Correlations

Relationships

- What is the relationship between IV and DV?
- Measuring relationships depend on type of measurement
- You have primarily been working with categorical IVs (*t*-test, ANOVA)
- We are moving towards continuous predictor models

Review of Dispersion

Variation (sum of squares)

$$SS = \sum (x - \bar{x})^2$$

$$SS = \sum (x - \mu)^2$$

Review of Dispersion

Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{N - 1}$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Review of Dispersion

Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Population Variability

Sums of squares

$$SS = \Sigma(X_i - \mu_x)^2$$

Variance

$$\sigma^2 = \frac{\Sigma(X_i - \mu_x)^2}{N} = \frac{SS}{N}$$

Standard devation

$$\sigma = \sqrt{\frac{\Sigma(X_i - \mu_x)^2}{N}} = \sqrt{\frac{SS}{N}} = \sqrt{\sigma^2}$$

Sample variability

Sums of squares

$$SS = \Sigma(X_i - \bar{X})^2$$

Variance

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N-1} = \frac{SS}{N-1}$$

Standard deviation

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N-1}} = \sqrt{\frac{SS}{N-1}} = \sqrt{s^2}$$

Review of Dispersion

Formula for standard error of the mean?

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

$$\sigma_M = \frac{\hat{s}}{\sqrt{N}}$$

Associations

- i.e., relationships
- to look at continuous variable associations we need to think in terms of how variables relate to one another

Bi-variate descriptives

Covariation

"Sum of the cross-products"

Population

$$SP_{XY} = \Sigma(X_i - \mu_X)(Y_i - \mu_Y)$$

Sample

$$SP_{XY} = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})$$

Covariance

Sort of like the variance of two variables

Population

$$\sigma_{XY} = \frac{\Sigma(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Sample

$$s_{XY} = cov_{XY} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

Covariance table

$$\mathbf{K_{XX}} = \begin{bmatrix} \sigma_X^2 & cov_{XY} & cov_{XZ} \\ cov_{YX} & \sigma_Y^2 & cov_{YZ} \\ cov_{ZX} & cov_{ZY} & \sigma_Z^2 \end{bmatrix}$$

$cov_{xy} = cov_{yx}$

Covariance table

$$\mathbf{K_{XX}} = \begin{bmatrix} \sigma_X^2 & 126.5 & 5.2 \\ 126.5 & \sigma_Y^2 & cov_{YZ} \\ 5.2 & cov_{ZY} & \sigma_Z^2 \end{bmatrix}$$

Which variable, Y or Z, does X have greater relationship with?

Can't know because you don't know what units they're measured in!

Correlations

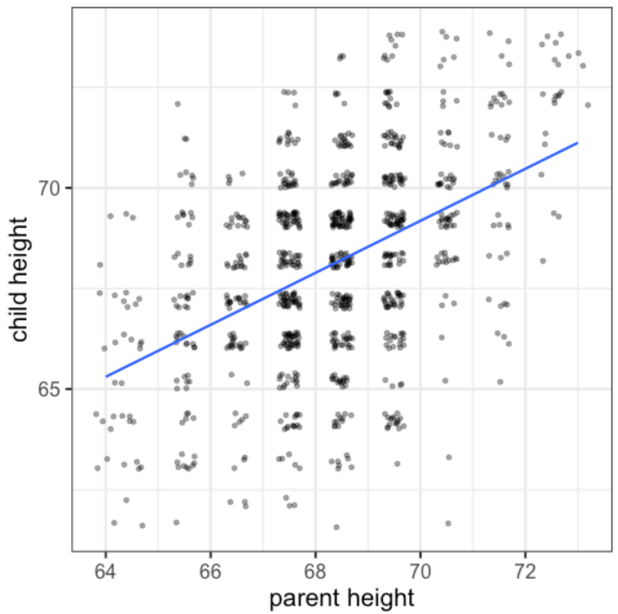
- How much two variables are linearly related
- -1 to 1
- Invariant to changes in mean or scaling
- Most common (and basic) effect size measure
- Will use to build our regression model

Conceptually

Ways to think about a correlation:

- How two vectors of numbers co-relate
- Product of z-scores
    - Mathematically, it is
- The average squared distance between two vectors in the same space

Correlations



- This will become our regression line. Right now, it is our correlation line. At this point **they are the same!**
- You can do a lot of things with just correlations.

Correlation

Pearson product moment correlation

Population

$$\rho_{XY} = \frac{\Sigma z_X z_Y}{N} = \frac{SP}{\sqrt{SS_X}\sqrt{SS_Y}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Sample

$$r_{XY} = \frac{\Sigma z_X z_Y}{n-1} = \frac{SP}{\sqrt{SS_X}\sqrt{SS_Y}} = \frac{s_{XY}}{s_X s_Y}$$

Effect size

- The strength of relationship between two variables
- η^2, Cohen's d, R^2 , Risk-ratio, etc
- Sig is a function of effect size and sample size
- Statistical sig $\neq$ practical sig
- *z*-scores allow us to compare across units of measure
- The correlation coefficient is a **standardized effect size**.
  - Ex: $r_{age\ and\ height\ in\ kids} = .70$
    $r_{self-\ and\ other-ratings\ extraversion} = .25$

What is a large correlation?

- Cohen (1988): .1 (small), .3 (medium), .5 (large)
  - Often forgot: Cohen said only to use them when you had nothing else to go on, and has since regretted even suggesting benchmarks to begin with.
- r^2: Proportion of variance "explained"
  - as Ozer & Funder (2019) discuss, we're not really explaining anything and the change in scale can mess up our interpretations if we're not careful.

What is the size of the correlation?

- Chemotherapy and breast cancer survival?
- Batting ability and hit success on a single at bat?
- Antihistamine use and reduced sneezing/runny nose?
- Combat exposure and PTSD?
- Ibuprofen on pain reduction?
- Gender and weight?
- Therapy and well being?
- Observer ratings of attractiveness?
- Gender and arm strength?

What is the size of the correlation?

- Chemotherapy and breast cancer survival? (.03)
- Batting ability and hit success on a single at bat? (.06)
- Antihistamine use and reduced sneezing/runny nose? (.11)
- Combat exposure and PTSD? (.11)
- Ibuprofen on pain reduction? (.14)
- Gender and weight? (.26)
- Therapy and well being? (.32)
- Observer ratings of attractiveness? (.39)
- Gender and arm strength? (.55)

Questions to ask yourself:

- What is your N?
- What is the typical effect size in the field?
- Study design?
- What is your DV?
- Importance?

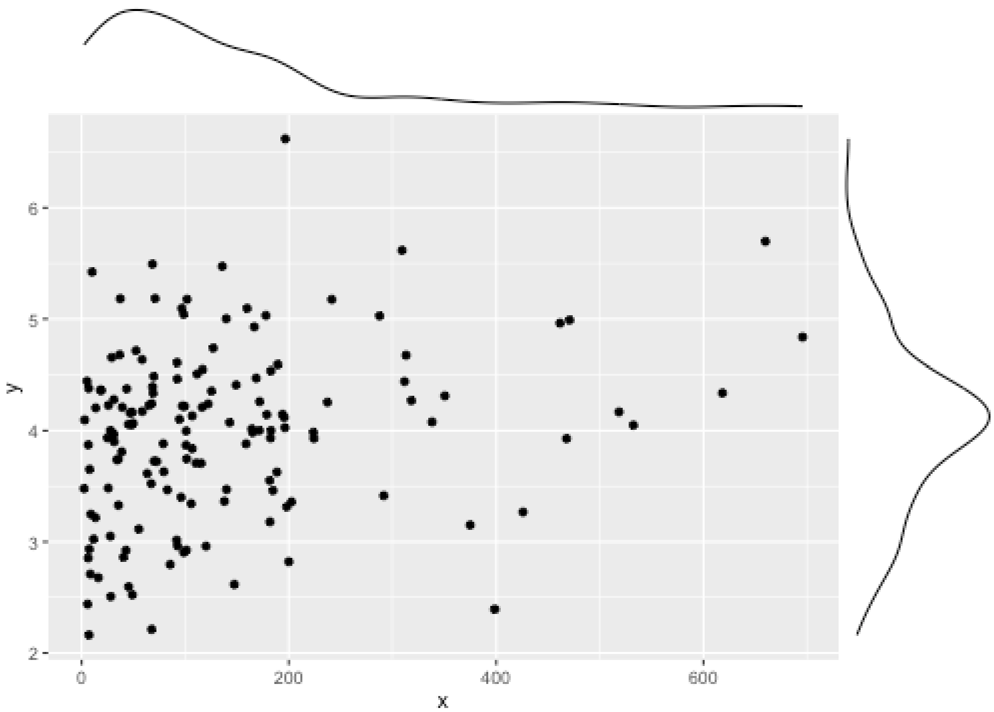
What affects correlations?

Correlations can be easily fooled by qualities of your data, like:

- Skewed distributions
- Outliers
- Restriction of range
- Nonlinearity
- Reliability
- Measurement Overlap (redundancy)

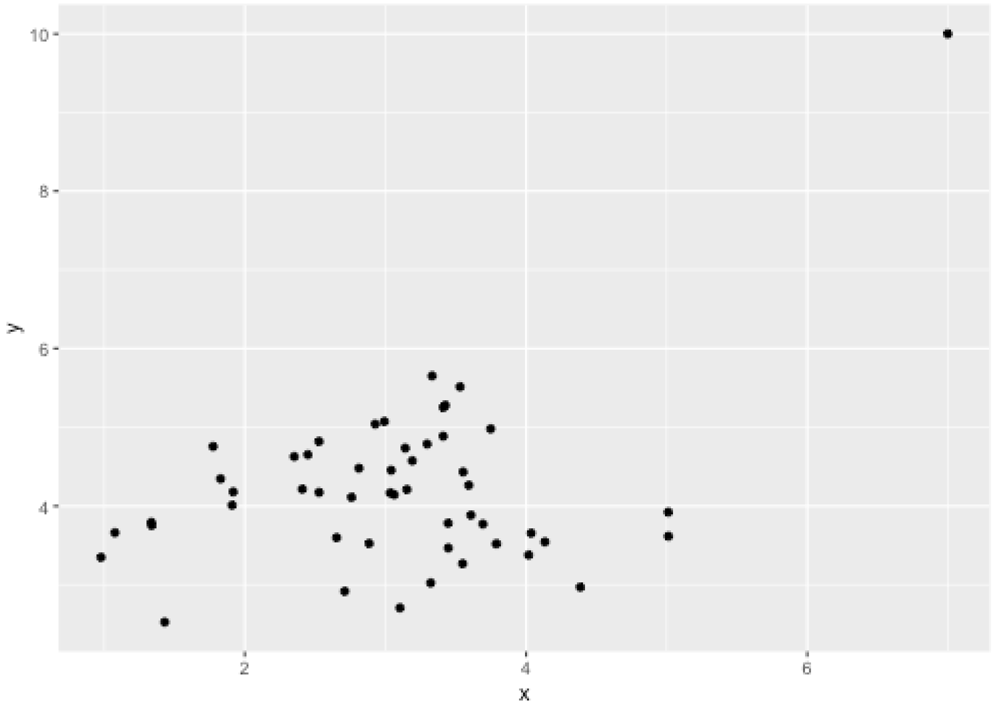
Skewed distributions

```
p = data %>% ggplot(aes(x=x, y=y)) + geom_point()
ggMarginal(p, type = "density")
```



