PSY 3307

Correlations

Jonathan A. Pedroza PhD

Cal Poly Pomona

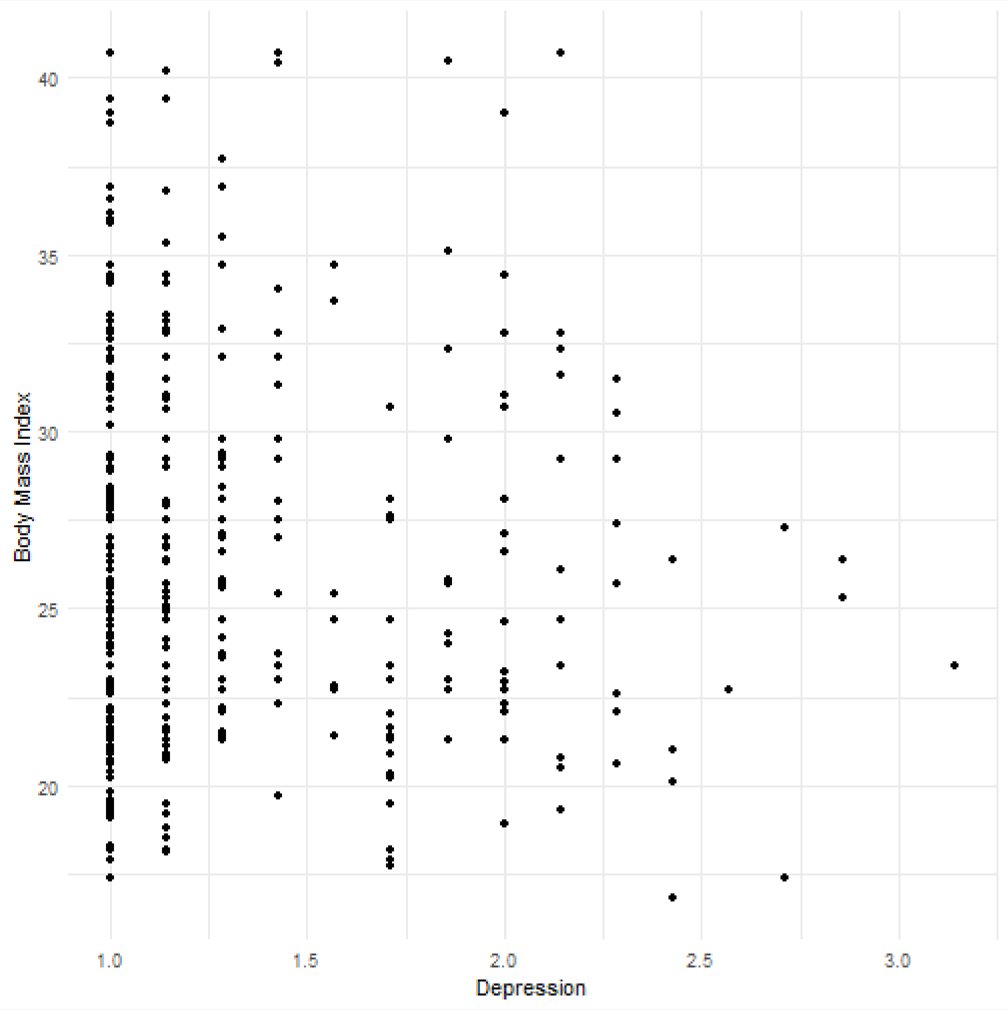
2021-11-16

Agenda

- Recap of Relationships
  - now with continuous variables
- What is a Correlation?
- Different Correlation Coefficients
- Partial and semi-partial correlations
- Reporting correlation coefficients

Relationships

- how related/associated two variables are
  - now between a continuous IV and a continuous DV
- three different relationships
  - positive relationship (as IV goes up, DV goes up)
  - negative relationship (as IV goes up, DV goes down)
  - also called inverse relationship
  - no relationship



40
35
30
25
20
Body Mass Index
1.0
1.5
2.0
2.5
3.0
Depression

```r
cor.test(jp$bmi, jp$depression)
```

```
## 
## 	Pearson's product-moment correlation
## 
## data:  jp$bmi and jp$depression
## t = -1.1019, df = 370, p-value = 0.2712
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.15794997  0.04474988
## sample estimates:
##         cor 
## -0.05718939
```

