two variables is *linear* (linearity)
- The variance is constant with the points distributed equally around the line (homoscedasticity)

# Transforming the variables as an alternative when Pearson's r correlation assumptions are not met

- One of the ways to deal with data that do not meet assumptions for Pearson's r is to use a data transformation and examine the relationship between the transformed variables.
- There are two types of transformations:
  - (1) **Linear transformations** keep existing linear relationships between variables, often by multiplying or dividing one or both of the variables by some amount
  - (2) **Nonlinear transformations** increase (or decrease) the linear relationship between two variables by applying an exponent (i.e., **power transformation**) or other function to one or both of the variables
- Different transformations are appropriate in different settings.

# The logit transformation

- For variables that are percentages or proportions, a **logit transformation** or **arcsine transformation** is often used to account for the floor and ceiling effects.
- The **logit transformation** uses the equation below to make percentage data more normally distributed.

$$y_{logit} = log(\frac{y}{1-y})‘$$

- y is a percent ranging from 0 to 1.
- The arcsine transformation is also used to normalize percentage or proportion data to transform the variable y.

$$y_{arcsine} = arcsin(\sqrt{y})‘$$

# Transforming the variables in R

- The function for arcsine is `asin()`.
- Use `mutate()` to add new transformed variables to the data frame, `logit.female.school`, `logit.perc.basic.water`, `arcsin.female.school`, and `arcsin.perc.basic.water`.

```
# create new variables
water.educ.new <- water.educ %>%
  mutate(logit.female.school = log(x = (female.in.school/100)/(1-female.
  mutate(logit.perc.basic.water = log(x = (perc.basic2015water/100)/(1-p
  mutate(arcsin.female.school = asin(x = sqrt(female.in.school/100))) %>
  mutate(arcsin.perc.basic.water = asin(x = sqrt(perc.basic2015water/100

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

```
##    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
```

# Folded power transformation

- Folded power transformations are suggested by Tukey.
  - $y_{folded.power} = y^{\frac{1}{p}} - (1-y)^{\frac{1}{p}}$
- The p in the formula is for the power to raise it to.
- The rcompanion package can be used to choose the value of p.

```
# use Tukey transformation to get power for transforming
# female in school variable to more normal distribution
p.female <- rcompanion::transformTukey(x = water.educ$female.in.school,
                           plotit = FALSE,
                           quiet = TRUE,
                           returnLambda = TRUE)
p.female
```

```
## lambda
##   8.85
```

```
# use Tukey transformation to get power for transforming
# basic2015 water variable to more normal distribution
p.water <- rcompanion::transformTukey(x = water.educ$perc.basic2015water
                           plotit = FALSE,
                           quiet = TRUE,
                           returnLambda = TRUE)
p.water
```

# Folded power transformation in R

- The best value for p, which is called lambda ( $\lambda$ ) by the package, is 8.85 for the female in school variable and 9.975 for the water variable.

```
# create new transformation variables
water.educ.new <- water.educ %>%
  mutate(arcsin.female.school = asin(x = sqrt(female.in.school/100))) %>
  mutate(arcsin.perc.basic.water = asin(x = sqrt(perc.basic2015water/100
  mutate(folded.p.female.school = (female.in.school/100)^(1/p.female) -
  mutate(folded.p.basic.water = (perc.basic2015water/100)^(1/p.water) -

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

```
##    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
```

# Examine normality for arcsine transformed variable

```
# histogram of arcsin females in school
water.educ.new %>%
  ggplot(aes(x = arcsin.female.school)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 14) +
  labs(x = "Arcsine transformation of females in school",
       y = "Number of countries",
       title = "Distribution of arcsine transformed percentage of\nfemal
```

# Examine normality for folded power transformed variable

```
# histogram of folded power transf females in school
water.educ.new %>%
  ggplot(aes(x = folded.p.female.school)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 14) +
  labs(x = "Folded power transformation of females in school",
       y = "Number of countries",
       title = "Distribution of folded power transformed percentage\nof
```

# Histogram of variable transformed with arcsin()

```
# histogram of arcsine of water variable
water.educ.new %>%
  ggplot(aes(x = arcsin.perc.basic.water)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 14) +
  labs(x = "Arcsine transformed basic water access",
       y = "Number of countries",
       title = "Distribution of arcsine transformed basic\nwater access
```

