Correlation

Research Methods for Political Science

Thomas Chadefaux

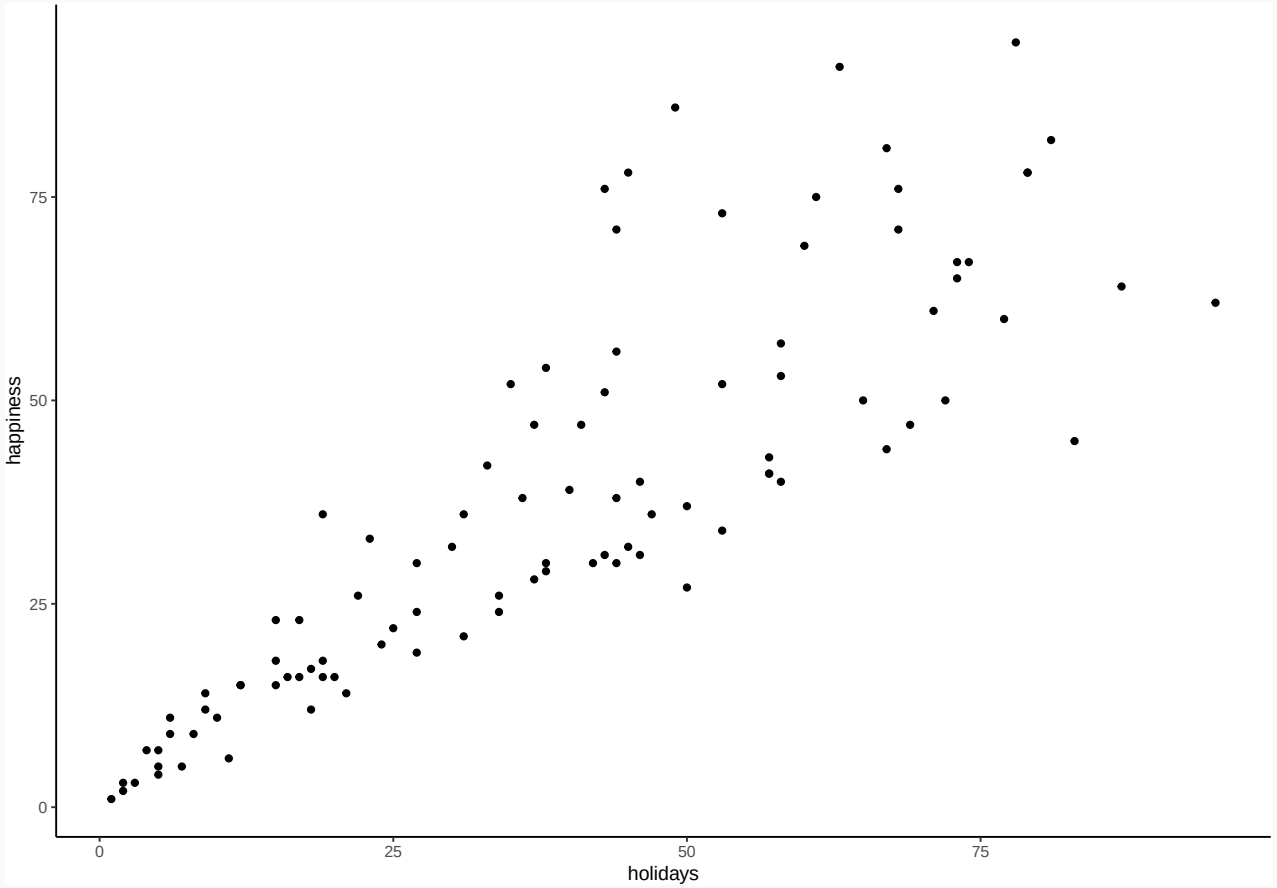
Trinity College Dublin

Correlation: introduction

Sample data

subject	holidays	happiness
1	1	1
2	2	2
3	3	3
4	2	3
5	5	5
6	5	4
7	4	7
8	7	5
9	5	7
10	6	9

Sample data



It looks like the more holidays you have the happier you will be, and vice-versa $\rightarrow$ A **positive correlation**.