Modeling with ANOVA

$$X_{ij} = \mu + \gamma_j + \epsilon_{ij}$$

- mu is the grand mean
- gamma is the specific treatment effect for group j (which group you are interested in looking at)
- epsilon if the error/residual of a specific individual (how much an individual deviates from the group's mean)

Modeling Correlations

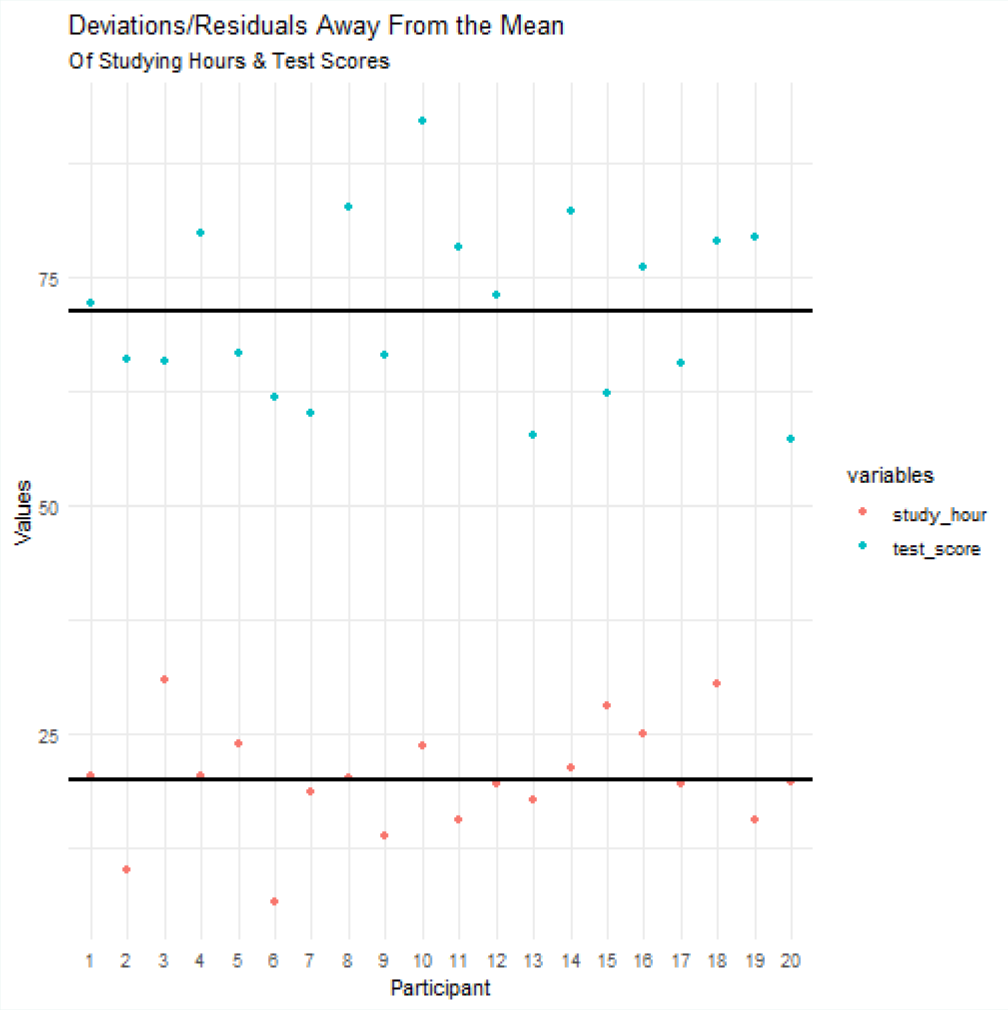
$$Y_i = (model) + error_i$$

$$Y_i = (bX_i) + error_i$$

Variance

- **variance** is the average of the squared deviations of the scores around the sample mean

