# Biometry. Lecture 20

Alexey Shipunov

Minot State University

April 20, 2015

# Outline

1. Two-dimensional statistics
   - Analysis of covariation
   - Exact and approximate tests

# Starting...

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

or

“`Change dir`”

in menu!

On Mac, be sure that startup option is working: `getwd()` (`getwd()` checks if R is in working folder, `dir()` checks the folder content)

# Two-dimensional statistics

## Analysis of covariation

# Analysis of covariation (ANCOVA)

- ANCOVA integrates several regression lines together and checks the full model
- Model formula is
  `response ~ influence * factor`
- The ANCOVA will check if there is any difference between intersection and slope of the first line and intersections and slopes of all other lines (each line corresponds with one factor level)

# Grazing data

- 40 plants were treated in two groups: with grazing (in first two weeks) and without grazing
- Rootstock diameter was also measured
- At the end of season, fruit production was measured (dry weight in mg)

# Visualization first

```
> ipo <- read.table(
+ "http://ashipunov.info/data/ipomopsis.txt", h=T)
> head(ipo)
> with(ipo, plot(Root, Fruit,
+ pch=as.numeric(Grazing)))
> abline(lm(Fruit ~ Root, data=subset(ipo,
+ Grazing=="Grazed")))
> abline(lm(Fruit ~ Root, data=subset(ipo,
+ Grazing=="Ungrazed")), lty=2)
> legend("topleft", lty=1:2,
+ legend=c("Grazed","Ungrazed"))
```

# Model output

```
> ipo.lm <- lm(Fruit ~ Root * Grazing, data=ipo)
> summary(ipo.lm)
```

Two equations:

`Fruit = -125.174 + 23.24 * Root` (for grazed)

`Fruit = (-125.174 + 30.806) + (23.24 + 0.756) * Root` (for ungrazed)

# ANCOVA model tuning

```
> ipo.lm2 <- update(ipo.lm, . ~ . - Root:Grazing)
> summary(ipo.lm2)
> ipo.lm3 <- lm(Fruit ~ Root + Grazing, data=ipo)
> summary(ipo.lm3) # same as ipo.lm2: additive model
> AIC(ipo.lm)
> AIC(ipo.lm2)
> Let us return to women data
> AIC(women.lm1)
> AIC(women.lm2) # better!
```

AIC stands for “Akaike Information Criterion”

“ `. - something`” means “everything in the model is the same except something which has been taken out”

# Analysis of covariation, example II

Islands of two types: islet-like and stone-like

```
> it <- read.table("http://ashipunov.info/data/it.txt",
+ h=T, sep="\t")
> str(it)
> it$SQ <- log10(it$SQ)
> plot(SP ~ SQ, data=it, type="n")
> text(it$SQ, it$SP, labels=abbreviate(it$TYPE, 1))
> abline(lm(SP ~ SQ, data=subset(it, TYPE=="islet-like")))
> abline(lm(SP ~ SQ, data=subset(it, TYPE=="stone-like")),
+ lty=2)
```

# Analysis of covariation, example II

```
> it.ancova <- lm(SP ~ SQ * TYPE, data=it)
> summary(it.ancova)
> it.ancova2 <- update(it.ancova, ~ . - SQ:TYPE)
> summary(it.ancova2)
> AIC(it.ancova)
> AIC(it.ancova2) # better!
> summary(lm(SP ~ SQ + TYPE, data=it)) # like second
```

Interceptions are different but slopes are the same. In statistical language, we may say that in this case, additive model is better. Square and type are two independent terms.

# Two-dimensional statistics

## Exact and approximate tests

# Chi-squared and Fisher exact

- Chi-squared proportion tests will **estimate** the p-value from theoretical distribution. As a consequence, it may say "*Chi-squared approximation may be incorrect*".
- Fisher exact and binomial tests will **calculate** p-value directly. That is why they are sometimes preferable.
- Somehow similar difference exists between t-test and Wilcoxon test. The later sometimes says "*Cannot compute exact p-values with ties*". Pearson (default) and Spearman correlations are also different this way.

# Fisher's tea drinker

A British woman claimed to be able to distinguish whether milk or tea was added to the cup first. To test, she was given 8 cups of tea, in four of which milk was added first.

```
> tea <- matrix(c(3,1,1,3), nrow=2)
> colnames(tea) <- row.names(tea) <- c("Milk", "Tea")
> tea
> chisq.test(tea) # warning!
> fisher.test(tea)
```

# How to avoid the approximation with simulation

```
> eq <- read.table("http://ashipunov.info/data/eq.txt", h=T)
> table(eq$N.REB, eq$N.ZUB) # less than 5 in cells
> chisq.test(eq$N.REB, eq$N.ZUB)
> chisq.test(tea, simulate.p.value=T) # no warning!
> chisq.test(eq$N.REB, eq$N.ZUB, simulate.p.value=T)
```

When some cells contain less than 5 items, `simulate.p.value=T` is recommended.

# Finishing...

Save your commands!

(`savehistory(<todaysdate>.r)` or `File -> Save as...` on Mac)

# Summary: most important commands

- `lm()`—estimates the linear regression model and many other models (like ANCOVA)
- `predict()`—predict values with model
- **Exact tests** will not need approximations but they are normally less powerful than “regular” tests

# For Further Reading

A. Shipunov.
*Biometry* [Electronic resource].
2012—onwards.
Mode of access:
`http://ashipunov.info/shipunov/school/biol_240`

A. Shipunov, and others.
*Visual statistics. Use R!*
Ongoing translation from Russian.