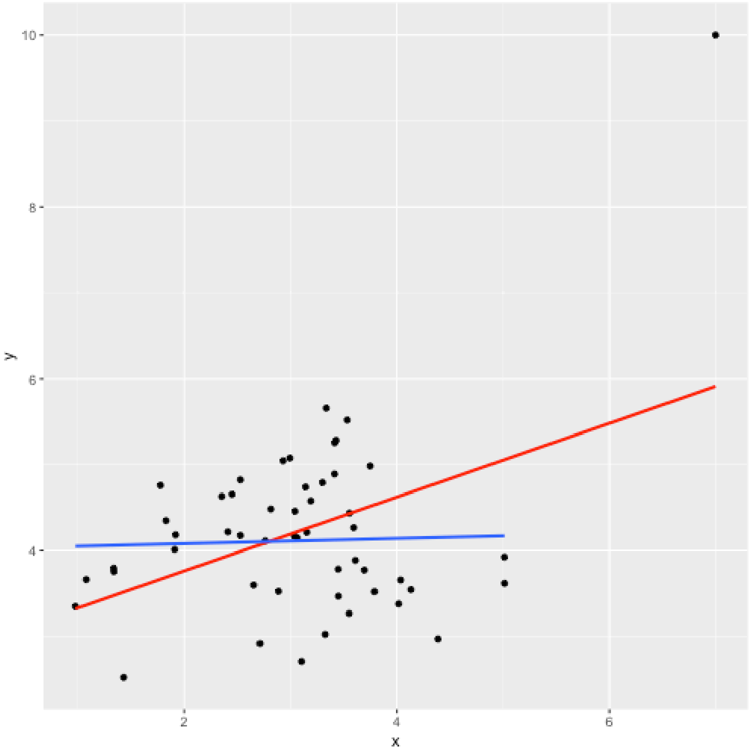
Outliers

```
data %>% ggplot(aes(x=x, y=y)) + geom_point()
```



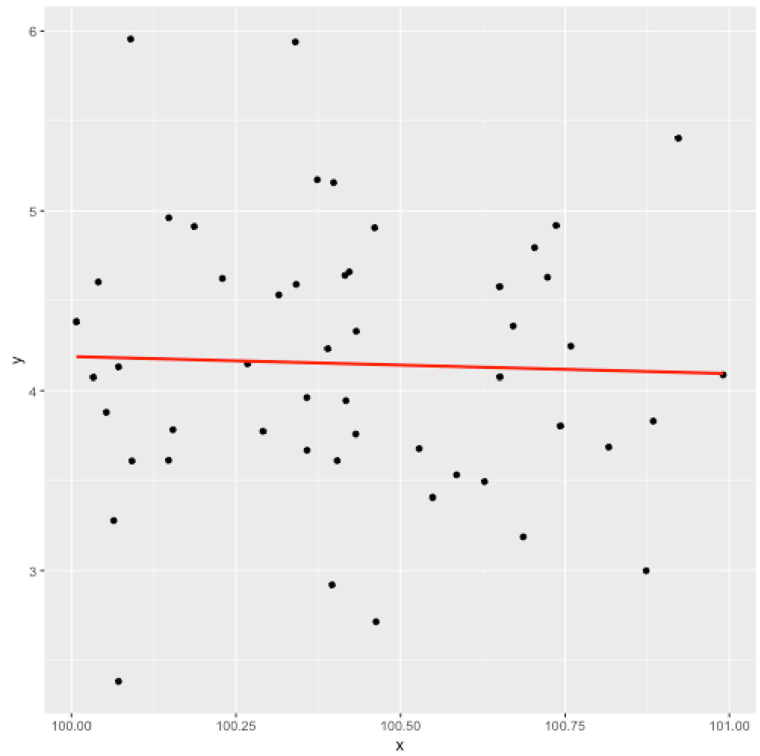
Outliers

```
data %>% ggplot(aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE,
              color = "red") +
  geom_smooth(data = data[-51,],
              method = "lm",
              se = FALSE)
```



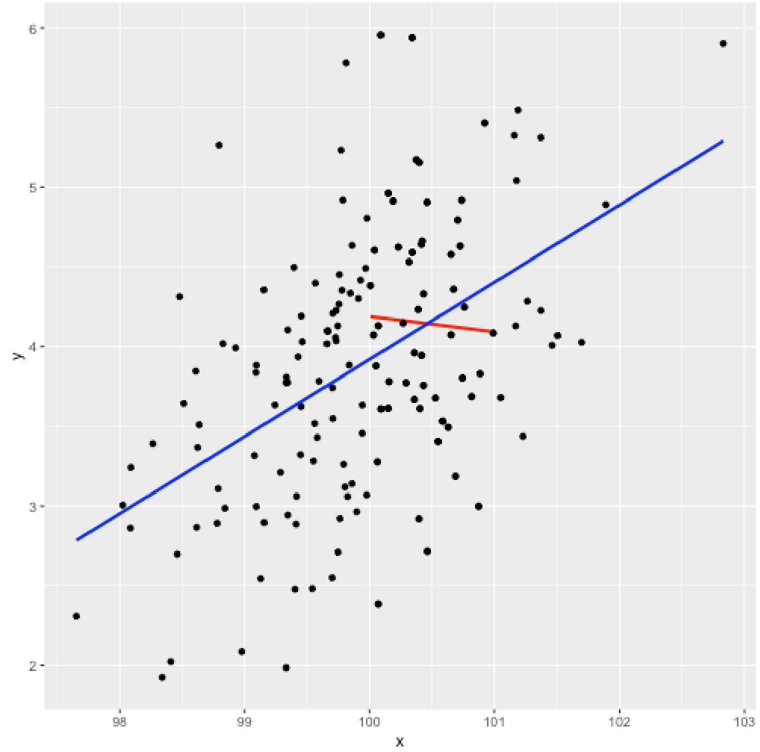
Restriction of range

```
data %>%
ggplot(aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE,
              color = "red")
```

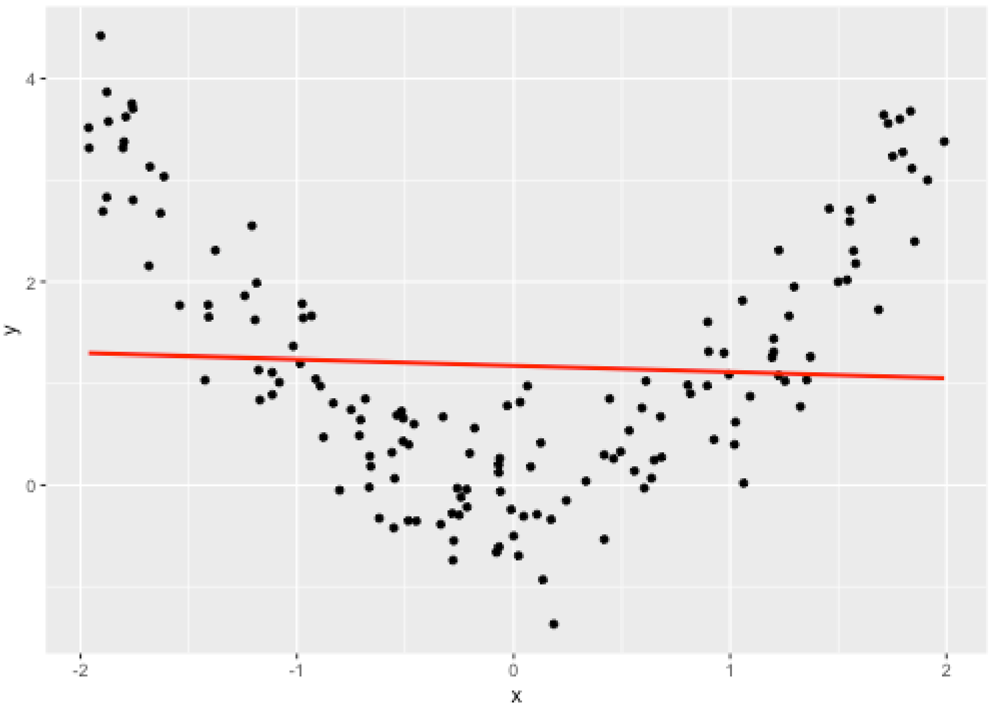


Restriction of range

```
data %>%
ggplot(aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE,
              color = "red") +
  geom_point(data = real_data) +
  geom_smooth(method = "lm",
              se = FALSE,
              data = real_data,
              color = "blue")
```



Nonlinearity

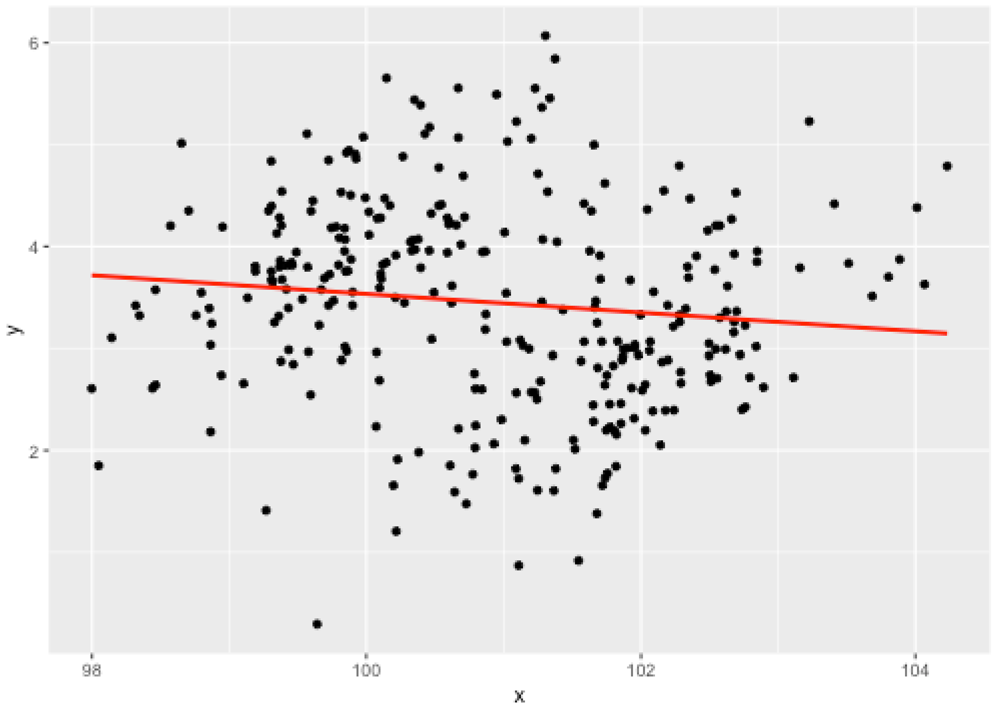


It's not always apparent

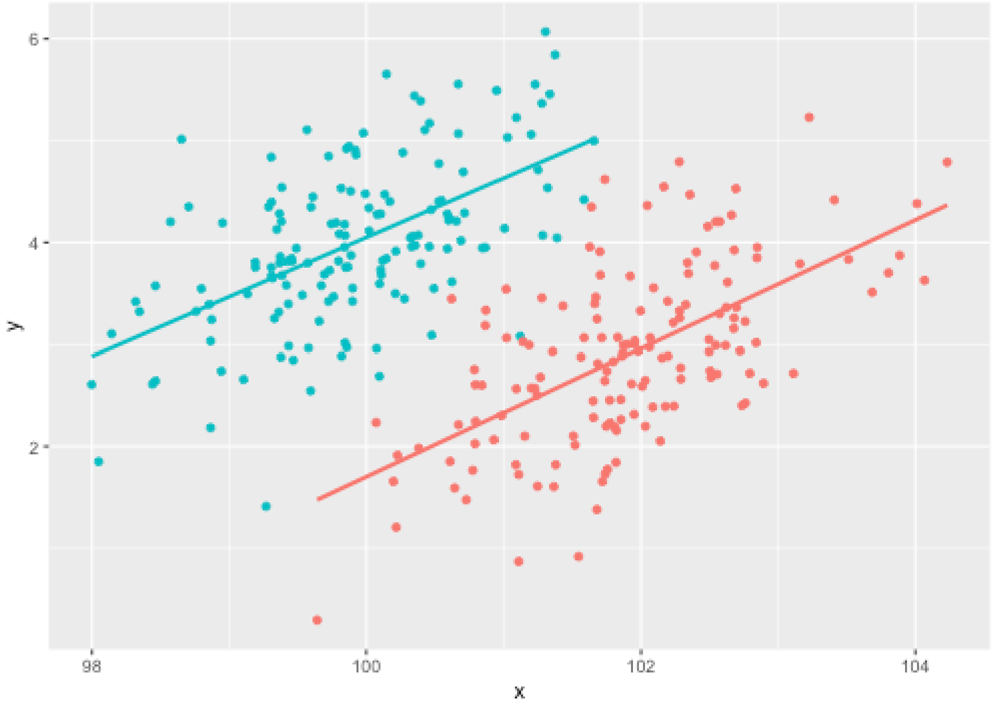
Sometimes issues that affect correlations won't appear in your graph, but you still need to know how to look for them.