$$s^2 = \frac{\Sigma(x - \overline{x}^2)}{N - 1}$$



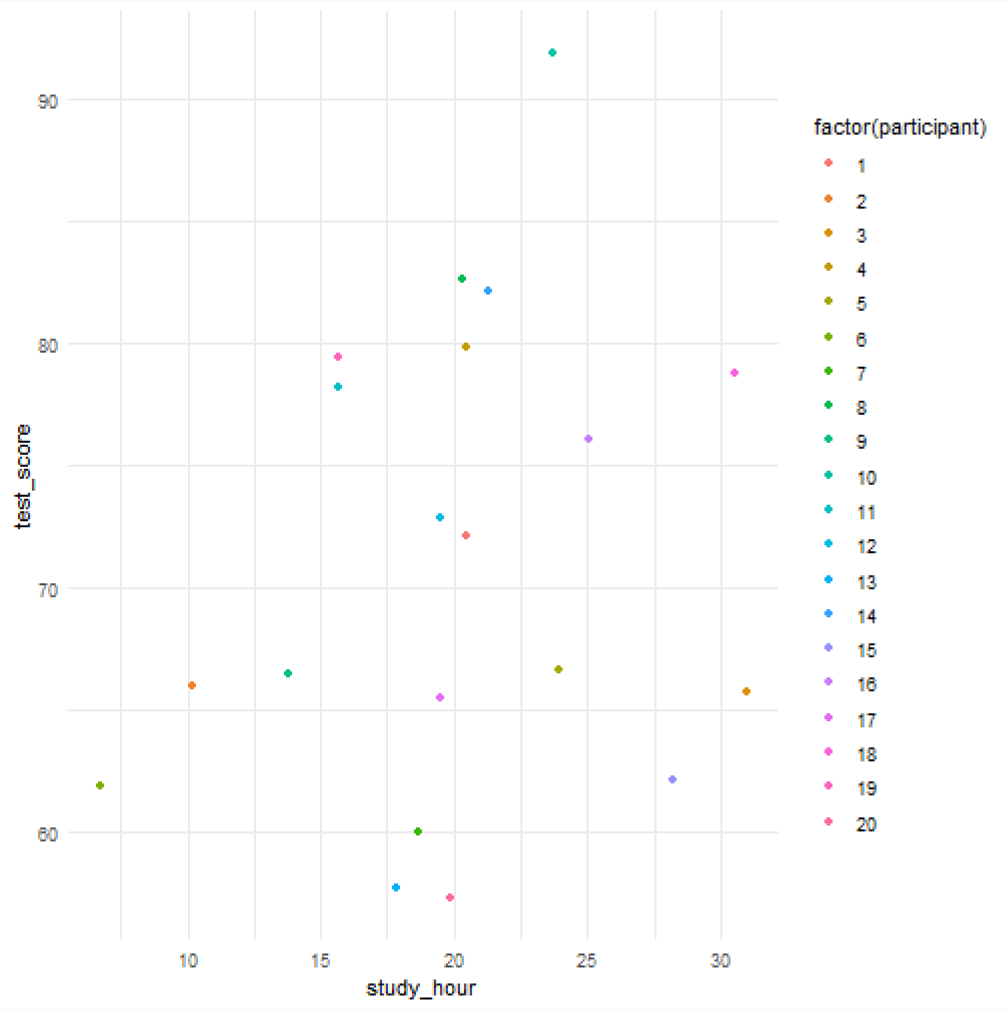
Deviations/Residuals Away From the Mean
Of Studying Hours & Test Scores
Values
75
50
25
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Participant
variables
study_hour
test_score

Variance

- if there is a relationship between two variables, as one variable deviates from the mean, the other variable would deviate from the mean
  - either in the same direction or opposing directions
- to eliminate values from zeroing itself out, we square our deviations
- however, when we multiply the deviations of one variable by the deviations of the second variable, we get a **cross-production deviation**
- when we average the combined deviations/cross-product deviations, we get covariance

Covariance

$$covariance(x, y) = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{N - 1}$$


test_score
90
80
70
60
10
15
20
25
30
study_hour
factor(participant)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Covariance Example

```
example <- data.frame(x = c(1, 4, 5, 6, 7),
                      y = c(10, 9, 8, 6, 8))
example
```

```
##   x  y
## 1 1 10
## 2 4  9
## 3 5  8
## 4 6  6
## 5 7  8
```

```
mean(example$x)
```

```
## [1] 4.6
```

```
mean(example$y)
```

```
## [1] 8.2
```

```r
example$x_deviations <- example$x - 4.6
example$y_deviations <- example$y - 8.2

example
```

```
##   x  y x_deviations y_deviations
## 1 1 10         -3.6          1.8
## 2 4  9         -0.6          0.8
## 3 5  8          0.4         -0.2
## 4 6  6          1.4         -2.2
## 5 7  8          2.4         -0.2
```

$$covariance(x, y) = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{N - 1}$$