Positive, Negative, and No-Correlation

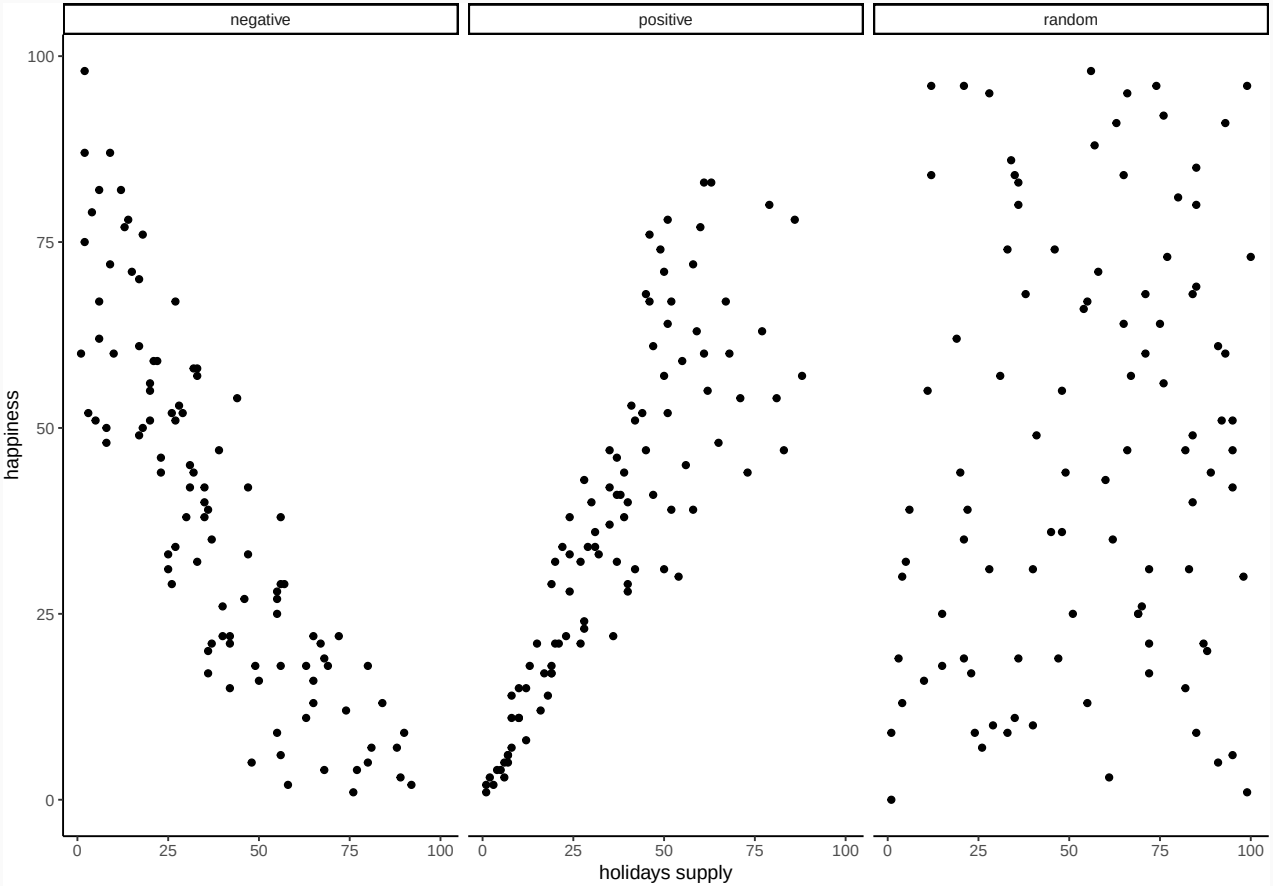


Figure 1: Three scatterplots showing negative, positive, and zero correlation

Put a number on it

We'd like to assign a number to these correlations.

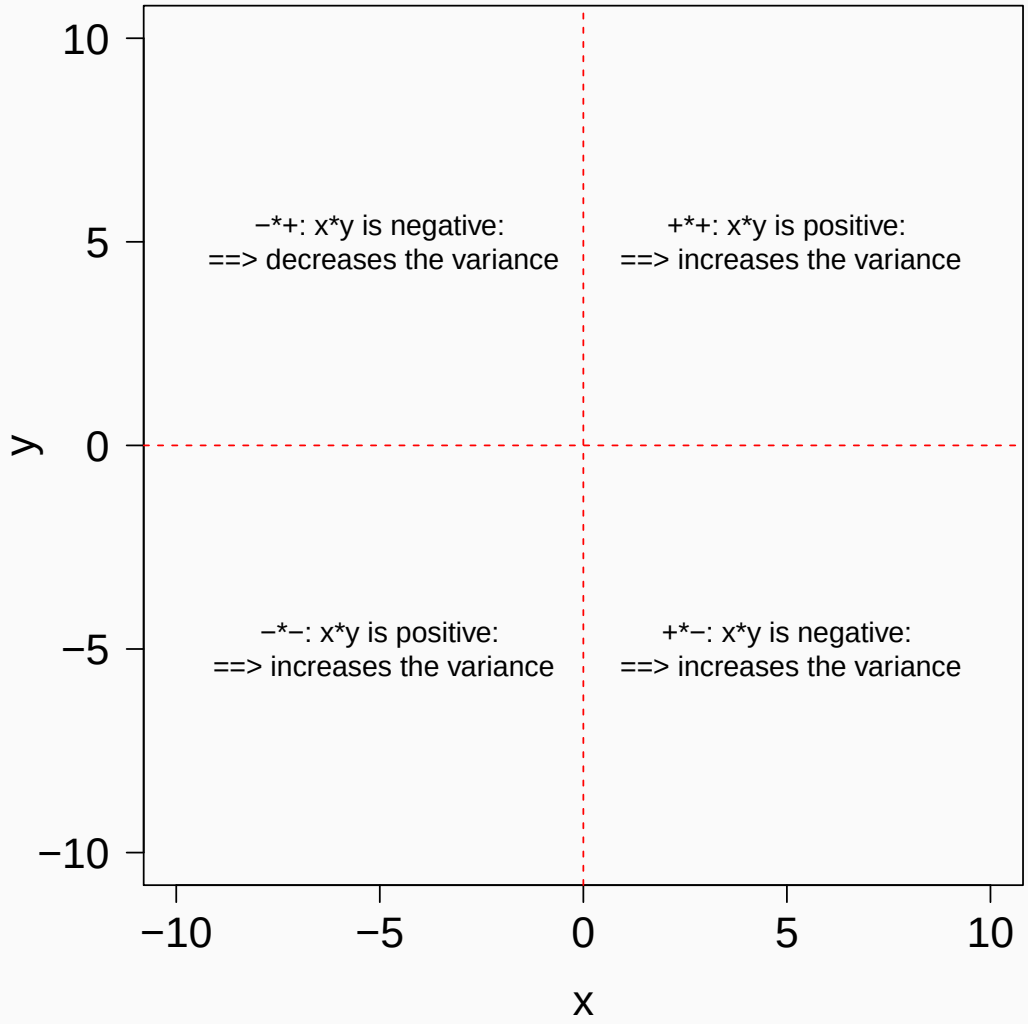
- $+1$ would mean a large *positive* correlation
- -1 would mean a large *negative* correlation
- 0 would mean no correlation
- 0.5 would mean some *positive* correlation, and so on for every number between 0 and 1

Pearson's r

The good news

Correlation is really just covariance, with a small twist.