# Histogram of variable transformed with folded power

```
# histogram of folded power transformed water variable
water.educ.new %>%
  ggplot(aes(x = folded.p.basic.water)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 14) +
  labs(x = "Folded power transformed basic water access",
       y = "Number of countries",
       title = "Distribution of folded power transformed\npercentage of
```

# NHST Step 1: Write the null and alternate hypotheses

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

HA: There is no correlation between the transformed values of percentage of females in school and percentage of citizens with basic water access ($r \neq 0$)

# NHST Step 2: Compute the test statistic

- The `cor.test()` function is then used with the transformed variables:

```
# correlation test for transformed variables
cor.test(water.educ.new$folded.p.female.school,
         water.educ.new$folded.p.basic.water)
```

```
##
##     Pearson's product-moment correlation
##
## data:  water.educ.new$folded.p.female.school and water.educ.new$folded.p.bas
## t = 8.8212, df = 94, p-value = 5.893e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5461788 0.7696207
## sample estimates:
##       cor
## 0.6729733
```

- The test statistic is t = 8.82 for the correlation of r = .67 between the two transformed variables.

# 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)

The p-value shown in the output of `cor.test()` is very tiny. The probability that the t-statistic would be 8.82 or larger if there were no relationship is very tiny, nearly zero.

# NHST Step 4 & 5: Reject or retain the null hypothesis based on the probability

- With a very tiny p-value, reject the null hypothesis.
- There was a statistically significant relationship between the transformed variables for percentage of females in school and percentage of citizens with basic water access in a country. The relationship was positive and moderate to strong (r = .67).
- As the percentage of citizens with basic water access goes up, the percentage of females in school also goes up. The correlation is .67 in the sample and the 95% confidence interval shows that it is likely between .55 and .77 in the sampled population.

# Testing assumptions for Pearson's r between the transformed variables

The four assumptions to test with the transformed variables:

- Both variables are continuous
- Both variables are normally distributed
- The relationship between the two variables is *linear* (linearity)
- The variance is constant with the points distributed equally around the line (homoscedasticity)
- The first assumption is *met*; the transformations resulted in continuous variables.
- The second assumption of normal distributions was *not met* based on the left-skewed histogram of the transformed water variable examined during data transformation.
- To test the third and fourth assumptions, make a scatterplot with the Loess curve and the linear model line to check linearity and homoscedasticity.

# Testing linearity and constant variance assumptions

```
# explore plot of transformed female education and water
# female education and water graph with linear fit line and Loess curve
water.educ.new %>%
  ggplot(aes(y = folded.p.female.school, x = folded.p.basic.water)) +
  geom_smooth(aes(color = "liner 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 = "Power transformed percent of females in school",
       x = "Power transformed percent with basic water access",
       title = "Transformed females in school and water\naccess variable
  scale_color_manual(name="Type of fit line", values=c("gray60", "deeppi
  scale_size_manual(values = 2)
```

# Testing linearity and constant variance assumptions

# Using Breusch-Pagan to test homoscedasticity

- The plot shows a pretty terrible deviation from linearity, which looks like it is mostly due to all the countries with 100% basic water access.
- The homoscedasticity looks better, but use Breusch-Pagan (BP) just for practice and to determine if this spread is considered equal around the line.
- The BP test is testing the null hypothesis that *the variance is constant* around the line.

```
# testing for homoscedasticity
bp.test.trans <- lmtest::bptest(formula = water.educ.new$folded.p.female
                                  water.educ.new$folded.p.basic.water)
bp.test.trans
```

```
##
##     studentized Breusch-Pagan test
##
## data:  water.educ.new$folded.p.female.school ~ water.educ.new$folded.p.basic
## BP = 6.3816, df = 1, p-value = 0.01153
```

- With a p-value of .01, the null hypothesis is rejected and the assumption fails.
- The data transformation worked to mostly address the problem of normality for the females in school variable, but the transformed data were not better for linearity or homoscedasticity.

# Interpreting the results

- Write a conclusion:
    - There was a statistically significant, positive, and strong ($r = .67$; $t = 8.82$; $p<.05$; 95% CI: .55 - .77) relationship between the transformed variables for percentage of females in school and percentage of citizens with basic water access in a sample of countries. As the percentage of citizens with basic water access increases, so does the percentage of school-age females in school. The data failed several of the assumptions for r and so these results should not be generalized outside the sample.
- Note: Although transformations may work to meet assumptions in some cases, transformations also make interpretation more complicated. Because the relationship is now between the transformed values, the interpretation is now with respect to the transformed values and not the original data. When possible, use the original untransformed data.