$$\frac{(1 - 4.6)(10 - 8.2) + (4 - 4.6)(9 - 8.2) + (5 - 4.6)(8 - 8.2) + (6 - 4.6)(6 - 8.2)}{5 - 1}$$

```r
(1 - 4.6)
```

```
## [1] -3.6
```

```r
(10 - 8.2)
```

```
## [1] 1.8
```

```r
(4 - 4.6)
```

```
## [1] -0.6
```

```r
(9 - 8.2)
```

```
## [1] 0.8
```

```r
(5 - 4.6)
```

```
## [1] 0.4
```

```r
(8 - 8.2)
```

```
## [1] -0.2
```

```r
(6 - 4.6)
```

```
## [1] 1.4
```

```r
(6 - 8.2)
```

```
## [1] -2.2
```

```r
(7 - 4.6)
```

```
## [1] 2.4
```

```r
(8 - 8.2)
```

```
## [1] -0.2
```

```r
5 - 1
```

```
## [1] 4
```

$$\frac{(-3.6)(1.8) + (-.6)(.8) + (.4)(-.2) + (1.4)(-2.2) + (2.4)(-.2)}{4}$$

```r
(-3.6)*(1.8)
```

```
## [1] -6.48
```

```r
(-.6)*(.8)
```

```
## [1] -0.48
```

```r
(.4)*(-.2)
```

```
## [1] -0.08
```

```r
(1.4)*(-2.2)
```

```
## [1] -3.08
```

```r
(2.4)*(-.2)
```

```
## [1] -0.48
```

$$\frac{(-6.48) + (-.48) + (-.08) + (-3.08) + (-.48)}{4}$$

```
(-6.48) + (-.48) + (-.08) + (-3.08) + (-.48)
```