- Multiple Groups
- Low reliability
- Content overlap

Multiple groups



Multiple groups



Known as **Simpson's Paradox**

Reliability

Which would you rather have?

- 1-item final exam versus 30-item?
- assessment via trained clinician vs tarot cards?
- fMRI during minor earthquake vs no earthquake?

All measurement includes error

- Score = true score + measurement error (CTT version)

Reliability

- Error is random; it cannot correlate with something
- Because we don't measure our variables perfectly, we get lower correlations compared to true correlations
- If we want to have a valid measure it better be a reliable measure

Reliability

- Think of reliability as a correlation with a measure and itself in a different world, at a different time, or a different but equal version

$$r_{XX}$$

Reliability

- True score variance divided by observed variance

$$r_{XX} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

- How do you assess true score variance?

$$r_{XY} = r_{X_T Y_T} \sqrt{r_{XX} r_{YY}}$$

- We can look at our observed correlation (r_{XY}) and our reliability coefficients (r_{XX} and r_{YY}), and solve

Reliability

$$r_{X_T Y_T} = \frac{r_{XY}}{\sqrt{r_{XX} r_{YY}}}$$

$$r_{X_T Y_T} = \frac{.30}{\sqrt{(.70)(.70)}} = .42$$

- CAVEAT: Reliabilities are also estimates. They can be wrong and there are lots of ways to get a reliability coefficient. The correlation you calculate ($r_{X_T Y_T}$) is the highest correlation it could possibly be -- not the actual "theoretically true" correlation.

Most common ways to assess

- Cronbach's α

```r
library (psych)
alpha(measure)
## Gives average split half correlation
## Can tell you if you are assessing a single construct
```

- Test - retest reliability (r, ICC)
- Inter-rater reliability (kappa, ICC)

Reliability

- If you are going to measure something, do it well
- Applies to ALL IVs and DVs, and all designs
- Remember this when interpreting other research

Statistical test

Hypothesis testing

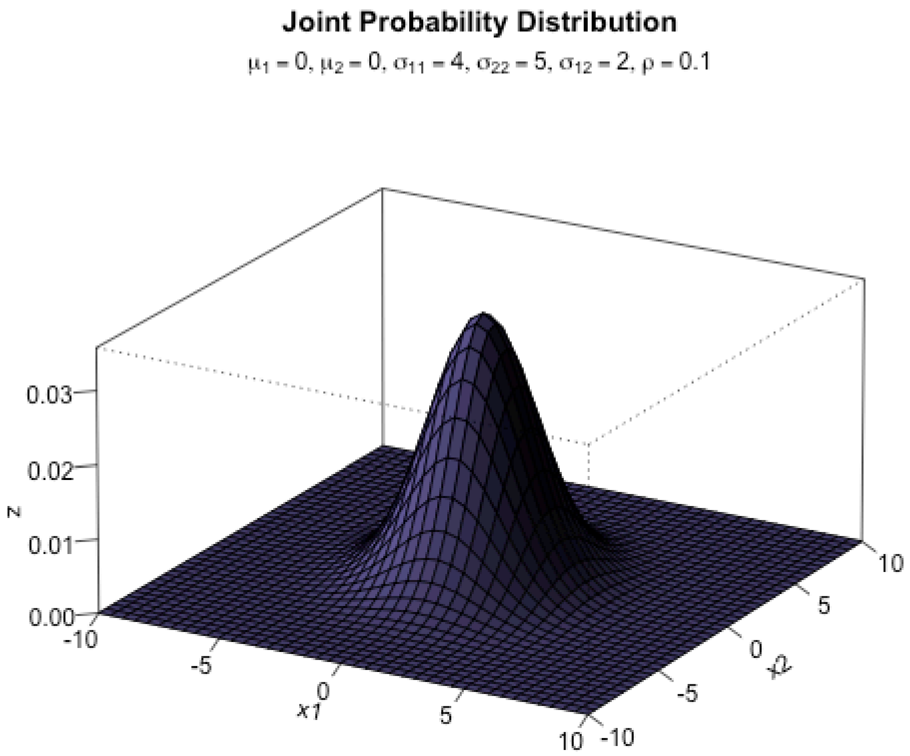
$$H_0 : \rho_{xy} = 0$$

$$H_A : \rho_{xy} \neq 0$$

Assumes:

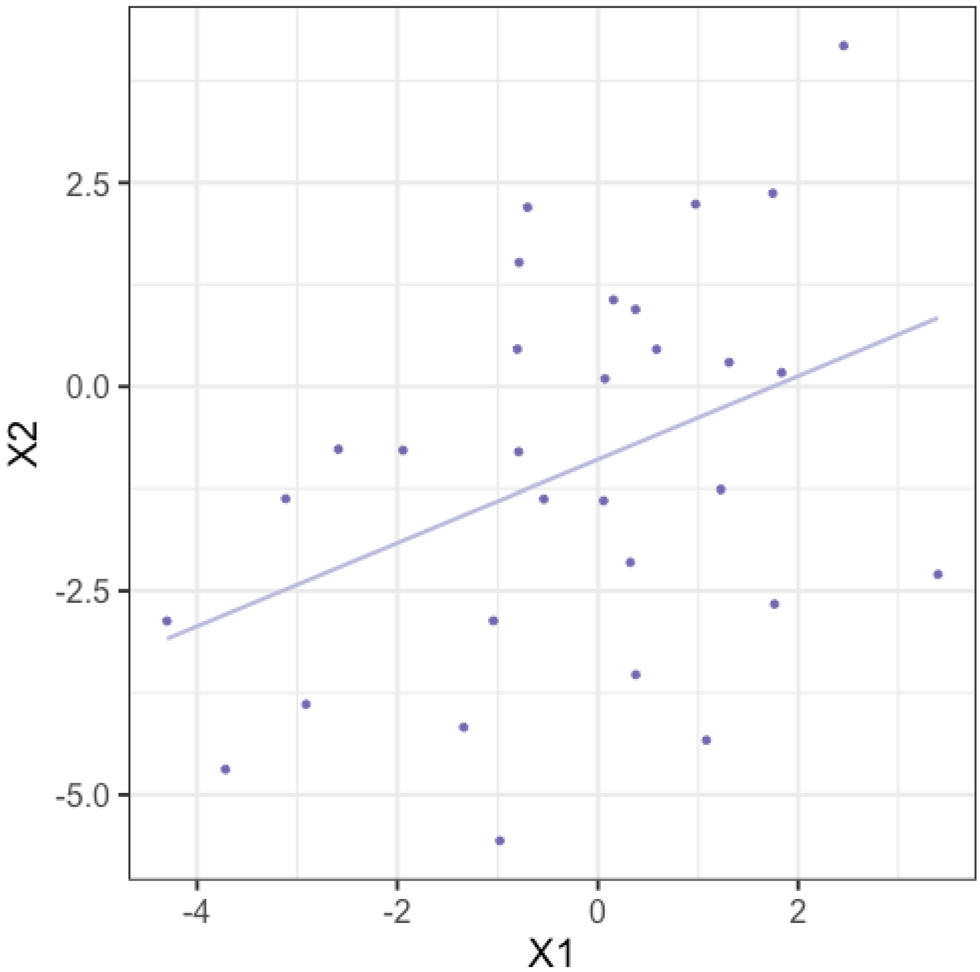
- Observations are independent
- Symmetric bivariate distribution (joint probability distribution)

Population



$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_{11}} - 2\rho\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}}\right]\right\}$$

Sample



Sampling distribution?

The sampling distribution we use depends on our null hypothesis.

If our null hypothesis is that $(\rho = 0)$, then we can use a ***t*-distribution** to estimate the statistical significance of a correlation.

Test Statistic

Signal divided by noise

$$t = \frac{r}{SE_r}$$

$$SE_r = \sqrt{\frac{1-r^2}{N-2}} \qquad DF = N - 2$$

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

Power calculations

What sample size do you need in order to have enough power to detect a **.1** correlation?

```r
library(pwr)
pwr.r.test(n = , r = .1, sig.level = .05 , power = .8)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 781.7516
##               r = 0.1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Power calculations

What sample size do you need in order to have enough power to detect a **.3** correlation?

```
library(pwr)
pwr.r.test(n = , r = .3, sig.level = .05 , power = .8)
```

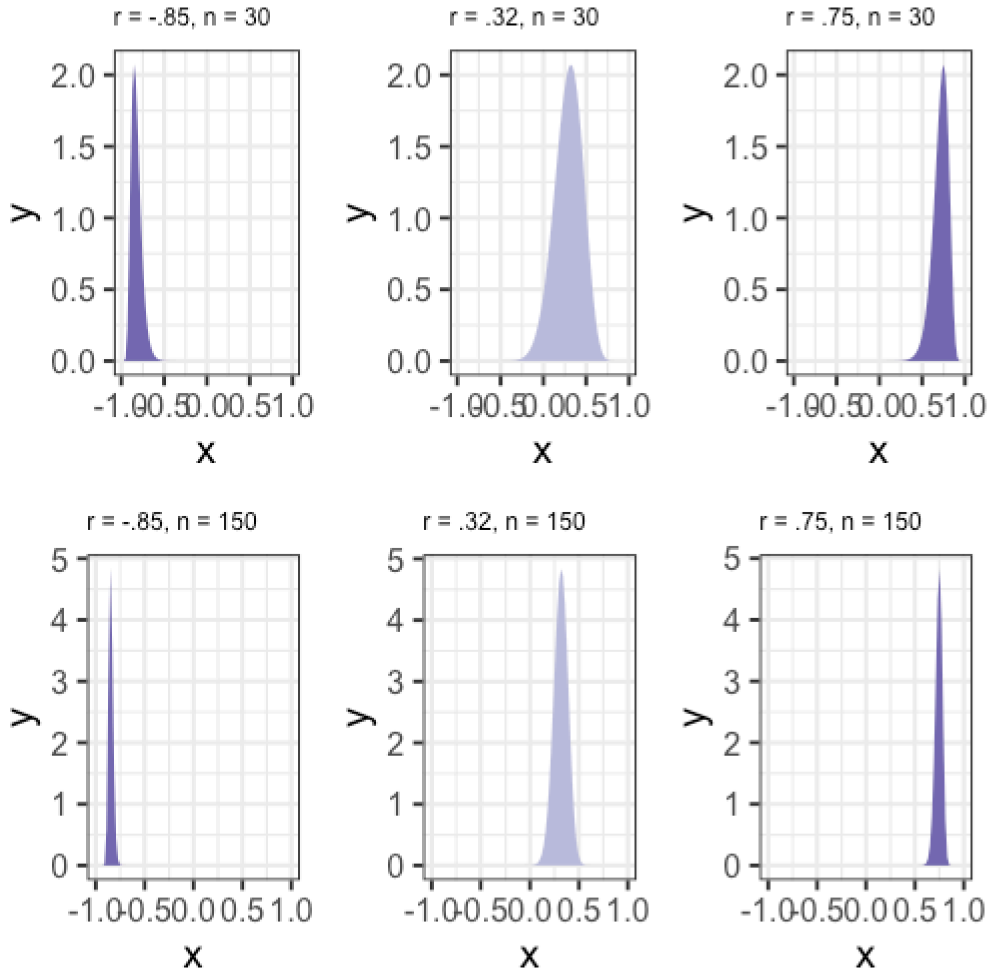
```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 84.07364
##               r = 0.3
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Power calculations

- But what is your confidence?
- N = 84 gives you CI[.09, .48]
- Schönbrodt & Perugini (2013) suggest correlations 'stabilize' at 250+ regardless of effect size

Fisher's r to z' transformation

If we want to make calculatio based on $\rho \neq 0$ then we will run into a skewed sampling distributio.