Remember when we discussed the covariance?



The problem with the covariance

Problem: the covariance has no 'natural' scale.

- If I double the size of every value x and y, the covariance increases, even though the linear relationship is the same.
- If I add points, the covariance changes.

As a result, you cannot easily compare the covariance of two different samples

The problem with the covariance

An example:

```
x <- c(1,2,3,4)
y <- c(4,3,2,6)
cov(x,y)
```

```
## [1] 0.8333333
```

```
cov(2*x, 2*y)
```

```
## [1] 3.333333
```

Correlation: Pearson's r

- Pearson's r, or coefficient of correlation, is a "normalized" variance. It is defined as:

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

- Note that this is just the covariance divided by the standard deviations of x and y. Doing this normalizes the covariance to a number between -1 and 1. A negative number means that x and y are negatively correlated (a negative slope).

Correlation: Pearson's *r*

In R, all you need is 'cor(x,y)' Note that the correlation coefficient is not affected by the units.

```
x <- c(1,2,3,6)
y <- c(3,4,2,5)
cor(x,y)
```

```
## [1] 0.5976143
```

```
cor(2*x, 2*y)
```

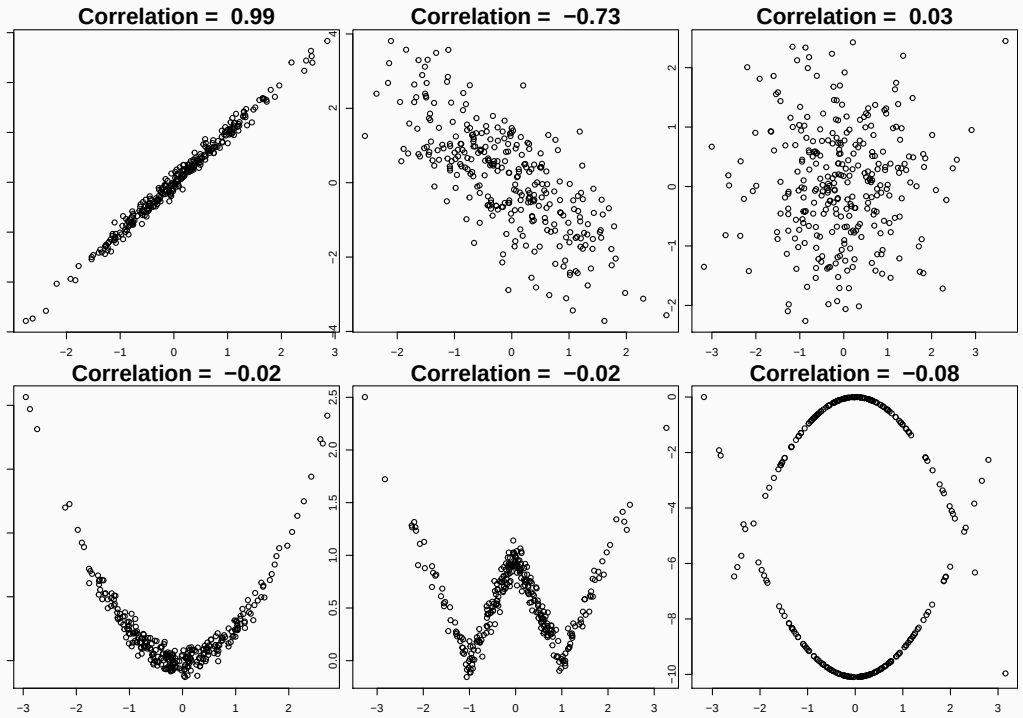
```
## [1] 0.5976143
```

```
# BUT a non-linear change does affect the correlation
cor(x^2, y^2)
```

```
## [1] 0.7632795
```

