CORRELATION TESTS

UNIVERSITÄT
LEIPZIG

Erik Kusch

erik.kusch@i-solution.de

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

Introduction

These approaches are extremely useful in data exploration and for preliminary analyses!

Prominent correlation tests include:

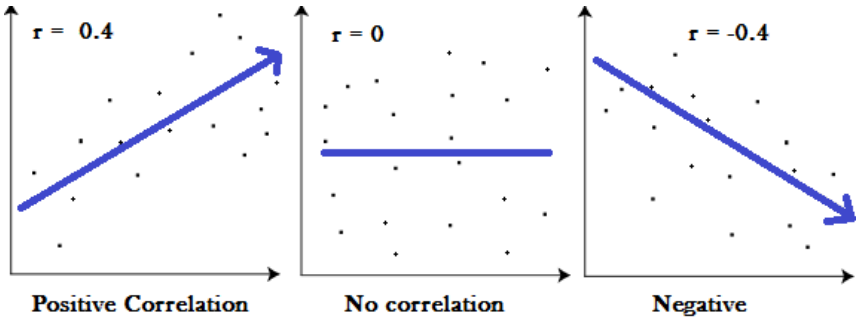
- **Contingency Coefficient**
- **Kendall's Tau**
- **Spearman Correlation**
- **Pearson Correlation**
- Cramer's V
- ...

Terminology

Correlation is **not** necessarily **causation**.

Correlation tests yield two measurements:

- *r* value (measure of correlation)
  - $r \approx 1$ (strong, positive correlation)
  - $r \approx 0$ (no correlation)
  - $r \approx -1$ (strong, negative correlation)
- *p* value (measure of statistical significance)



→ Get a feeling for it here http://guessthecorrelation.com/

Purpose And Assumptions

Contingency Coefficent

`ContCoef()` in the `DescTools` package

Purpose: To test whether variables are associated.

Assumptions:

- Variables must be categorical.

This test does not yield a significance assessment and only makes a statement of whether variables are associated ($|c| \approx 1$) or not ($|c| \approx 0$).

Minimal Working Example

Let's see if the `ContCoef()` function will identify the association between two categorical variables whose records are identical.

```
Samples <- c("Yes", "No")
set.seed(42)
counts <- sample(Samples, size = 1000, replace = TRUE)
table(counts, counts)
```

```
##        counts
## counts  No Yes
##    No  501   0
##    Yes   0 499
```

```
ContCoef(table(counts, counts))
```

```
## [1] 0.71
```

The association has been identified.

Purpose And Assumptions

Kendall's Tau

`cor.test()` in base `R` using the `method = "kendall"` specification

Purpose:	To test whether the ranks of values of two variables are correlated.
H_0	*Ranks of variable values are not correlated.*
Assumptions:	■ Variable value ranks are recorded as 'numeric'. ■ Variable values are ordinal, interval or ratio scaled.

Minimal Working Example

Let's have a look at what happens when we supply the `cor.test(..., method = "kendall")` with two correlated data sets:

```
cor.test(x = 1:1000, y = 1:1000, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  1:1000 and 1:1000
## z = 47, p-value <2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
## tau
##   1
```

As expected, a strong correlation can be found.

Purpose And Assumptions

Spearman Correlation

`cor.test()` in base `R` using the `method = "spearman"` specification

Purpose: To test whether the values of two variables are correlated in a **non-parametric** way.

H_0 *Values of variables are not correlated.*

Assumptions:

- Variable values are recorded as 'numeric'.
- Variable values are ordinal, interval or ratio scaled.
- Pairs of variable values are monotonically related.

Minimal Working Example

Let's have a look at what happens when we supply the `cor.test(..., method = "spearman")` with two non-correlated data sets:

```
set.seed(42)
cor.test(x = c(1:1000), y = sample(c(1:1000), size = 1000, replace = FALSE),
         method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  c(1:1000) and sample(c(1:1000), size = 1000, replace = FALSE)
## S = 2e+08, p-value = 0.7
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##  rho
## 0.01
```

As expected, no correlation can be found.

Purpose And Assumptions

Pearson Correlation

`cor.test()` in base `R` (`method = "pearson"` is the default)

Purpose: To test whether the values of two variables are correlated in a **parametric** way.

H_0 *Values of variables are not correlated.*

Assumptions:

- Values of each variable follow a **normal distribution**.
- Variable values are recorded as ‘numeric‘.
- Variable values are ordinal, interval or ratio scaled.
- Pairs of variable values are monotonically related.

Minimal Working Example

Let's have a look at what happens when we supply the `cor.test()` with two identical (hence correlated) data sets:

```
set.seed(42)
Data <- rnorm(n = 1000, mean = 500, sd = 50)
cor.test(x = Data, y = Data)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Data and Data
## t = Inf, df = 998, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  1 1
## sample estimates:
## cor
##   1
```

As expected, a strong correlation can be found.

Variables We Can Use

Which variables in our *Passer domesticus* data set are usable?

In general, correlation analyses can make use of all kinds of data as long as they are recorded as `numeric` and we are able to convert categorical values into `numerical` records if need be.

What about Pearson correlation and normal distributed data?
We need to check whether the assumption of normal distributed data records holds true for each variable before applying Pearson correlation. If that is not the case, we may wish to use Spearman correlation.

When dealing with climate types, remember to reduce confounding factors by only using data belonging to the stations `SI`, `UK`, `RE`, and `AU`.

What If There Are No Ranks?

We *order* our data!

```
set.seed(42)
Data <- sort(round(rnorm(n=10, mean=50, sd=2), 0))
Data # the data does not need to be sorted for ranking!
```

```
##  [1] 49 50 50 50 51 51 51 53 53 54
```

The `ties.method` argument in the `ranks()` function controls what to do when multiple values are the same.

```
rank(Data, ties.method = "average")
```

```
##  [1]  1.0  3.0  3.0  3.0  6.0  6.0  6.0  8.5  8.5 10.0
```

```
rank(Data, ties.method = "min")
```

```
##  [1]  1  2  2  2  5  5  5  8  8 10
```

Research Questions And Hypotheses

So which of our major research questions (seminar 6) can we answer?

Contingency Coefficient

- *Predation:* Are colour/nesting site and predators associated?
- *Sexual Dimorphism:* Are climate records and sex ratios associated?

Kendall's Tau

- *Climate Warming/Extremes:* Do heavier sparrows have heavier/less eggs? You need to rank female sparrow weight and egg weight for this.

Spearman

- *Climate Warming/Extremes:* Do sparrow weight/height and wing chord correlate with latitude?
- *Climate Warming/Extremes:* Do egg weight/number of eggs correlate with latitude?

Pearson

- *Fitness Constraints:* Dies psarrow weight and height correlate? Can you even run this analysis?

Use the `1 - Sparrow_Data_READY.rds` data set for these analyses.