# Biometry. Lecture 1

Alexey Shipunov

Minot State University

January 13, 2016

# Outline



# Outline



# Outline



# Outline



# Course in general

## Description

# Course description

Course will cover introductory statistic concepts in a form designed specifically for biology majors, its goal is to strengthen Biology and Chemistry students statistical knowledge and abilities. It is a practical, software-based examination of the concepts of sampling, hypotheses testing (non-parametric and parametric), descriptive statistics, contingency, correlation, analysis of variation, linear models and basic multivariate techniques. Only biological, real-world data will be used. Course will concentrate on underlying principles, applicability and practical use of methods covered. R statistical environment will be used as a main software tool.
The course relies on the computer literacy: file system and basic file operations, basic text operations, spreadsheets, vector and raster graphics, Internet file formats and protocols.

# Main concepts

- What is data and how to process it
- What are statistical hypotheses and how to prove them
- How to get answers from one-, two- and multidimensional data

# What should be your skills by May: Exam 4

1. Open R, download the data file from Internet (address is `http://ashipunov.info/data/[illegible].txt`), load it into the R object.
2. Explore the data frame, **check normality** for every measurement character (5 points).
3. Answer the following questions (do not forget to supply numerical arguments):
    1) Do these "species" grow on the different distances from sea? (15 points)
    2) Does the association exist between species and substrate type? (15 points)
    3) Which pair of **morphological measurement characters** are most correlated? Is this correlation significant? (15 points)
    4) Make the linear model for these two most correlated characters. How good is the model, is it significant? (15 points)
    5) Make the logistic regression of being "[illegible]", taking into account the distance from sea. How reliable is that model? (20 points)
    6) There are three types of substrate. Does the length of leaf depend on the substrate? (15 points)
4. You may want to supply graphs. Every reliable graph = 5 extra points.

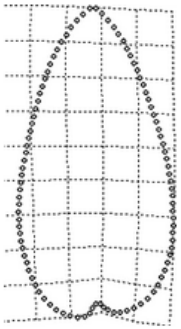
# Instructor

- Dr. Alexey Shipunov
- Office: Moore 229
- Office Hours: Mondays, Wednesdays, 1 p.m. to 3 p.m.
- Phone: 858-3116
- E-mail: `alexey.shipunov@minotstateu.edu` — this is the preferrable way of communication.

# Know your Syllabus!



## BIOL 240: Biometry



Course materials:

- Syllabus (PDF, 0.15 Mb)
- Points and grades (Excel, 0.01 Mb)

---

- ~~Lecture 1 (PDF, 0.3 Mb), R script for Lecture 1 (Text)~~

---

- Old lectures (2012)
- Old lectures (2014)
- Old lectures and scripts (2015)

---

- Data files
- R reference card (PDF, 0.1 Mb)
- Shipunov, A., and many others. Visual statistics. Use R! (PDF, 1 Mb)

Back

**http://ashipunov.info/shipunov/school/biol_240/**

# Computer literacy

## Computer knowledge and skills needed

# Checklist of the necessary computer skills

- File system and basic file operations, working with file manager:
  use only lowercase letters, numbers and underscore (dot for extension), learn how to use ZIP folders
- Understanding of the simple and formatting text:
  use Notepad, Text or other simple text editors; be aware of different line endings on Mac, Windows and Unix/Linux; be aware of invisible symbols including tabulation
- basic text operations (copy/paste etc.)
- Spreadsheets:
  know basic operations, use LibreOffice Calc instead of Excel if you like
- Vector and raster graphics:
  will be explained due course
- Internet file formats and protocols:
  HTML, PDF, `http://`, `ftp://`, `mailto:`

# Statistics
## What is statistics

# Definition of Statistics

**Data collection** Collecting any numerical data, e.g. unemployment rate per state.

**Sampling** Working with any subsets (samples) of data, like voting polls.

**Data analysis** Procedures used to analyze data, such as ANOVA or chi-square statistic.

**Research** Science that develops mathematical procedures to describe data.

In all, statistics is about data.

# Statistics
## Data

# Small data

- Small data is often self-explanatory.
- Experiments with cognition show that it is easy to operate with 5-9 objects in mind.
- Visual inspection gives an average value close to 2.

```
2 3 4 2 1 2 2 0
```

# Uniform data

2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

- Visual inspection again gives an average value close to 2.
- Uniform data could be (relatively) big, but understandable without special tools.