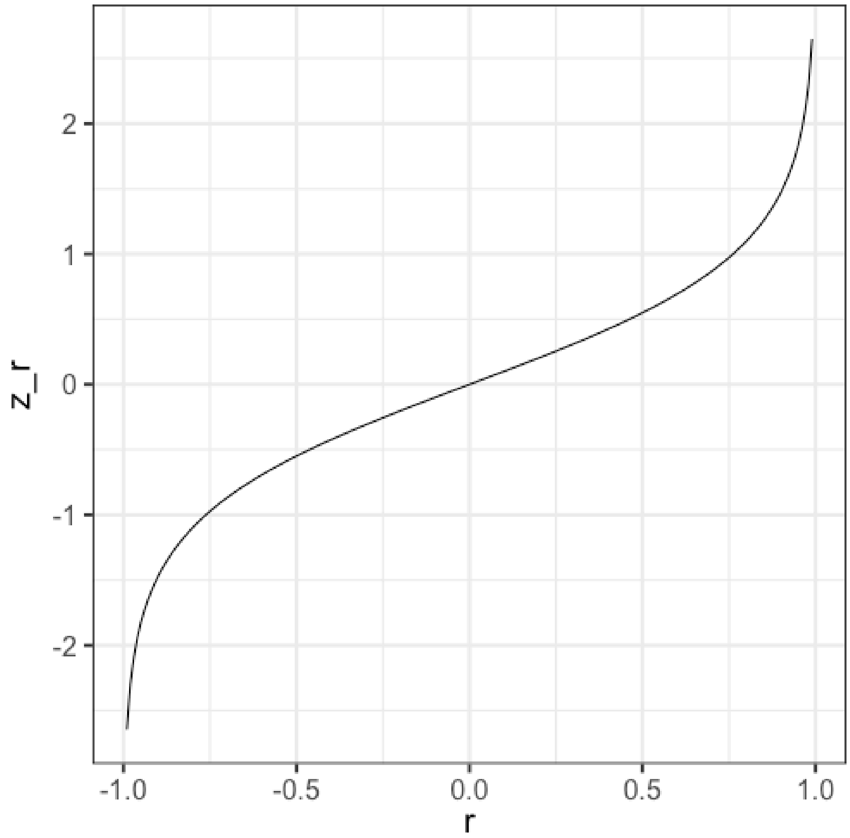


Fisher's r to z' transformation

- Skewed sampling distribution will rear its head when:
    - $H_0 : \rho \neq 0$
    - Calculating confidence intervals
    - Testing two correlations against one another
- r to z':

$$z^{'} = \frac{1}{2} ln \frac{1+r}{1-r}$$

Fisher's r to z' transformation



No longer bounded by 1 & -1

Computing confidence interval

1. Transform r into z'
2. Compute CI as you normally would using z'
3. Revert back to r

$$SE_z = \frac{1}{\sqrt{N-3}}$$

$$r = \frac{e^{2z'} - 1}{e^{2z'} + 1}$$

Note, `e` here stands for Euler's number. `exp(1)` is straight Euler's number or `e`, `exp(2)` is Euler's number squared or `e^2`

In a sample of 42 students, you calculate a correlation of 0.44 between hours spent outside on Saturday and self-rated health. What is the precision of your estimate?

$$z^{'} = \frac{1}{2} ln \frac{1 + .44}{1 - .44} = 0.47$$

$$SE_z = \frac{1}{\sqrt{42 - 3}} = 0.16$$

$$CI_{Z_{LB}} = 0.47 - (2.021)0.16 = 0.15$$

$$CI_{Z_{UB}} = 0.47 + (2.021)0.16 = 0.8$$

Why is it 2.021 instead of 1.96? Because we're using a *t* distribution with N-2 degrees of

$$CI_{r_{LB}} = \frac{e^{2(0.15)} - 1}{e^{2(0.15)} + 1} = 0.15$$

$$CI_{r_{UB}} == \frac{e^{2(0.8)} - 1}{e^{2(0.8)} + 1} = 0.66$$

How to do in R

```r
library(psych)
fisherz(r)
fisherz2r(z)
```

Two independent group test

- Does the correlation in group 1 differ from the correlation in group 2?

$$H_0 : \rho_1 = \rho_2$$

$$H_A : \rho_1 \neq \rho_2$$

- Normally distributed

$$Z = \frac{z_1^{'} - z_2^{'}}{se_{z_1 - z_2}}$$

Comparing two correlations

Again, we use Fisher's r to z' transformation. Here, we're transforming the correlations into z's, then using the difference between z's to calculate the test statistic.

$$Z = \frac{z_1^{'} - z_2^{'}}{se_{z_1 - z_2}}$$

$$se_{z_1 - z_2} = \sqrt{se_{z_1} + se_{z_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

- But probably best to do this test in another framework (e.g., GLM via interaction or SEM)

Example

Replication of Hill et al. (2012) where they found that the correlation between narcissism and happiness was greater for young adults compared to older adults

Young adults

$$N = 327$$

$$r = .402$$

Older adults

$$N = 273$$

$$r = .283$$

$$H_0 : \rho_1 = \rho_2$$

$$H_1 : \rho_1 \neq \rho_2$$

$$z_1^{'} = \frac{1}{2} ln \frac{1 + .402}{1 - .402} = 0.426$$

$$z_2^{'} = \frac{1}{2} ln \frac{1 + .283}{1 - .283} = 0.291$$

$$se_{z_1 - z_2} = \sqrt{\frac{1}{327 - 3} + \frac{1}{273 - 3}} = 0.082$$

$$\text{Test statistic} = \frac{z_1^{'} - z_2^{'}}{se_{z_1 - z_2}} = \frac{0.426 - 0.291}{0.082} = 1.639$$

```
pnorm(abs(zstat), lower.tail = F)*2
```

```
## [1] 0.1011256
```

Note: more examples at end of slides, after "next time"

Summary of NHST with Correlations

- If "is *r* different from the number 0, *t*-test" -- **where have you seen this before?**
- Use r to z if:
  - Need a CI
  - "is *r* different from another number that is not 0"
  - Compare 2 correlations against each other

Correlation matrices

Correlations are both a descriptive and an inferential statistic. As a descriptive statistic, they're useful for understanding what's going on in a larger dataset.

Like we use the `summary()` or `describe()` (psych) functions to examine our dataset *before we run any infernetial tests*, we should also look at the correlation matrix.

```r
library(psych)
data(bfi)
head(bfi)
```

```
##       A1 A2 A3 A4 A5 C1 C2 C3 C4 C5 E1 E2 E3 E4 E5 N1 N2 N3 N4 N5 O1 O2 O3 O4
## 61617  2  4  3  4  4  2  3  3  4  4  3  3  3  4  4  3  4  2  2  3  3  6  3  4
## 61618  2  4  5  2  5  5  4  4  3  4  1  1  6  4  3  3  3  3  5  5  4  2  4  3
## 61620  5  4  5  4  4  4  5  4  2  5  2  4  4  4  5  4  5  4  2  3  4  2  5  5
## 61621  4  4  6  5  5  4  4  3  5  5  5  3  4  4  4  2  5  2  4  1  3  3  4  3
## 61622  2  3  3  4  5  4  4  5  3  2  2  2  5  4  5  2  3  4  4  3  3  3  4  3
## 61623  6  6  5  6  5  6  6  6  1  3  2  1  6  5  6  3  5  2  2  3  4  3  5  6
##       O5 gender education age
## 61617  3      1        NA  16
## 61618  3      2        NA  18
## 61620  2      2        NA  17
## 61621  5      2        NA  17
## 61622  3      1        NA  17
## 61623  1      2         3  21
```

```r
cor(bfi)
```

```
##           A1 A2 A3 A4 A5 C1 C2 C3 C4 C5 E1 E2 E3 E4 E5 N1 N2 N3 N4 N5 O1
## A1         1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## A2        NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## A3        NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## A4        NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## A5        NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## C1        NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## C2        NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## C3        NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA NA
## C4        NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA NA
## C5        NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA NA
## E1        NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA NA
## E2        NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA NA
## E3        NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA NA
## E4        NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA NA
## E5        NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA NA
## N1        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA NA
## N2        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA NA
## N3        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA NA
## N4        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA NA
## N5        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1 NA
## O1        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  1
## O2        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## O3        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## O4        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## O5        NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## gender    NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## education NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```r
round(cor(bfi, use = "pairwise"),2)
```

```
##              A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1
## A1         1.00 -0.34 -0.27 -0.15 -0.18  0.03  0.02 -0.02  0.13  0.05  0.11
## A2        -0.34  1.00  0.49  0.34  0.39  0.09  0.14  0.19 -0.15 -0.12 -0.21
## A3        -0.27  0.49  1.00  0.36  0.50  0.10  0.14  0.13 -0.12 -0.16 -0.21
## A4        -0.15  0.34  0.36  1.00  0.31  0.09  0.23  0.13 -0.15 -0.24 -0.11
## A5        -0.18  0.39  0.50  0.31  1.00  0.12  0.11  0.13 -0.13 -0.17 -0.25
## C1         0.03  0.09  0.10  0.09  0.12  1.00  0.43  0.31 -0.34 -0.25 -0.02
## C2         0.02  0.14  0.14  0.23  0.11  0.43  1.00  0.36 -0.38 -0.30  0.02
## C3        -0.02  0.19  0.13  0.13  0.13  0.31  0.36  1.00 -0.34 -0.34  0.00
## C4         0.13 -0.15 -0.12 -0.15 -0.13 -0.34 -0.38 -0.34  1.00  0.48  0.09
## C5         0.05 -0.12 -0.16 -0.24 -0.17 -0.25 -0.30 -0.34  0.48  1.00  0.06
## E1         0.11 -0.21 -0.21 -0.11 -0.25 -0.02  0.02  0.00  0.09  0.06  1.00
## E2         0.09 -0.23 -0.29 -0.19 -0.33 -0.09 -0.06 -0.08  0.20  0.26  0.47
## E3        -0.05  0.25  0.39  0.19  0.42  0.12  0.15  0.09 -0.08 -0.16 -0.33
## E4        -0.06  0.28  0.38  0.30  0.47  0.14  0.12  0.09 -0.11 -0.20 -0.42
## E5        -0.02  0.29  0.25  0.16  0.27  0.25  0.25  0.21 -0.24 -0.23 -0.30
## N1         0.17 -0.09 -0.08 -0.10 -0.20 -0.07 -0.02 -0.07  0.22  0.21  0.02
## N2         0.14 -0.05 -0.09 -0.14 -0.19 -0.04 -0.01 -0.06  0.16  0.25  0.01
## N3         0.10 -0.04 -0.04 -0.07 -0.14 -0.03  0.00 -0.07  0.21  0.24  0.05
## N4         0.05 -0.09 -0.13 -0.17 -0.20 -0.10 -0.05 -0.11  0.26  0.34  0.23
## N5         0.02  0.02 -0.04 -0.01 -0.08 -0.05  0.05 -0.01  0.20  0.17  0.05
## O1         0.01  0.13  0.15  0.06  0.16  0.17  0.16  0.09 -0.09 -0.08 -0.10
## O2         0.08  0.02  0.00  0.04  0.00 -0.11 -0.04 -0.03  0.21  0.14  0.04
## O3        -0.06  0.16  0.22  0.07  0.24  0.19  0.19  0.06 -0.08 -0.08 -0.22
## O4        -0.08  0.09  0.04 -0.04  0.02  0.11  0.06  0.02  0.05  0.14  0.08
## O5         0.11 -0.09 -0.05  0.02 -0.05 -0.12 -0.05 -0.01  0.20  0.06  0.10
## gender    -0.16  0.18  0.14  0.13  0.10  0.01  0.07  0.05 -0.08 -0.09 -0.13
## education -0.14  0.01  0.00 -0.02  0.01  0.03  0.00  0.05 -0.04  0.03  0.00
```

```r
round(cor(bfi, use = "complete"),2)
```

```
##              A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1
## A1         1.00 -0.34 -0.26 -0.14 -0.19  0.02  0.01 -0.01  0.10  0.02  0.12
## A2        -0.34  1.00  0.48  0.34  0.38  0.09  0.13  0.19 -0.14 -0.11 -0.24
## A3        -0.26  0.48  1.00  0.38  0.50  0.10  0.14  0.13 -0.12 -0.15 -0.22
## A4        -0.14  0.34  0.38  1.00  0.32  0.08  0.22  0.13 -0.16 -0.24 -0.14
## A5        -0.19  0.38  0.50  0.32  1.00  0.12  0.11  0.13 -0.12 -0.16 -0.25
## C1         0.02  0.09  0.10  0.08  0.12  1.00  0.43  0.32 -0.35 -0.25 -0.03
## C2         0.01  0.13  0.14  0.22  0.11  0.43  1.00  0.36 -0.38 -0.30  0.02
## C3        -0.01  0.19  0.13  0.13  0.13  0.32  0.36  1.00 -0.35 -0.35 -0.02
## C4         0.10 -0.14 -0.12 -0.16 -0.12 -0.35 -0.38 -0.35  1.00  0.48  0.10
## C5         0.02 -0.11 -0.15 -0.24 -0.16 -0.25 -0.30 -0.35  0.48  1.00  0.07
## E1         0.12 -0.24 -0.22 -0.14 -0.25 -0.03  0.02 -0.02  0.10  0.07  1.00
## E2         0.08 -0.24 -0.29 -0.20 -0.33 -0.10 -0.07 -0.09  0.21  0.26  0.47
## E3        -0.04  0.25  0.38  0.20  0.41  0.13  0.15  0.10 -0.09 -0.17 -0.33
## E4        -0.07  0.30  0.39  0.33  0.48  0.14  0.12  0.10 -0.12 -0.21 -0.42
## E5        -0.02  0.30  0.26  0.16  0.27  0.26  0.25  0.22 -0.23 -0.24 -0.31
## N1         0.16 -0.08 -0.07 -0.09 -0.19 -0.06 -0.02 -0.08  0.21  0.21  0.01
## N2         0.13 -0.04 -0.08 -0.15 -0.19 -0.03  0.00 -0.06  0.15  0.24  0.01
## N3         0.09 -0.02 -0.03 -0.07 -0.13 -0.01  0.01 -0.07  0.20  0.23  0.05
## N4         0.04 -0.09 -0.13 -0.16 -0.21 -0.09 -0.04 -0.13  0.28  0.35  0.23
## N5         0.01  0.02 -0.04  0.00 -0.08 -0.05  0.05 -0.04  0.21  0.18  0.04
## O1         0.00  0.11  0.14  0.04  0.15  0.18  0.16  0.09 -0.10 -0.09 -0.10
## O2         0.07  0.03  0.03  0.05  0.00 -0.13 -0.05 -0.03  0.21  0.12  0.06
## O3        -0.06  0.15  0.22  0.04  0.22  0.19  0.18  0.06 -0.07 -0.07 -0.21
## O4        -0.09  0.05  0.02 -0.06  0.00  0.08  0.03  0.00  0.07  0.14  0.08
## O5         0.11 -0.08 -0.04  0.04 -0.04 -0.13 -0.06  0.00  0.18  0.05  0.09
## gender    -0.17  0.21  0.16  0.13  0.11  0.00  0.06  0.04 -0.07 -0.09 -0.15
## education -0.14  0.02  0.00 -0.02  0.02  0.04  0.01  0.06 -0.04  0.04  0.00
```