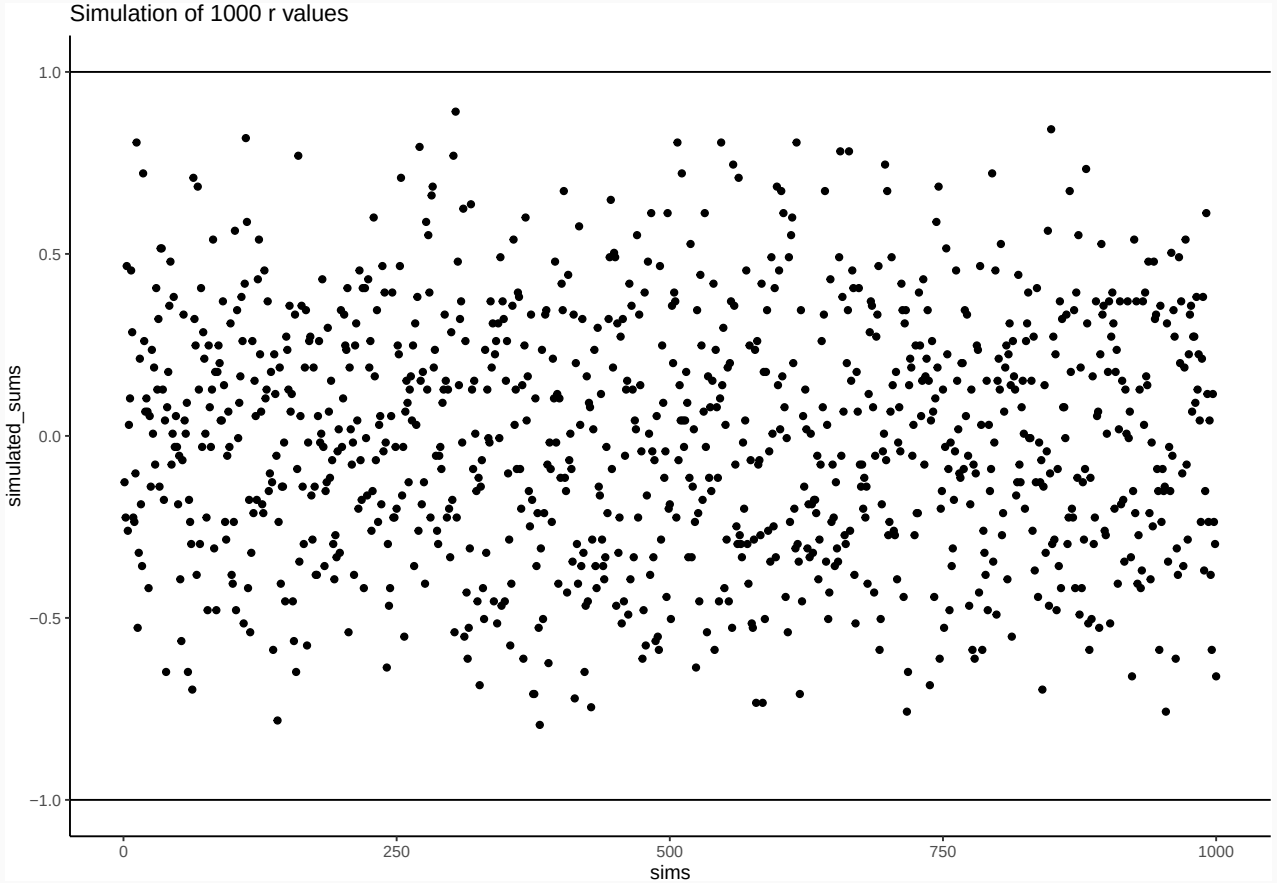
Correlation is a measure of linear relationship

Some examples where correlation fails to uncover a relationship between two variables:



test

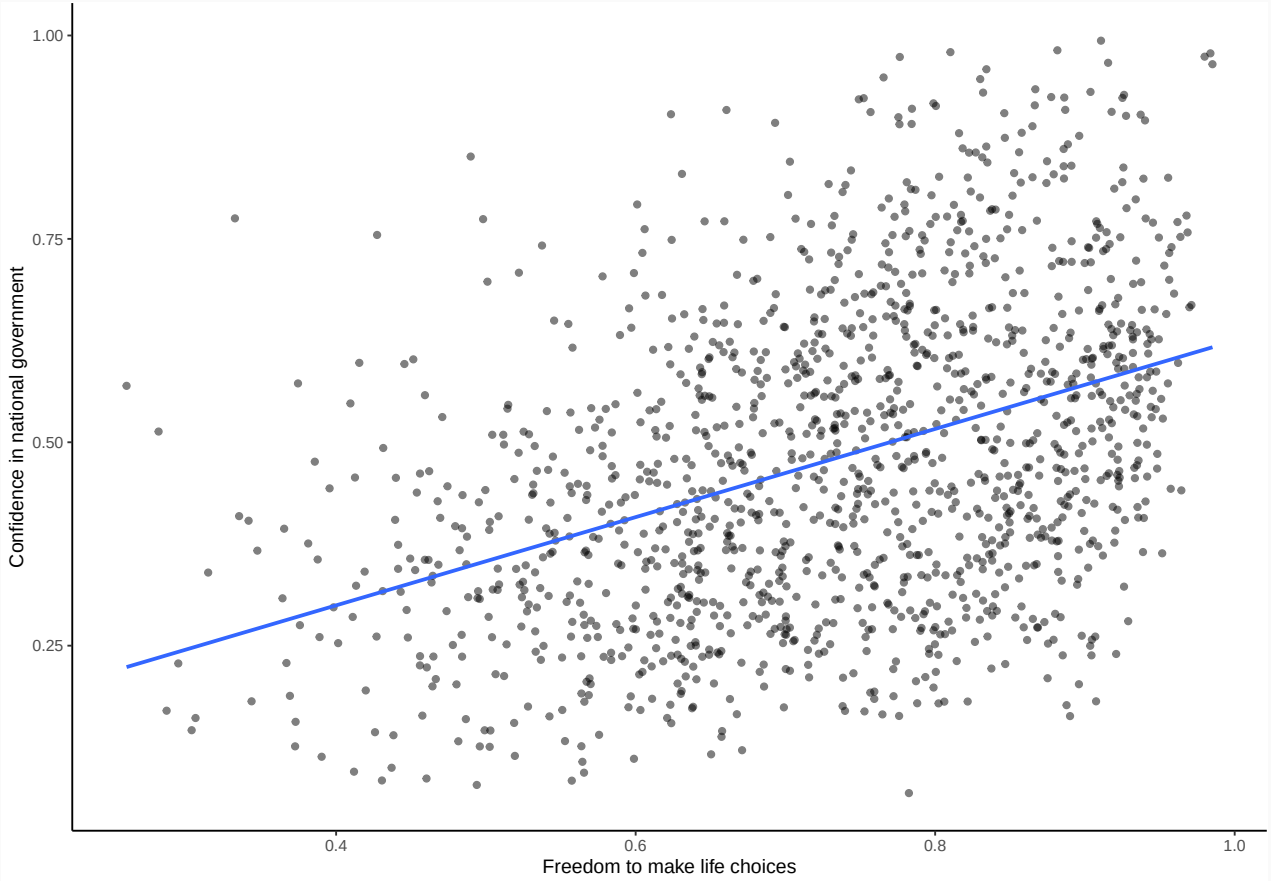
Does Pearson's *r* really stay between -1 and 1 no matter what? Let's generate 10 random "x" numbers, 10 random y numbers, and calculate their correlation. I did this 1000 times and plotted their correlation below. All of the dots are between -1 and 1!