# Real data

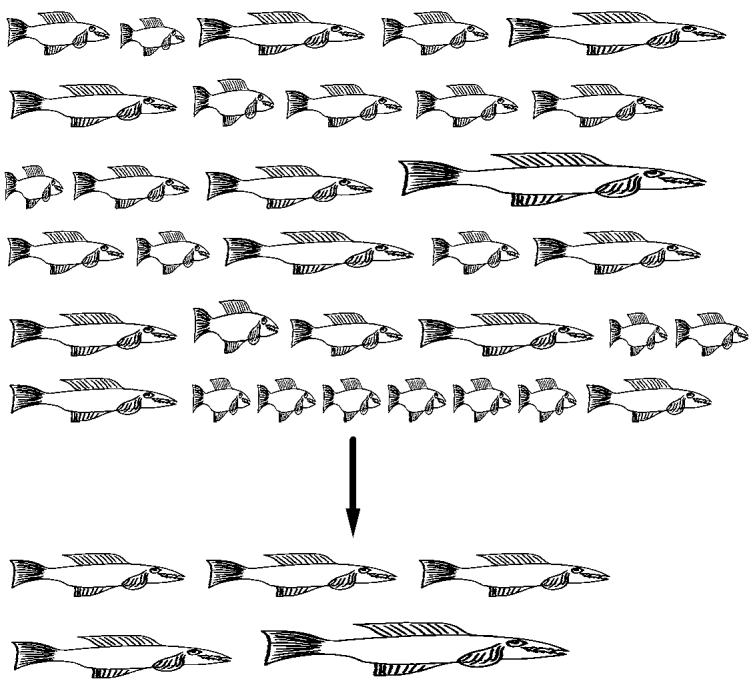
**Data from Shipunov et al., 2012**

```
88 22 52 31 51 63 32 57 68 27 15 20 26 3 33 7 35 17
28 32 8 19 60 18 30 104 0 72 51 66 22 44 75 87 95 65
77 34 47 108 9 105 24 29 31 65 12 82
```

- However, in most cases biological data is much more complicated.
- Therefore, we will need specific (statistical) tools even for preliminary description of data.

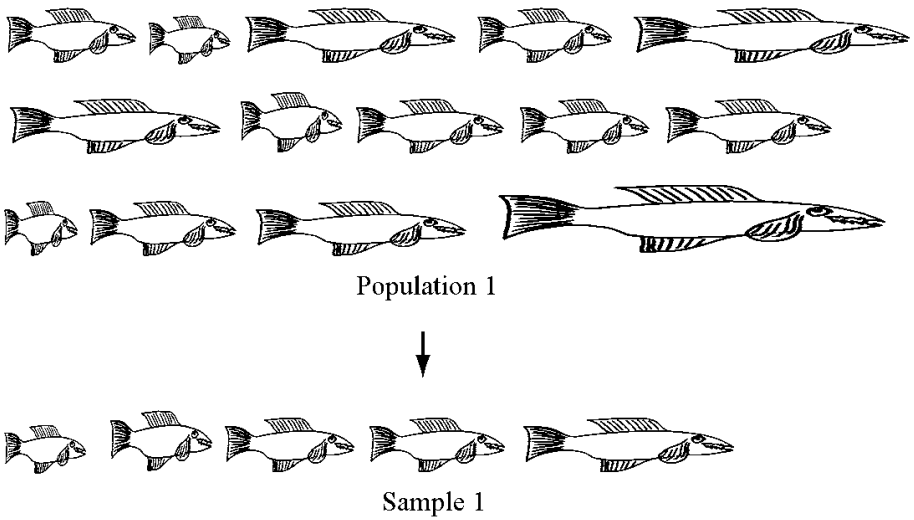
# Statistics
## Samples

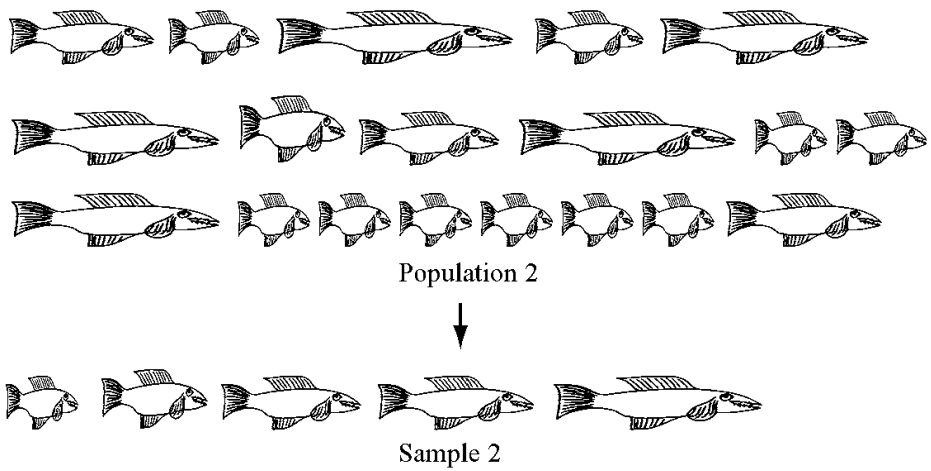
# Sampling



- Biologists often work with large numbers of objects and therefore need to sample (subset) initial population.
- Sampling gives you free hands, it is robust from errors and it is cheaper than full research. Moreover, philosophically, any research is based on sampling.
- However, the sample may not necessary be a good representative of a population. Only statistical tools will help to determine the reliability of the sample.