With **pairwise deletion**, different sets of cases contribute to different correlations. That maximizes the sample sizes, but can lead to problems if the data are missing for some systematic reason.

Listwise deletion ("complete cases") doesn't have the same issue of biasing correlations, but does result in smaller samples and potentially limited generalizability.

A good practice is comparing the different matrices; if the correlation values are very different, this suggests that the missingness that affects pairwise deletion is systematic.

```r
round(cor(bfi, use = "pairwise")- cor(bfi, use = "complete"),2)
```

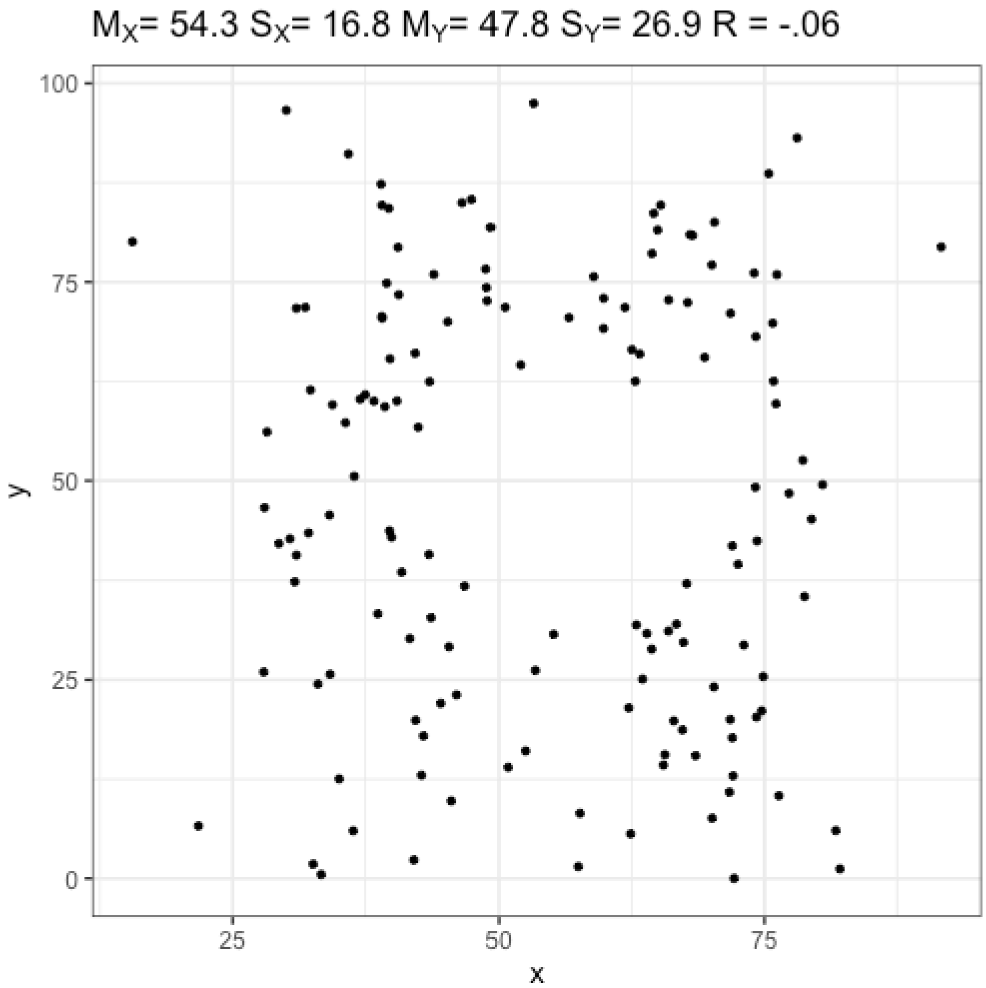
```
##              A1    A2    A3    A4    A5    C1    C2    C3    C4    C5    E1
## A1         0.00  0.00  0.00  0.00  0.00  0.01  0.00 -0.01  0.03  0.03 -0.01
## A2         0.00  0.00  0.00 -0.01  0.01  0.00  0.01  0.01 -0.01 -0.01  0.03
## A3         0.00  0.00  0.00 -0.02  0.00  0.00  0.00  0.00  0.00 -0.01  0.00
## A4         0.00 -0.01 -0.02  0.00 -0.01  0.01  0.01  0.00  0.01  0.00  0.03
## A5         0.00  0.01  0.00 -0.01  0.00  0.00  0.00  0.00 -0.01 -0.01  0.00
## C1         0.01  0.00  0.00  0.01  0.00  0.00  0.00 -0.01  0.01  0.00  0.00
## C2         0.00  0.01  0.00  0.01  0.00  0.00  0.00  0.00  0.00  0.00 -0.01
## C3        -0.01  0.01  0.00  0.00  0.00 -0.01  0.00  0.00  0.02  0.01  0.02
## C4         0.03 -0.01  0.00  0.01 -0.01  0.01  0.00  0.02  0.00 -0.01 -0.01
## C5         0.03 -0.01 -0.01  0.00 -0.01  0.00  0.00  0.01 -0.01  0.00  0.00
## E1        -0.01  0.03  0.00  0.03  0.00  0.00 -0.01  0.02 -0.01  0.00  0.00
## E2         0.01  0.01  0.00  0.01  0.00  0.01  0.01  0.01 -0.01  0.00  0.00
## E3         0.00  0.00  0.00 -0.01  0.00 -0.02  0.00 -0.02  0.01  0.01  0.01
## E4         0.01 -0.02 -0.02 -0.03 -0.01  0.00  0.00 -0.01  0.01  0.01  0.00
## E5         0.00  0.00 -0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.01  0.00
## N1         0.01 -0.01 -0.02  0.00  0.00 -0.01  0.00  0.01  0.01  0.01  0.01
## N2         0.01 -0.01  0.00  0.00  0.00 -0.01 -0.01  0.00  0.01  0.01  0.01
## N3         0.01 -0.02 -0.01  0.00 -0.01 -0.02 -0.01  0.01  0.01  0.01  0.00
## N4         0.01  0.00  0.00 -0.01  0.01 -0.01 -0.01  0.02 -0.02 -0.01  0.00
## N5         0.01  0.00  0.00  0.00  0.00  0.00  0.00  0.02 -0.02 -0.01  0.01
## O1         0.01  0.02  0.00  0.02  0.02 -0.01  0.01  0.00  0.01  0.01  0.00
## O2         0.01 -0.02 -0.03 -0.01  0.00  0.02  0.01  0.00  0.00  0.02 -0.01
## O3         0.00  0.02  0.01  0.03  0.02  0.00  0.01  0.01 -0.01 -0.01  0.00
## O4         0.01  0.03  0.01  0.02  0.01  0.03  0.03  0.02 -0.02  0.00 -0.01
## O5         0.01 -0.01 -0.01 -0.01 -0.01  0.01  0.00 -0.01  0.01  0.01  0.01
## gender     0.01 -0.03 -0.02  0.00 -0.01  0.01  0.01  0.01 -0.01  0.00  0.00
## education  0.00 -0.01 -0.01  0.00  0.00 -0.01 -0.01 -0.01  0.00 -0.01  0.00
```