An example from the real world

Data from the world happiness report (2018).

Relationship between how much freedom people thought they had to make life choices and how confident people were in their national government. Each dot represents the mean for one country.



An example from the real world

The actual correlation (i.e., Pearson's r) is:

```
## [1] 0.4080963
```

Correlation does not equal causation

Ok but why not?

Correlation does not equal causation

1. Even when there is causation, there might not be obvious correlation

Correlation does not equal causation:

2. Confounding variables

Correlation does not equal causation

There is randomness in the world

You can find a positive or negative correlation between two measures, even when they have absolutely nothing to do with one another.

Simulation of random correlations

By pure chance, we can get high values of Person's r

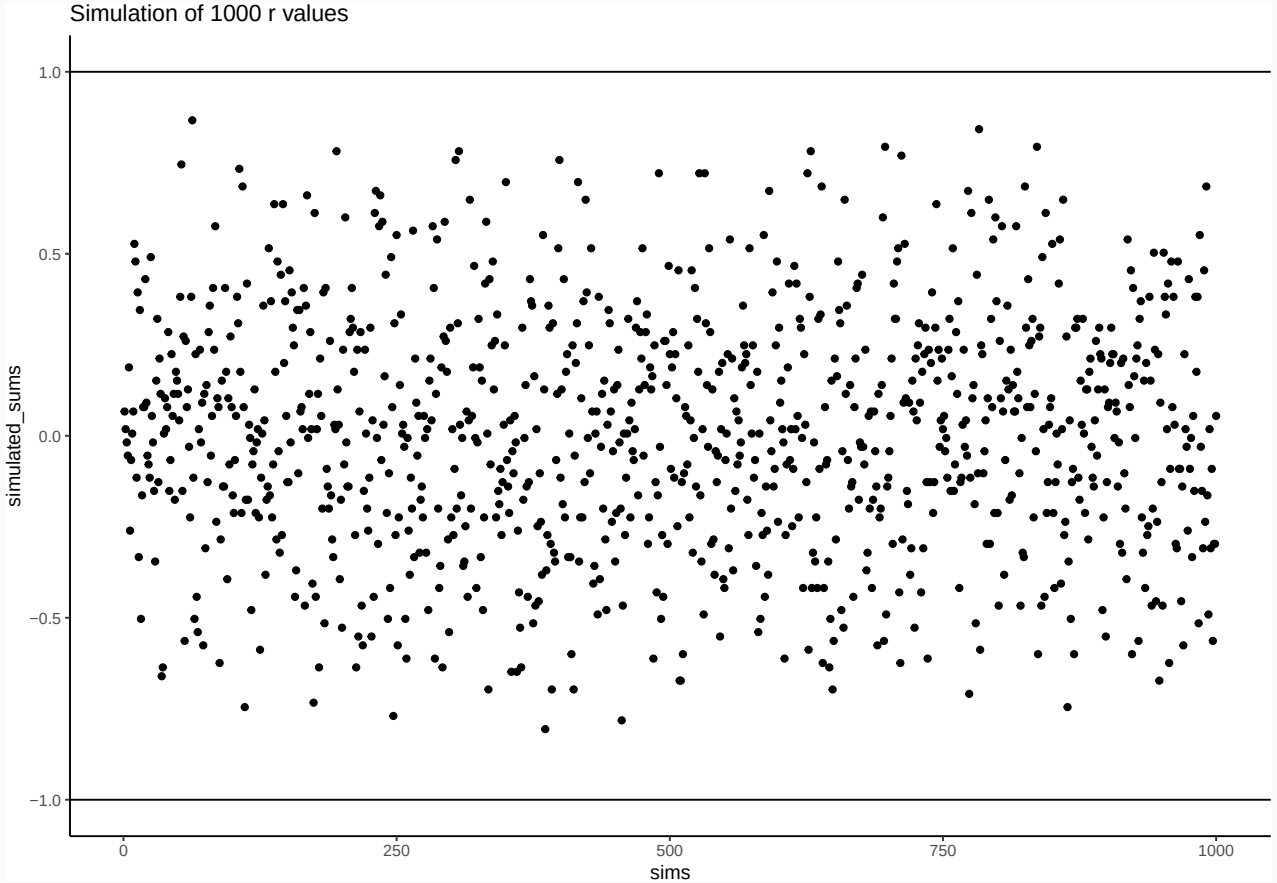


Figure 3: A simulation of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders

Is correlation useless then?