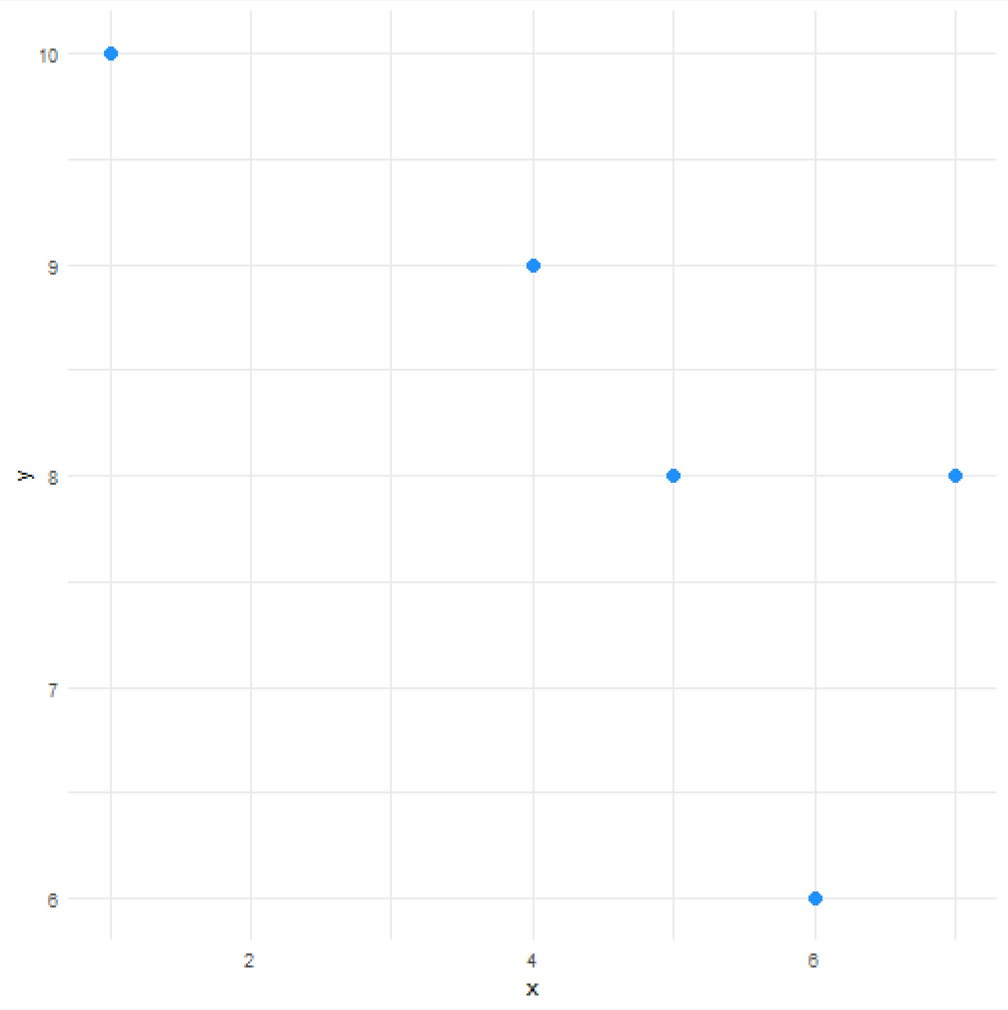
```
## [1] -10.6
```

$$covariance(x, y) = \frac{-10.6}{4}$$

```
-10.6/4
```

```
## [1] -2.65
```

$$covariance(x, y) = -\mathbf{2.65}$$


10
9
8
7
6
y
2
4
6
x

Covariance

- positive covariances means that when one variable deviates from the mean, the second variable deviates from the mean in the same direction as the first variable
    - negative covariances is when one variables deviates in one direction, the second variable deviates in the opposite direction
- covariance is not a standardized measure, the value can be as high or low as possible
    - **correlation coefficient** is the standardized equivalent of a covariance measure

Covariance

- **standardization** is converting the scale to a unit of measurement that can be equivalent between all relationships
  - this is in standard deviation units
- the most common correlation coefficient is `r` or the **Pearson product-moment correlation coefficient**, or simply **Pearson's correlation coefficient**

$$r = \frac{cov_{xy}}{s_x s_y}$$

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x s_y}$$

Correlation Coefficients

- to get the correlation coefficient, we calculate the covariance and divide by the standard deviations of both of our variables
- from our previous example, we had a covariance of -2.65 and need to get the standard deviations from our two variables
- then we can multiply the standard deviations and have -2.65 divided by the multiplied standard deviations of both variables to get a correlation coefficient

```r
sd(example$x)
```

```
## [1] 2.302173
```

```r
sd(example$y)
```

```
## [1] 1.48324
```

```r
2.30*1.48
```

```
## [1] 3.404
```

```r
-2.65/3.40
```

```
## [1] -0.7794118
```

Correlation Coefficient & Effect Sizes

- by standardizing the covariance, similar to our z-scores, we can only have correlations between -1 (perfect negative correlation) and +1 (perfect positive correlation)
- thankfully, we don't need to calculate anything new for our effect sizes
- correlation coefficients (**r**) are effect sizes
  - +- .1 small effect
  - +- .3 medium/moderate effect
  - +- .5 large effect

Different types of Correlation

- bivariate correlations
  - correlation between two variables
- partial correlations
  - quantifies the relationship between two variables while "controlling/adjusting" for the effect of one or more other variables

Significance of Correlation Coefficients

- similar to other test statistics we have tested (t-test, ANOVA, z-test), we can test to see if our correlation coefficient is statistically significant
  - we are testing to see if our correlation coefficient is different from zero
  - we are testing to see if our relationship is different from no relationship

Significance of Correlation Coefficients

- the problem with pearson's `r` is that the sampling distribution is not normally distributed
  - thanks to Fisher (1921), there is a calculation to make it normal

$$z_r = \frac{1}{2}log_e(\frac{1+r}{1-r})$$

Significance of Correlation Coefficients

- There is also the accompanying standard error

$$SE_{z_r} = \frac{1}{\sqrt{N-3}}$$

Significance of Correlation Coefficients

- you cal also get a normal z score

$$z = \frac{z_r}{SEz_r}$$

Significance of Correlation Coefficients

- now to come full circle, we don't typically use z-scores to get correlation significance values
  - we use t-tests

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Hypotheses

H0: There will be no relationship between x and y

H1: There will be a relationship between x and y

one-tail hypothesis

H1: There will be a positive relationship between x and y

H1: There will be a negative relationship between x and y

Confidence Intervals for r

- we use the z~r~ value and the corresponding SE to then calculation confidence interavls like we did previously

$$lower\ CI = \overline{X} - (1.96 * SE)$$

$$upper\ CI = \overline{X} + (1.96 * SE)$$

becomes

$$lower\ CI = z_r - (1.96 * SE)$$

$$upper\ CI = z_r + (1.96 * SE)$$

these can then be converted back to a correlation coefficient by

$$r = \frac{\epsilon^{2z_r} - 1}{\epsilon^{2z_r} + 1}$$

Better Option for Confidence Intervals

- we can bootstrap the correlation test to get bootstrapped confidence intervals that are useful for non-normal distributed data