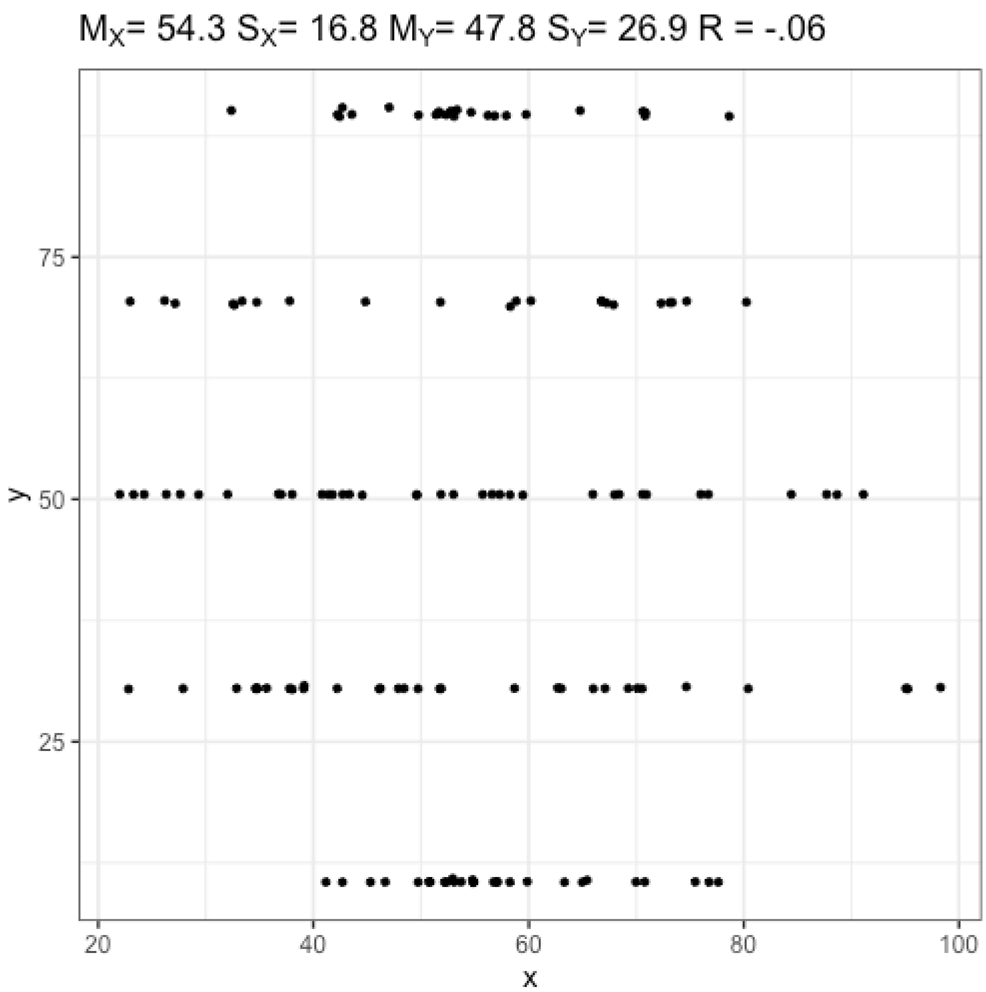
Visualizing correlations

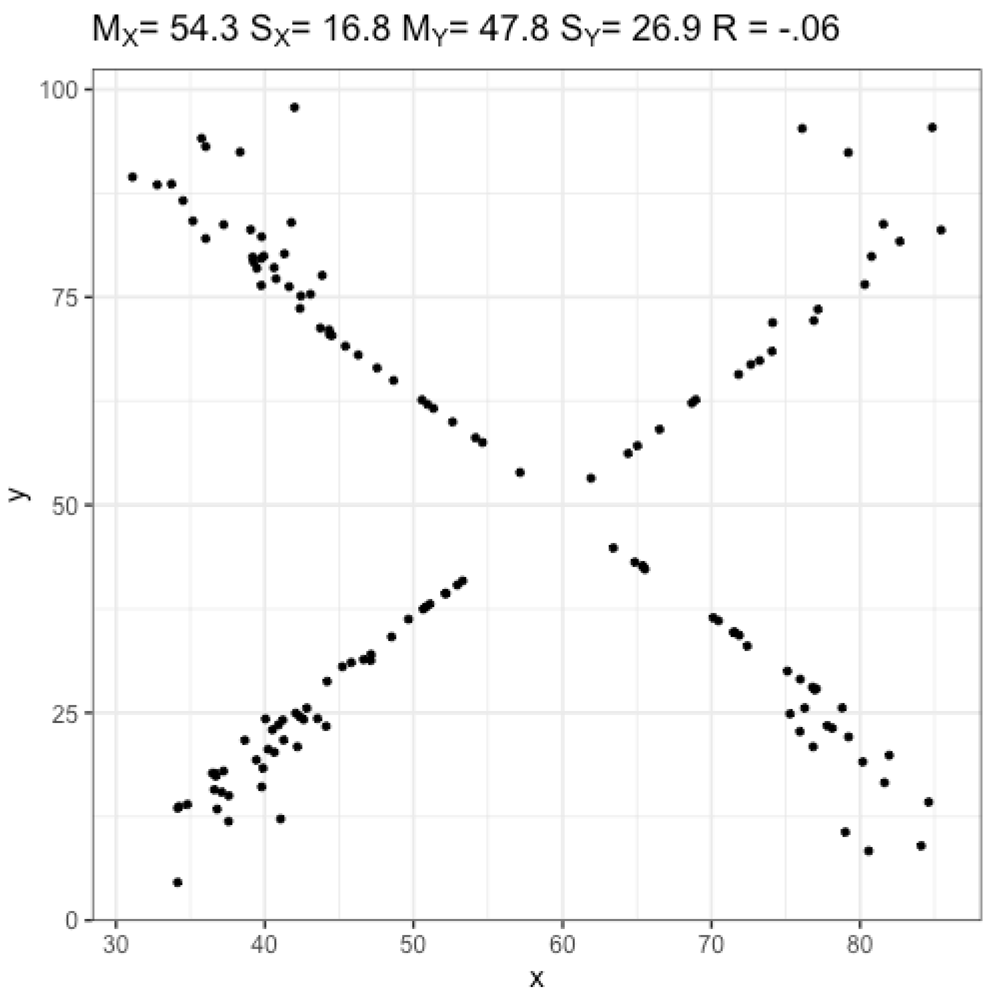
For a single correlation, best practice is to visualize the relationship using a scatterplot. A best fit line is advised, as it can help clarify the strength and direction of the relationship.