Does this mean correlation is useless, if correlation can all be due to chance?

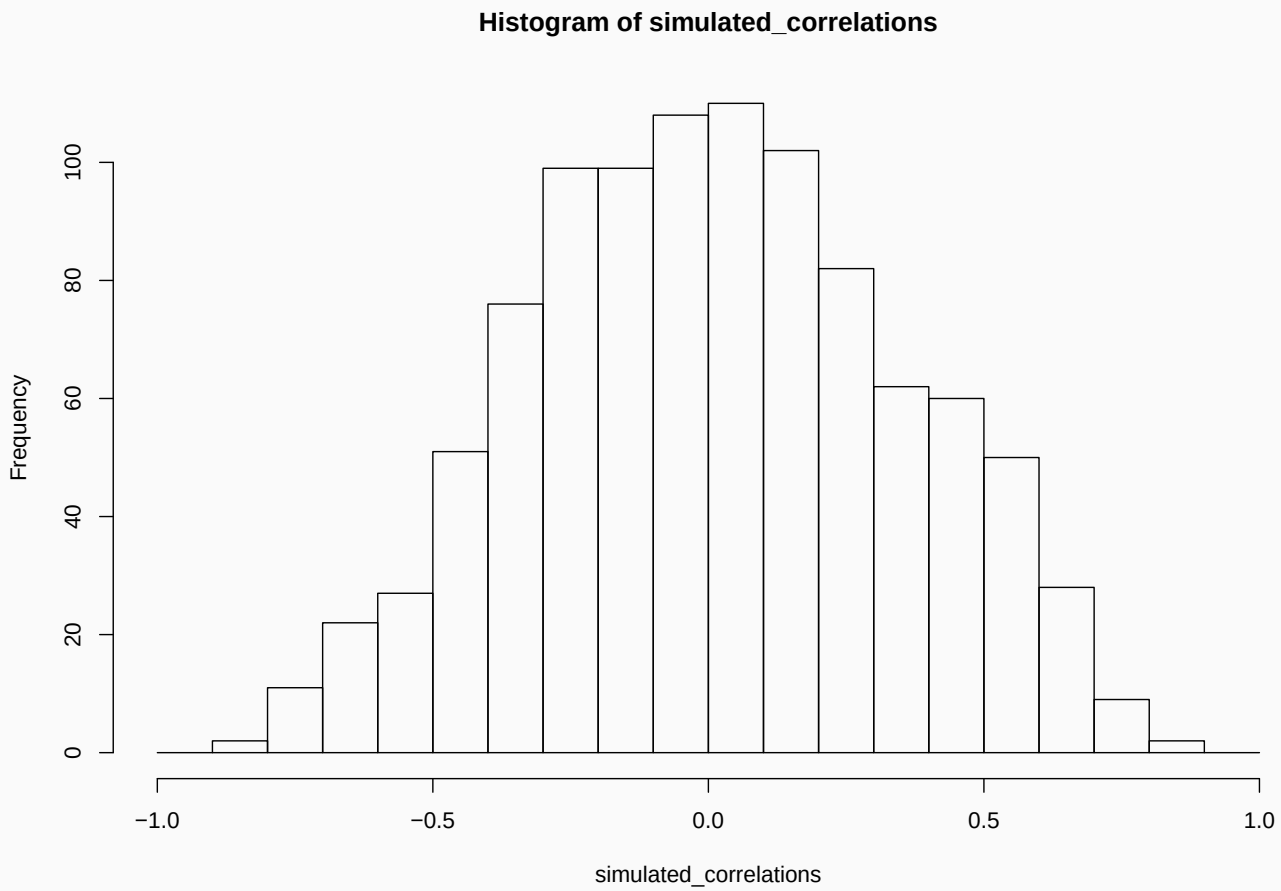


Figure 4: A histogram showing the frequency distribution of r-values for completely random values between an X and Y variable (sample-size=10). A full range of r-values can be obtained by chance alone. Larger r-values

Is correlation useless then?

Note that most of the values of r are close to 0. Large values are rare, so large values do give us confidence that there is indeed a relationship

You can think of this histogram as the window of chance.