Interpretation

- remember that when using correlational designs, we cannot infer causality from our findings
- our bivariate correlations cannot be used to infer causality
    - **third variable problem** (**tertirum quid**) there may be different variables not tested that could be influencing the relationship we are looking at
    - **direction of causality** we are not sure if x influences y or if y influences x
    - ex: depression and BMI/obesity measures

Assumptions of Bivariate Correlation

- outliers
- IV and DV need to be continuous
- the data should be able to be linear

Using R^2^ for Interpretation

- our correlation coefficient squared is a measure of the amount of variability in one variable that is shared by the other
- R^2^ is a measure of the variability is shared between your IV and your DV
- important way of stating this though, unlike the eta-squared, which can be used for experiments to show causality, R^2^ is different
  - "X shares _% of the variation in Y"

Spearman's Correlation Coefficient

- **Spearman's correlation coefficient** or `r~s~` is a non-parametric test that uses ranked data (ordinal data)
  - by using ranked data, we can remove the influence of extreme scores (outliers)

$$rho = \rho$$

- the test works by ranking data (recoding continuous data into categorical data) and then applying Pearson's equation to ranked data

Kendall's tau

- non-parametric test used when you have a small sample size

$$tau = \tau$$

- **Kendall's tau** is used over Spearman's coefficient when you have a small dataset/sample size with a large number of tied ranks
  - if you have high frequencies in many categories then you would use Kendall's tau

Biserial & Point-Biserial Correlations

- Biserial and point-biserial correlation coefficients are similar in that they are correlations where one variable is dichotomous (2 categories)
    - the difference is that dichotomous variable is either discrete or continuous
- a **point-biserial correlation coefficient** is used when one variable is a discrete dichotomous variable (sex)
- a **biserial correlation coefficient** is used when one variable is a continuous dichotomous variable (passing an exam = 1, failing an exam = 0)

$$point - biserial = r_{pb}$$

$$biserial = r_b$$

Partial Correlation

- remember that when we look at the variance "explained" by one variable on the second variable (DV), we are talking about R^2^
- however, sometimes we want to look at the influence of several variables on your DV
  - from this, we may want to see how much unique influence each variable has on your DV
- a **partial correlation** is when we are looking at the unique relationship between a IV and a DV while other included variables are held constant
  - this is somewhat like multiple regression (which we'll get to in the next slide)
- holding constant is another way of controlling for or adjusting for
- **zero-order correlation** is a pearson correlation coefficient without controlling for any other variable

Semi-partial Correlations

- also referred to as **part correlation**
- partial correlation is the unique relationship between two variables when controlling for a third variable
  - that means we are controlling for the effect of the third variable on both variables
- **semi-partial** correlation only controls for the effect that the third variable has on one of the variables in the correlation

Comparing Independent & Dependent rs

- independent rs
- you can compare correlation coefficients for different groups to see if the correlation coefficients are significantly different from one another
  - correlation between depression and BMI between males and females
- transform them into z values and then compare the converted scores using a z-test to see if the differences are significantly different from one another
- dependent rs
- to compare dependent conditions/levels, you would use a t-test to see differences between two dependent correlations
  - if 3 conditions, you would test every correlation and compare each correlation to another

Calculating Effect Sizes

- correlation coefficients are effect sizes
- r = effect size because it is standardized (0 to +-1)
- to get the proportion of variance you would square the correlation coefficient

$$R^2 = r^2$$

- R^2^ can be used for other correlation coefficients other than Pearson's (Spearman's)
    - for Spearman's the calculation is the same, however the interpretation is the proportion of variance in the ranks between the two variables
- Kendall's Tau is not comparable to the other two coefficients
    - tau can be used as an effect size but it is not comparable to Pearson's or Spearman's correlation coefficients and should not be squared

Reporting Correlation Coefficients

- reporting correlation coefficients includes the two variables that you conducted a correlation of
    - there was a significant association/relationship between X and Y
    - there was no evidence of a statistically significant relationship/assocaition between X and Y
- It is best practice to not state that **there was no significant association**
    - this is supporting your null hypothesis and by the rules of probability, we are not sure whether or not we found a true relationship
    - we can only say that in our sample, there was either evidence of a statistically significant relationship or no evidence of a significant relationship

Reporting Correlation Coefficients

- There was a statistically significant relationship between depression levels and body mass index; $r = .23$, $p = .015$.
- There was no evidence of a significant relationship between depression levels and test scores ($r = .03$, $p = .425$).