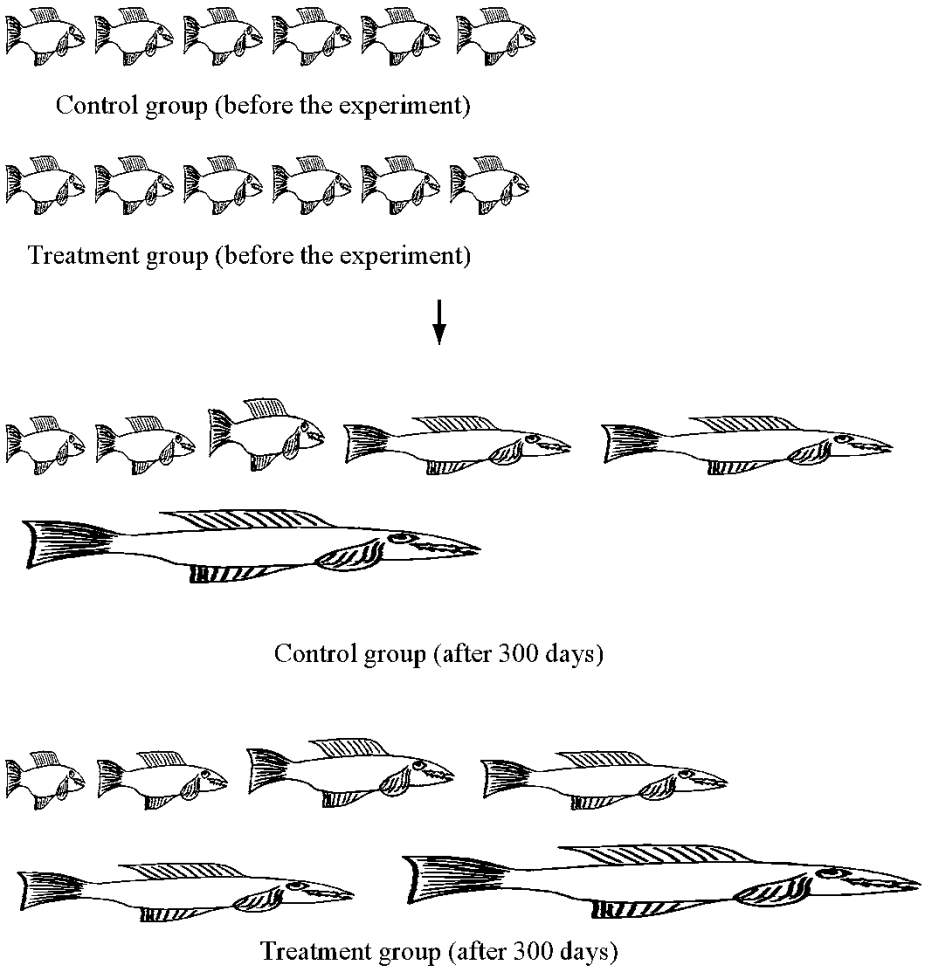
# Typical problem of sampling




- Even samples chosen at random from two different populations may not necessary be different.
- Whereas experiment requires simpler statistical tools, observation frequently needs things like data mining.

# Experiments vs. observation




- Experiment requires controlled conditions whereas observation minimizes the influence.
- Again, only careful examination of samples with appropriate tools will make results of experiment robust.

# R
# Non-R software

# Calculators

- Calculator is almost always embedded into OS
- Too elaborative if we use samples

# Spreadsheets

- MS Excel, OpenOffice.org/LibreOffice Calc, Gnumeric
- Very handy for data input and visualization
- Do not contain advanced and optimized statistical methods
- Are not able to conduct complex calculations

# Graphical statistical software

- SPSS, MiniTab and many others
- Have a high diversity of different graphs and plots
- Will fail if you need to repeat the complex procedures with different datasets

# Statistical environments

- SAS, S-Plus and R
- Full control: it is possible to implement *every* statistical method
- User should remember commands

# R
# Starting with R

# R history

- Started in 1993 as non-commercial analog of S-Plus
- R is just another implementation of S statistical language developed in AT&T
- In last five years, became a standard for statistical research
- Has more than 7,700 extension packages

# R pros and cons

- Extremely flexible, open source
- No GUI: which command?

# Final question (2 points)

# Final question (2 points)

What is sampling?

*Together with name and answer, supply your 4-digit class ID*

# Summary

Statistics is:

- Gathering data
- Making samples
- Applying tools
- Develop new ways of things above

# For Further Reading

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

A. Shipunov, and many others.
*Visual statistics. Use R!*
2015—onwards.
Mode of access: http://ashipunov.info/shipunov/school/biol_240/en/visual_statistics.pdf