Increasing sample-size decreases opportunity for spurious correlation

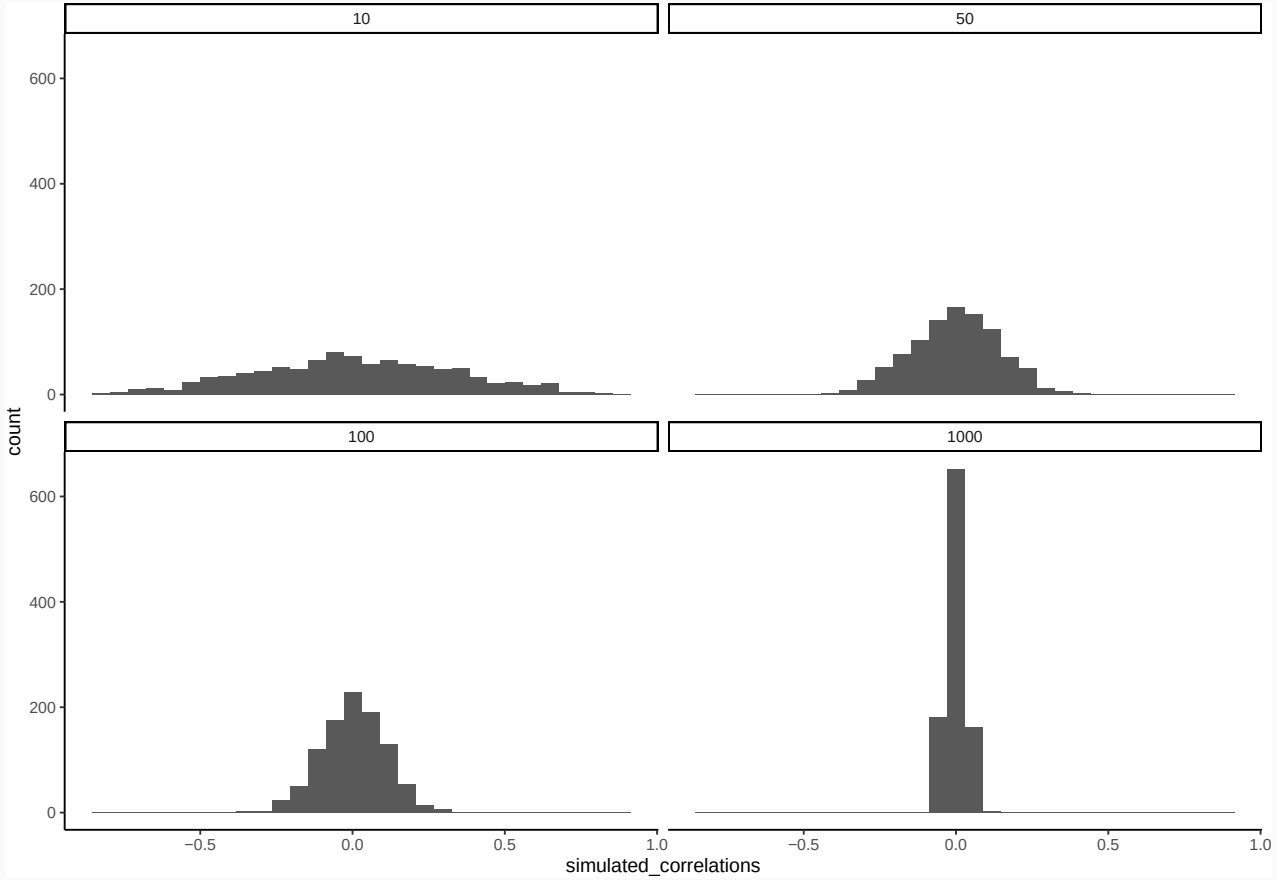


Figure 5: Four histograms showing the frequency distributions of r-values between completely random X and Y variables as a function of sample-size.

test

The width or range of each histogram shrinks as the sample-size increases.

Statistical significance of the correlation coefficient.

The null hypothesis for a correlation is that there is no correlation, i.e., r=0.

Three ways to calculate the statistical significance:

1. Bootstrap and get the p-value
2. Use the formula:

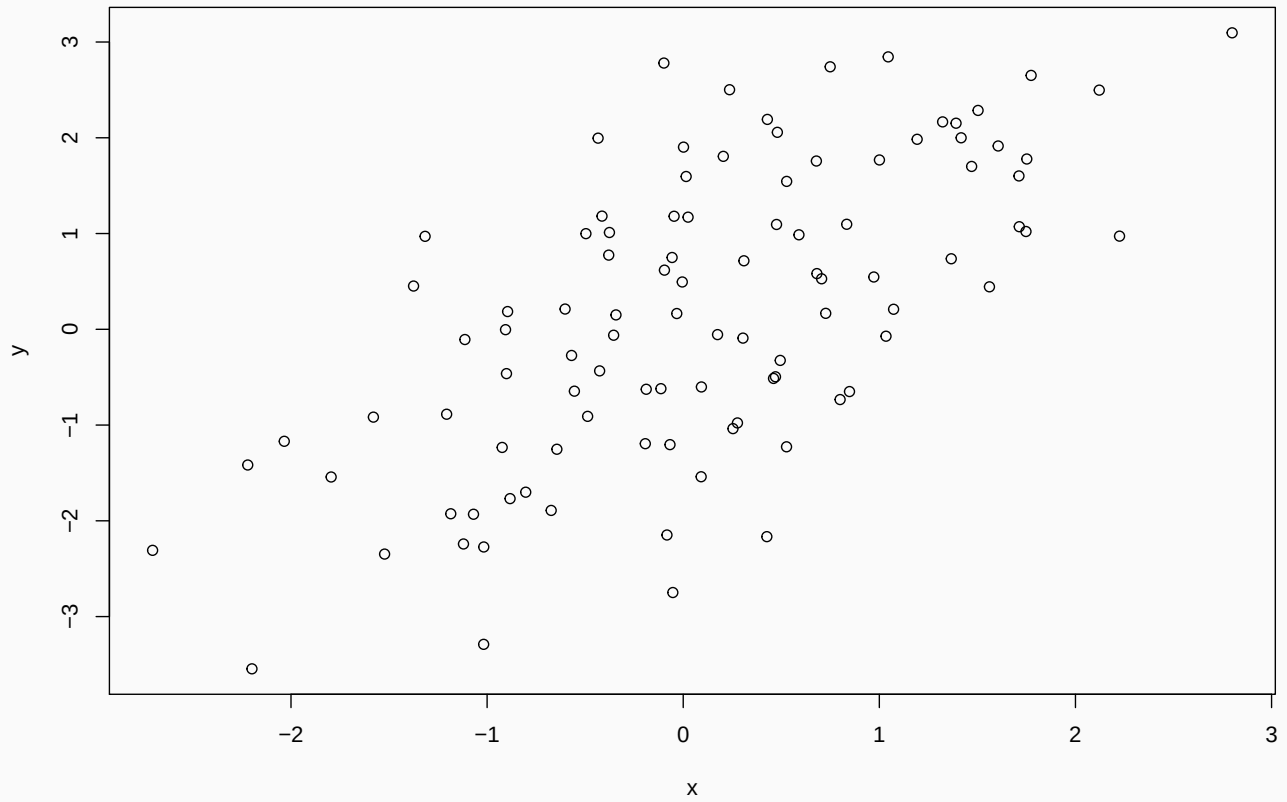
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

