# Biometry. Lecture 18

Alexey Shipunov

Minot State University

March 23, 2012

# Outline

1. Questions and answers
2. Two-dimensional statistics
    - Concordance and Cohen kappa
    - The anatomy of two-sample test: sign test
    - Correlation

# Outline

1. Questions and answers
2. Two-dimensional statistics
    - Concordance and Cohen kappa
    - The anatomy of two-sample test: sign test
    - Correlation

# Starting...

```
> setwd("<working folder>")
```

or

“`Change dir`”

in menu!

# Previous final question: the answer

What is a null hypothesis for the chi-squared test?

# Previous final question: the answer

What is a null hypothesis for the chi-squared test?

- Variables are distributed independently (no association)

# Two-dimensional statistics

## Concordance and Cohen kappa

- Concordance is a measure of “agreement” between two expert answer sheets
- The most common application are psychological tests
- Cohen kappa test is frequently used for understanding the degree of concordance; the null hypothesis for Cohen kappa is that two answer sheets are non-concordant

# Cohen kappa and island flora

```
> install.packages("concord")
> library(concord)
> isl <- read.table(
+ "http://ashipunov.info/data/pokorm_03.dat",
+ h=TRUE, sep=";")
> str(isl); head(isl)
> cohen.kappa(as.matrix(isl))
```

# Two-dimensional statistics

## The anatomy of two-sample test: sign test

# Sign test

- Idea is simple: to calculate differences between all pairs of values (paired test!)
- Then take only positive differences
- If two samples came from a same distribution, approximately 50% of differences should be positive—we can test with with, e.g., binomial test

# Making the sign test

We will take the same exam data we processed on lecture 16.

```
> first <- c(63, 72, 77, 76, 67, 56, 55, 51, 77, 64)
> second <- c(87, 86, 76, 79, 54, 60, 97, 80, 73, 97)
> dif <- second - first
> pos.dif <- dif[dif > 0]
> binom.test(length(pos.dif), length(dif))
```

# Two-dimensional statistics

## Correlation

# Covariance and correlation

- It is always interesting to know, **how much** are two random variables change together. Covariance show that but it is not easy to interpret.
- **Correlation coefficient** is a normalized version of covariance and therefore widely used as a measure of correlation. If correlation is close to 1 or $-1$, it is high.
- Therefore, correlation coefficient will show the strength of relation

# Features of correlation coefficient

- Correlation is a measure of **linear** relation. If relation is non-linear, correlation could be small or even zero. To check the linearity, it is recommended to make a `plot()` of two variables (scatterplot).
- Correlation may be positive or negative (from $-1$ to 1). If you need a sign-less measure, you may use determination coefficient $=$ correlation coefficient$^2$
- Correlation will only show that relation exists and has some strength, it will not show any other details about relation. For example, if correlation between A and B is high, it could mean that:
  - A depends on B
  - B depends on A
  - A and B depends on each other
  - A and B both independently depend on C and have nothing in common

# Calculation of correlation coefficient

```
> cor(5:15, 7:17)
> cor(5:15, c(7:16, 23))
> cor(5:15, c(7:16, 2))
> cor(5:15, 17:7)
> cor(trees)
```

`cor()` function works with vectors or tables (matrices and data frames). If `NA`s are present, one may use option `use="complete.obs"` (better) or `use="pairwise.complete.obs"`

# Non-parametric correlation

By default, cor() calculates parametric Pearson's correlation coefficient, it is possible to specify non-parametric (Spearman or Kendall) coefficients.

```
> cor(first, second, method="spearman")
```

# Visualization of correlation

```
> cor(longley)
> symnum(cor(longley))
> install.packages("ellipse")
> library(ellipse)
> plotcorr(cor(longley), type="lower")
```

# Correlation tests

- The null hypotheses for these tests is that correlation differs from zero
- There are both parametric and non-parametric tests

# Correlation tests

```
> with(trees, cor.test(Girth, Height))
> cor.test(first, second, method="spearman")
```

# Finishing...

```
> savehistory("20120323.r")
```

# Final question (10 points)

# Final question (10 points)

In the embedded data `USArrests`, there are numbers of murders and rapes per 100,000 for every state.

Are murders and rapes correlated? Is this correlation significant?

# Summary: most important commands

- `cor()`—calculates correlation coefficients
- `cor.test()`—run correlation tests

# For Further Reading

A. Shipunov.
*Biometry* [Electronic resource].
2012—onwards.
Mode of access: `http://ashipunov.info/shipunov/school/biol_299`

P. Dalgaard
*Introductory Statistics with R*. 2nd edition.
Springer, 2008.
*Chapter 6*.