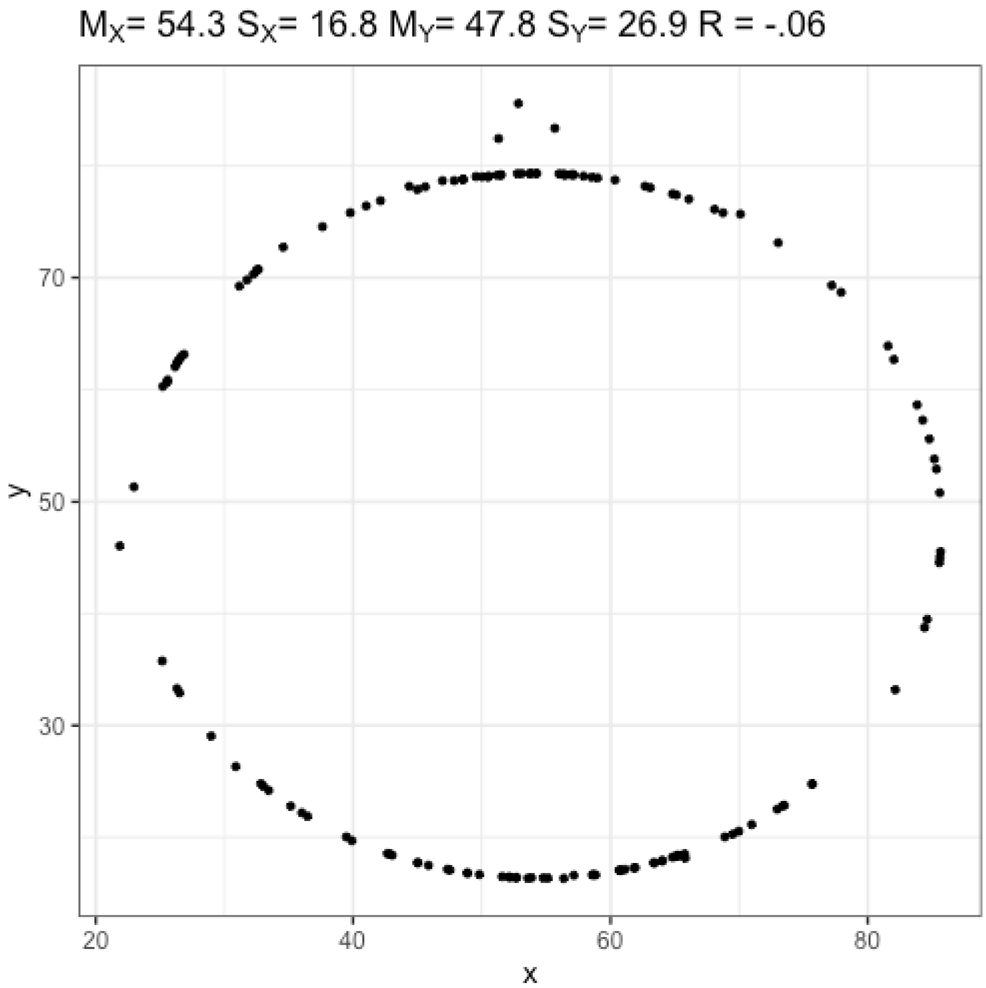
http://guessthecorrelation.com/

See also: Interpreting Correlations

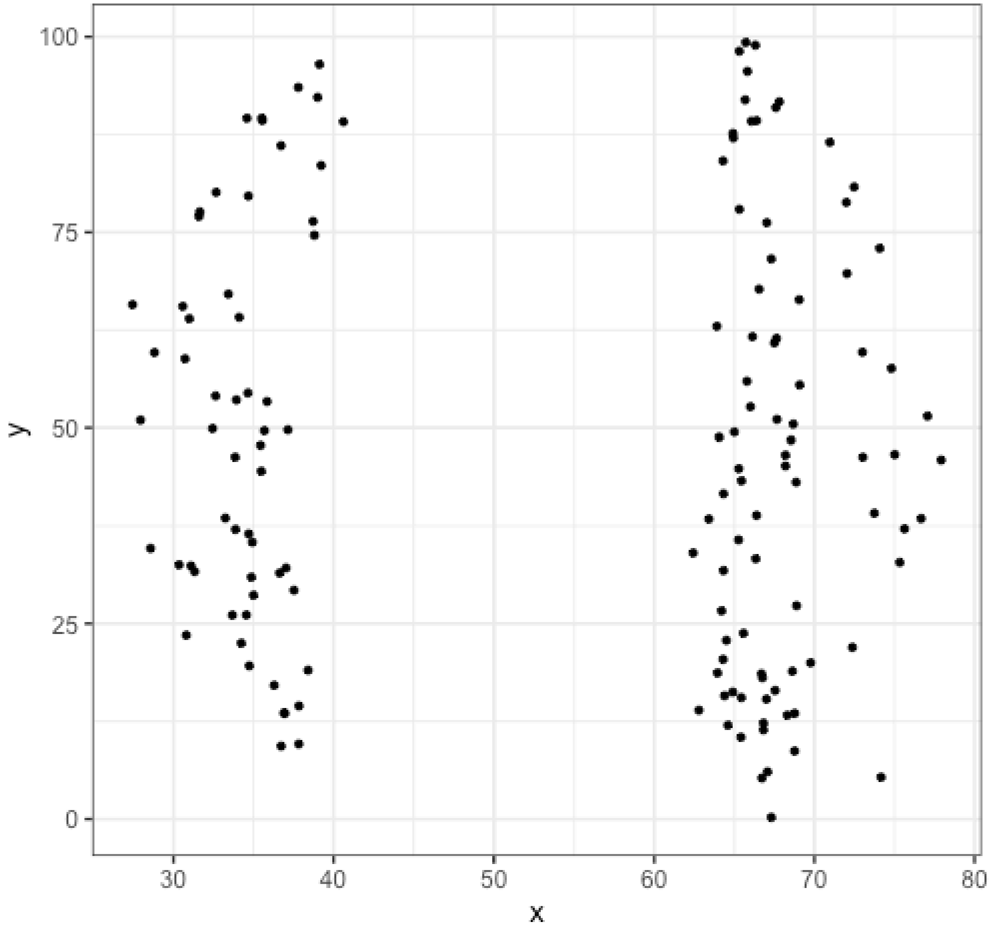

M_X= 54.3 S_X= 16.8 M_Y= 47.8 S_Y= 26.9 R = -.06
100
75
50
25
0
y
25
50
75
x

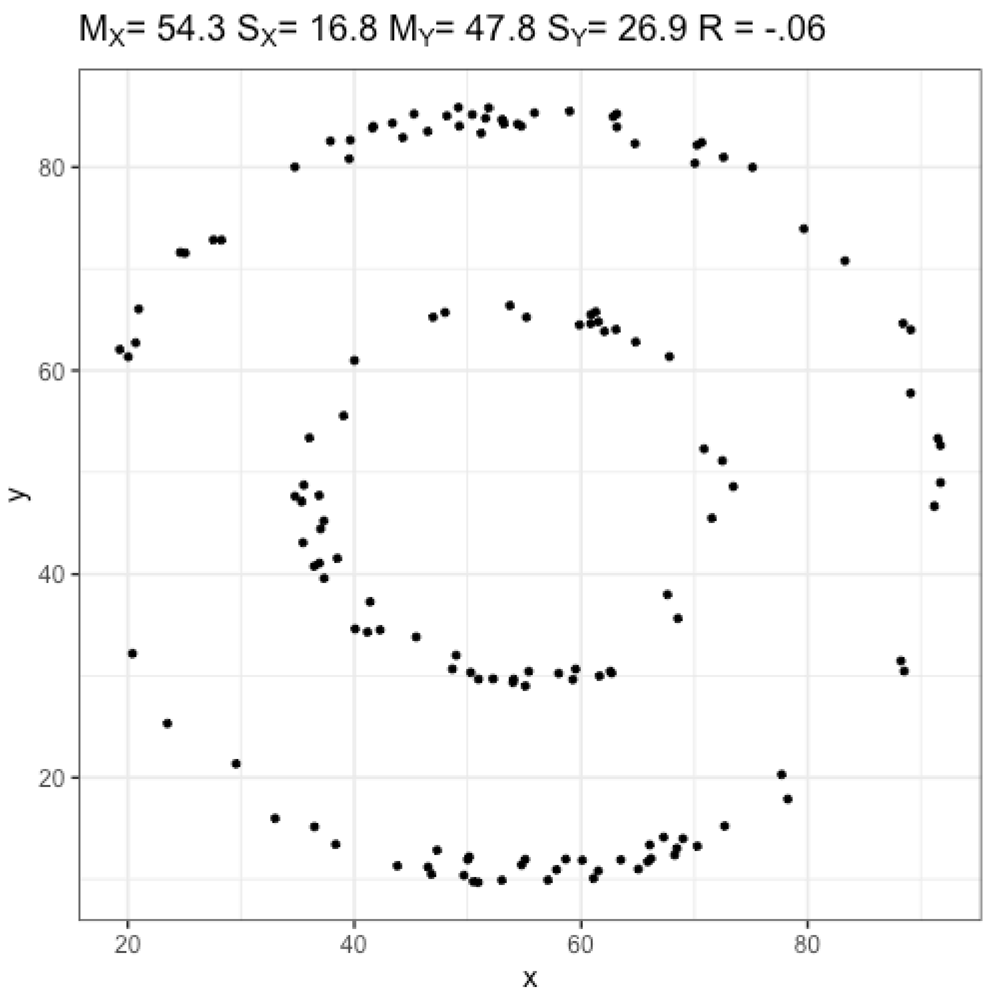

MX= 54.3 SX= 16.8 MY= 47.8 SY= 26.9 R = -.06
75
50
25
y
20
40
60
80
100
x


M_X= 54.3 S_X= 16.8 M_Y= 47.8 S_Y= 26.9 R = -.06
100
75
50
25
0
y
30
40
50
60
70
80
x

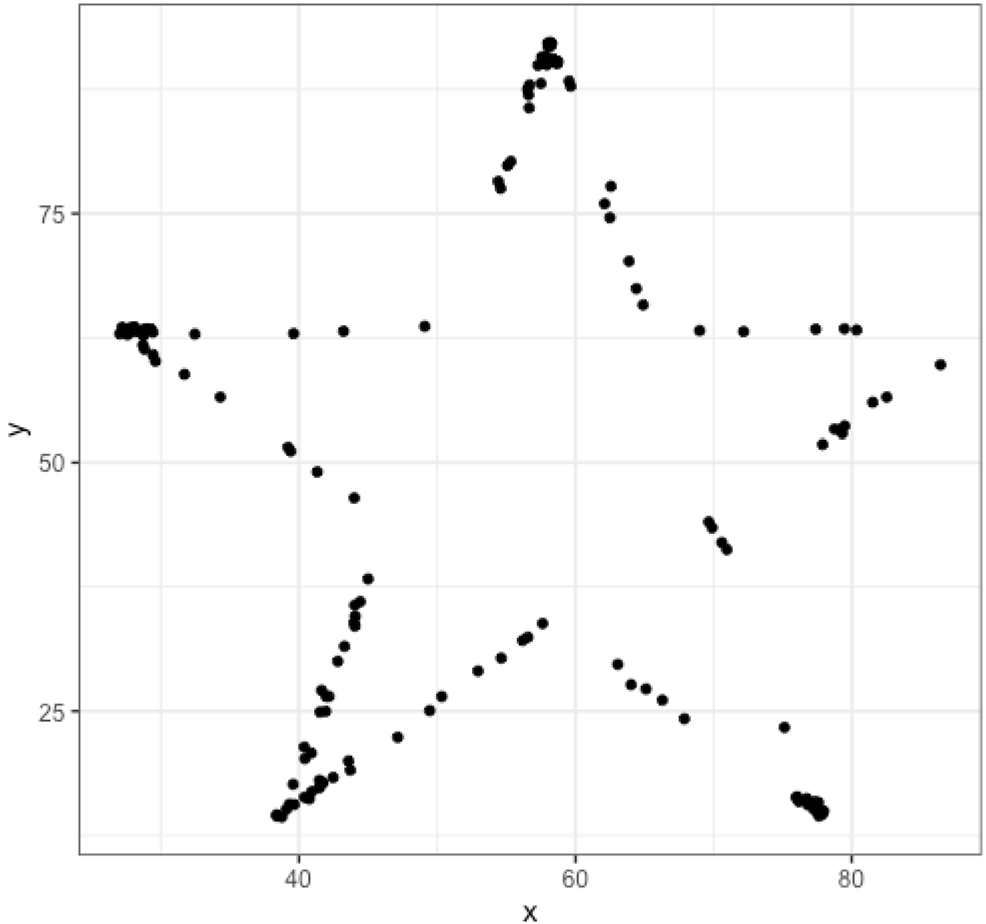

M_X= 54.3 S_X= 16.8 M_Y= 47.8 S_Y= 26.9 R = -.06
y
70
50
30
20
40
60
80
x

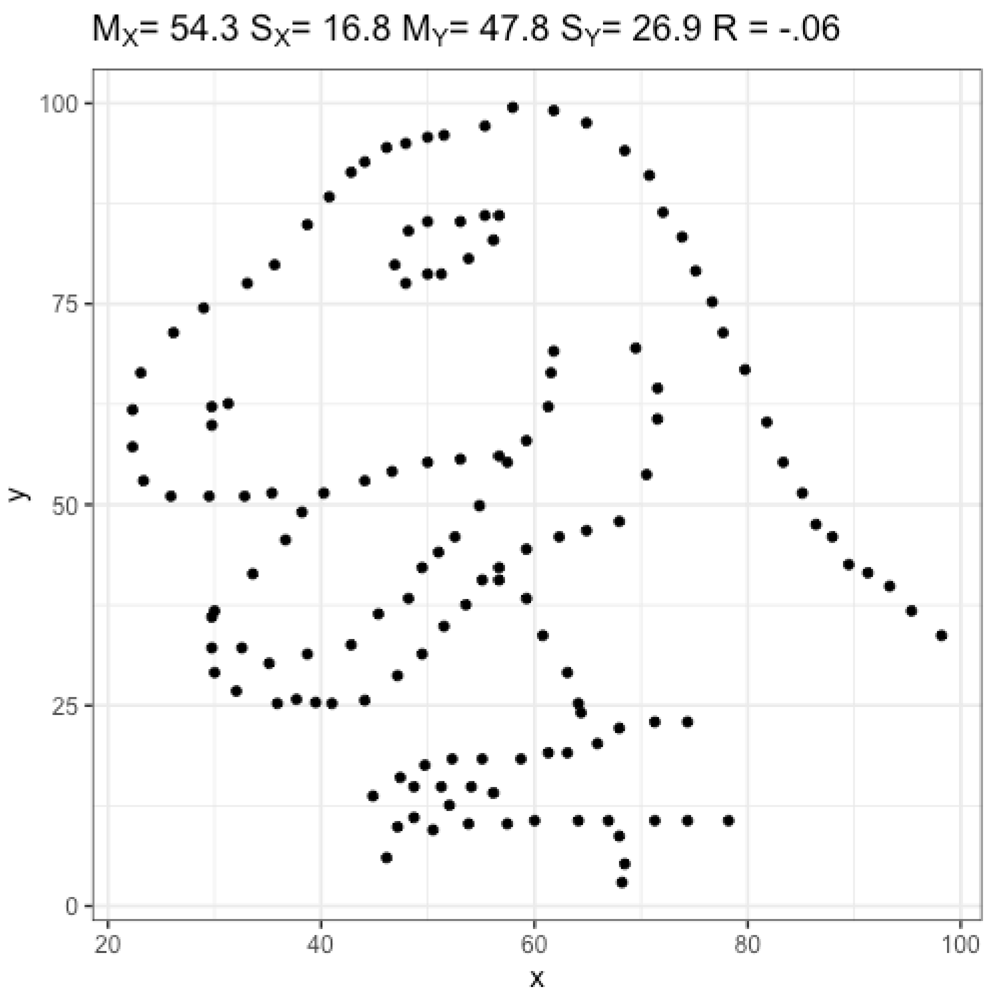
$M_X = 54.3$ $S_X = 16.8$ $M_Y = 47.8$ $S_Y = 26.9$ $R = -.06$