3. Ask R:

```
cor.test(x,y)
```

Statistical significance of the correlation coefficient: an example

```
x <- rnorm(100)
y <- x + rnorm(100)
plot(x,y)
```



Statistical significance of the correlation coefficient: an example

```
cor(x, y)
```

```
## [1] 0.6539086
```

```
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 8.5562, df = 98, p-value = 1.627e-13
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.5249097 0.7535422
## sample estimates:
```

Statistical significance of the correlation coefficient: an example

```
## [1] 0.6539086

##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 8.5562, df = 98, p-value = 1.627e-13
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.5249097 0.7535422
## sample estimates:
##       cor
## 0.6539086
```

The p-value is very low, therefore we can reject the null hypothesis

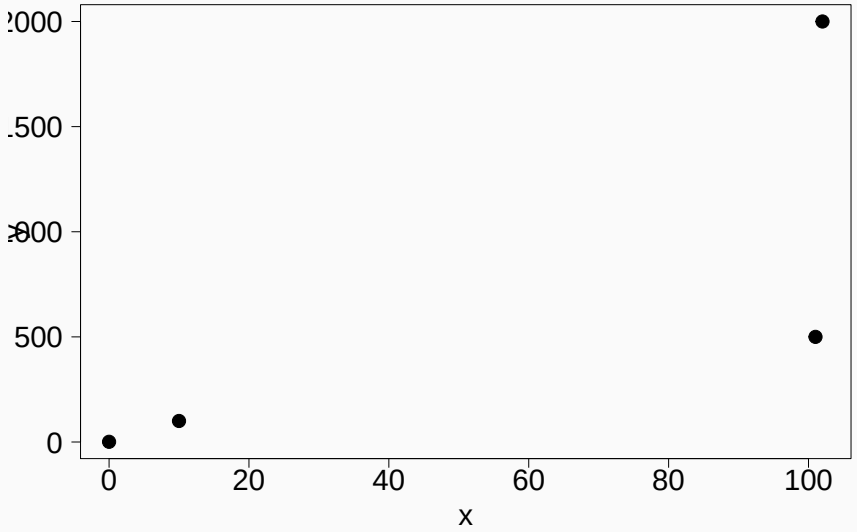
Correlation: Spearman's rank correlation coefficient

Correlation: Spearman's rank correlation coefficient

Sometimes we have non-linear data, or ordinal data, for which Pearson's r is not well suited. An alternative is Spearman's rank correlation coefficient, which is the same as Pearson' r but uses the correlation of the *ranks* rather than the correlation of the values. I.e.,

$$r_s = r_{rank_x, rank_y}$$

Correlation: Pearson vs Spearman: an example



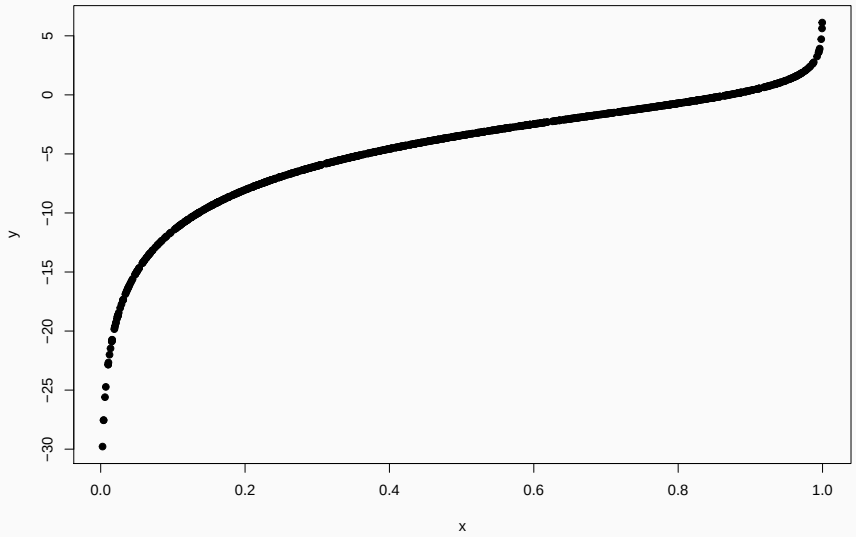

```
cor(x, y, method = 'pearson')
```

```
## [1] 0.7544237
```

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

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

Correlation: Pearson vs Spearman: another example




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

```
## [1] 0.9077051
```

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

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