M_X= 54.3 S_X= 16.8 M_Y= 47.8 S_Y= 26.9 R = -.06
80
60
40
20
y
20
40
60
80
x

M_X= 54.3 S_X= 16.8 M_Y= 47.8 S_Y= 26.9 R = -.06





M_X= 54.3 S_X= 16.8 M_Y= 47.8 S_Y= 26.9 R = -.06
100
75
50
25
0
y
20
40
60
80
100
x

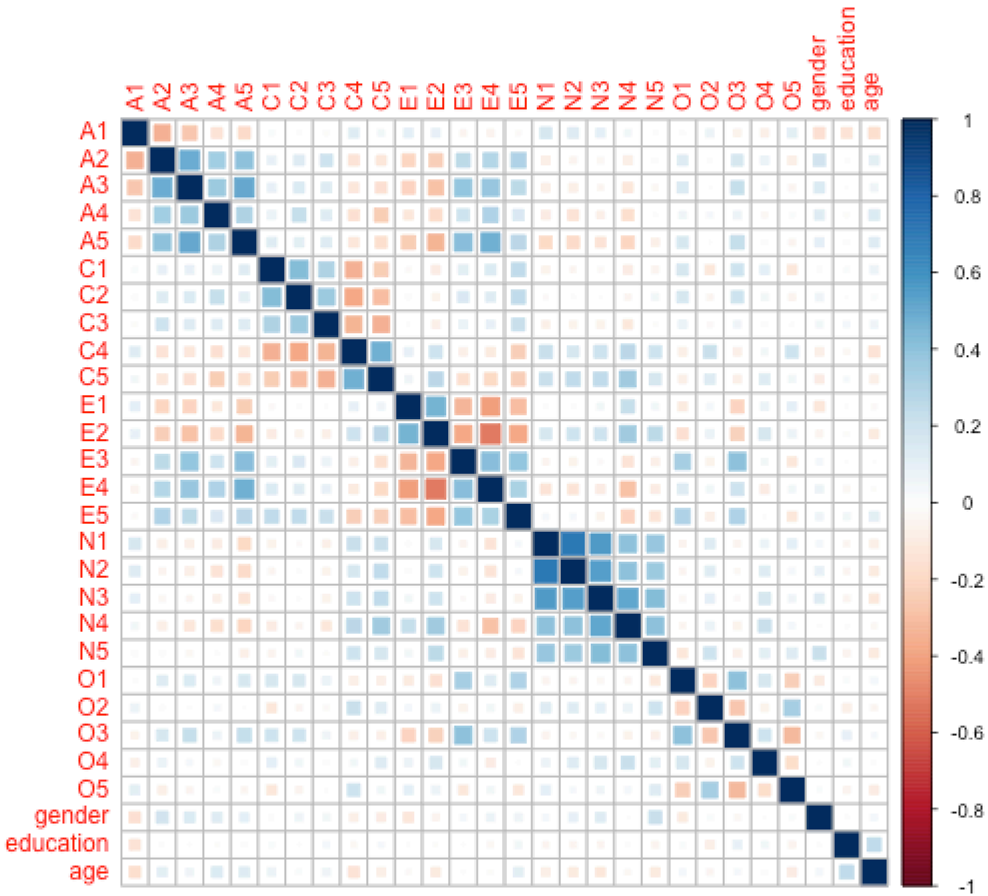
Visualizing correlation matrices

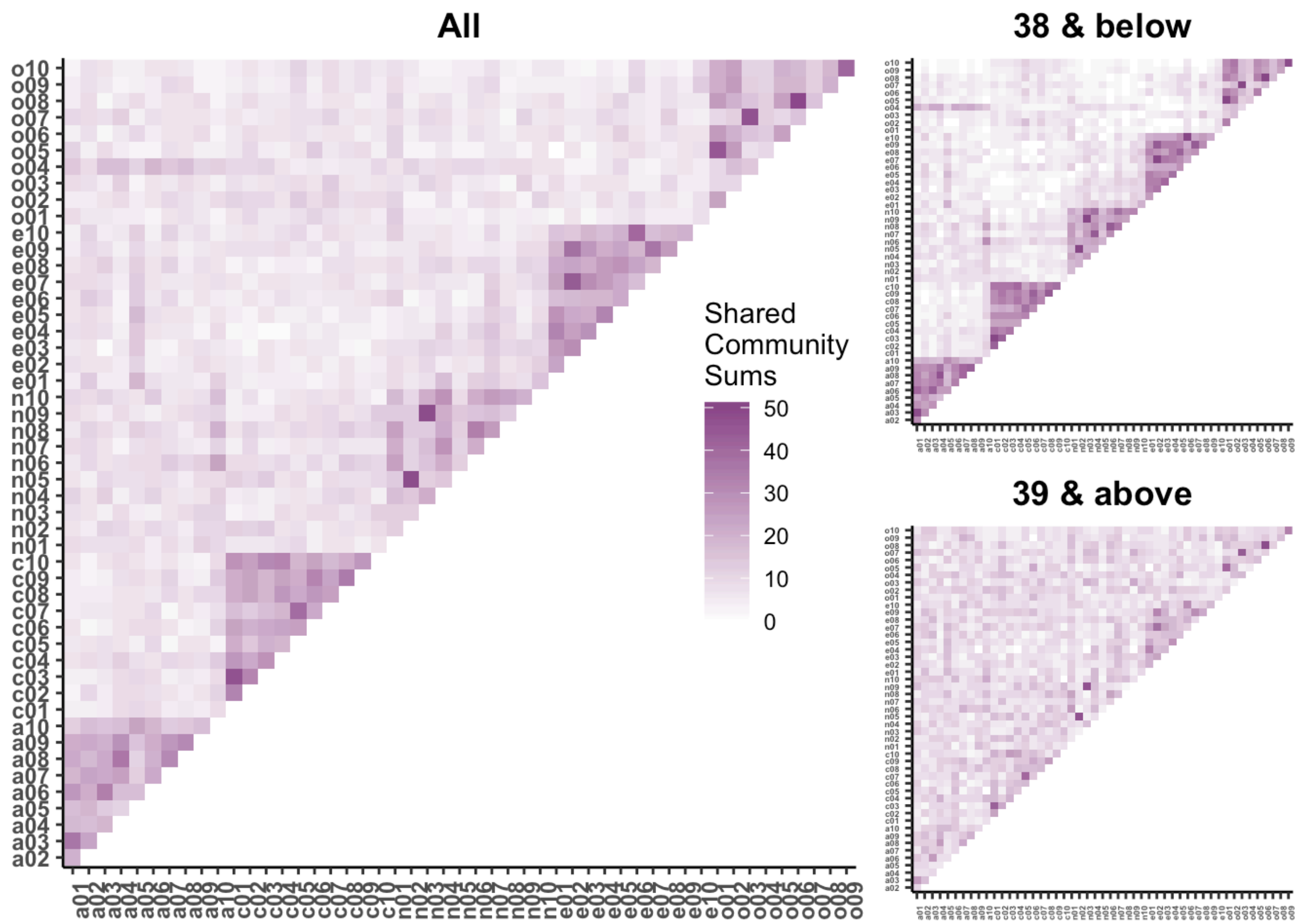
A single correlation can be informative; a correlation matrix is more than the sum of its parts.

Correlation matrices can be used to infer larger patterns of relationships. You may be one of the gifted who can look at a matrix of numbers and see those patterns immediately. Or you can use **heat maps** to visualize correlation matrices.

```
library(corrplot)
```

```r
corrplot(cor(bfi, use = "pairwise"), method = "square")
```





Beck, Condon, & Jackson, 2019

Other correlation tests:

1. Set of correlations
2. Dependent correlations (i.e., within same group). These are more easily tested via Structural Equation Modeling (SEM)
3. Intra Class Correlation (ICC)

- Again, best to do these tests in another framework (e.g., interaction, SEM, MLM)

Types of correlations

- Many ways to get at relationship between two variables
- Statistically the different types are almost exactly the same
- Exist for historical reasons

Types of correlations

1. Point Biserial
   - continuous and dichtomous
2. Phi coefficient
   - both dichotomous
3. Spearman & Kendall rank order
   - ranked data (nonparametric)
   - Spearman for larger samples, Kendall for smaller samples (or a lot of ties in rank ordering)
4. Biserial (assumes dichotomous is continuous)
5. Tetrachoric (assumes dichotomous is continuous)

Do the special cases matter?

For Spearman, you'll get a different answer.

```
x = rnorm(n = 10); y = rnorm(n = 10) #randomly generate 10 numbers f
```

```
head(cbind(x,y))
```

```
##               x          y
## [1,] -0.4088210  0.3471781
## [2,]  0.1629271  0.4593525
## [3,]  1.2753450  0.2569774
## [4,]  0.2815652 -0.2240819
## [5,]  1.3119315  0.1583264
## [6,]  0.6282713  1.2058852
```

```
cor(x,y, method = "pearson")
```

```
## [1] -0.1179297
```

```
head(cbind(x,y, rank(x), rank(y)
```

```
##               x          y
## [1,] -0.4088210  0.3471781  1 6
## [2,]  0.1629271  0.4593525  2 7
## [3,]  1.2753450  0.2569774  8 5
## [4,]  0.2815652 -0.2240819  3 1
## [5,]  1.3119315  0.1583264 10 3
## [6,]  0.6282713  1.2058852  6 9
```

```
cor(x,y, method = "spearman")
```

Do the special cases matter?

If your data are naturally binary, no difference between Pearson and point-biserial.

```
x = rnorm(n = 10); y = rbinom(n = 10, size = 1, prob = .3)
head(cbind(x,y))
```

```
##                x y
## [1,]  1.3142775 0
## [2,] -1.1809155 0
## [3,]  0.8890751 1
## [4,]  0.5232146 1
## [5,] -1.0808459 1
## [6,]  0.5172158 0
```

```
cor(x,y, method = "pearson")
```

```
## [1] 0.02748352
```

```
ltm::biserial.cor(x,y, level = 2
```

```
## [1] 0.02748352
```

Do the special cases matter?

If your data are artificially binary, there can be big differences. DON'T USE MEDIAN SPLITS!

```
x = rnorm(n = 10); y = rnorm(n = 10)
```

```
head(cbind(x,y))
```

```
##                x          y
## [1,]  0.2991754  0.1865549
## [2,]  2.3783854  0.1056268
## [3,] -2.5252925 -0.5656886
## [4,] -0.3590271 -0.0981752
## [5,]  0.5147804 -0.8599081
## [6,] -0.9583130 -0.4705694
```

```
cor(x,y, method = "pearson")
```

```
## [1] 0.4159568
```

```
d_y = ifelse(y < median(y), 0, 1
head(cbind(x,y, d_y))
```

```
##                x          y d_y
## [1,]  0.2991754  0.1865549   1
## [2,]  2.3783854  0.1056268   1
## [3,] -2.5252925 -0.5656886   0
## [4,] -0.3590271 -0.0981752   1
## [5,]  0.5147804 -0.8599081   0
## [6,] -0.9583130 -0.4705694   0
```

```
ltm::biserial.cor(x,d_y, level =
```

Next time....

Partial and Semi-partial correlations

(remaining slides include more examples, if you want some practice)

Example

The correlation between midterm exam grades and final exam grades was .56. The class size was 104. Is this statistically significant?

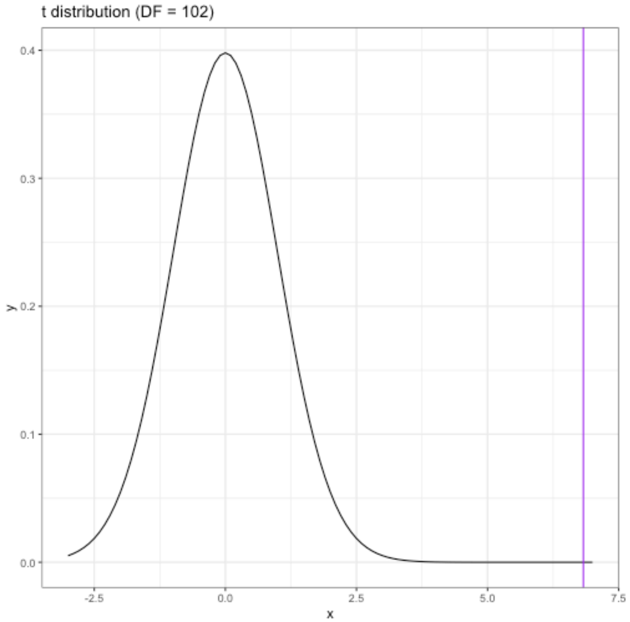
Using t-method

$$SE_r = \sqrt{\frac{1-r^2}{N-2}} = \sqrt{\frac{1-.56^2}{104-2}} = 0.08$$

$$t = \frac{r}{SE_r} = \frac{0.56}{0.08} = 6.83$$

Probability of getting a *t* statistic of 6.83 or greater is $3.19x10^{-10}$

Probability of getting *t* statistic of 6.83 or more extreme is $6.38x10^{-10}$



Example

The correlation between midterm exam grades and final exam grades was .56. The class size was 104. Is this statistically significantly different from .40?

$$z^{'} = \frac{1}{2} ln \frac{1+r}{1-r} = \frac{1}{2} ln \frac{1+0.56}{1-0.56} = 0.63$$

$$z^{'}_{H_0} = \frac{1}{2} ln \frac{1+r}{1-r} = \frac{1}{2} ln \frac{1+0.4}{1-0.4} = 0.42$$

$$SE_z = \frac{1}{\sqrt{104-3}} = 0.1$$

$$Z_{\text{statistic}} = \frac{z' - \mu}{SE_z} = \frac{0.63 - 0.42}{0.1} = 2.1$$

```
stat
```

```
## [1] 2.102276
```

```
pnorm(stat, lower.tail = F)
```

```
## [1] 0.01776456
```

```
pnorm(stat, lower.tail = F)*2
```

```
## [1] 0.03552913
```

pagedown::chrome_print("4-correlation.html")