Biometry. Lecture 1

Alexey Shipunov

Minot State University

January 14, 2015

Outline

1. Course in general
    - Description
    - Grading
2. Computer literacy
    - Computer knowledge and skills needed
3. Statistics
    - What is statistics
    - Data
    - Samples

Outline

1. Course in general
    - Description
    - Grading
2. Computer literacy
    - Computer knowledge and skills needed
3. Statistics
    - What is statistics
    - Data
    - Samples

Outline

1. Course in general
    - Description
    - Grading
2. Computer literacy
    - Computer knowledge and skills needed
3. Statistics
    - What is statistics
    - Data
    - Samples

Course in general

Description

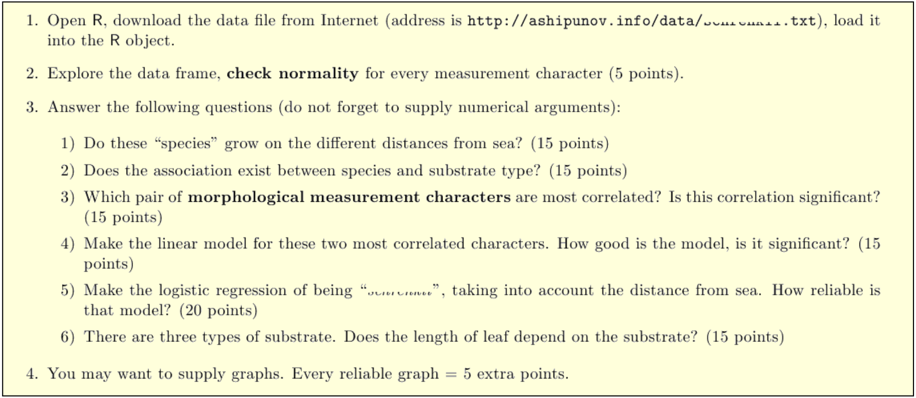
Course description

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 will be your skills by May: Exam 4



1. Open R, download the data file from Internet (address is `http://ashipunov.info/data/`[redacted]`.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 "[redacted]", 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 and Fridays, 10 a.m. to 11:00 a.m.
- Phone: 858-3116
- E-mail: `alexey.shipunov@minotstateu.edu` — this is the preferrable way of communication.

Details

А. Б. Шипунов, Е. М. Балдин, П. А. Волкова,
А. И. Коробейников, С. А. Назарова,
С. В. Петров, В. Г. Суфиянов

Наглядная статистика

Используем R!

Lectures	Mondays and Wednesdays, 11:35 a.m. to 12:50 a.m., Moore 213
Laboratories	Thursdays, 9 a.m. to 12 p.m., Moore 213
Textbook	Shipunov A., and others. Visual statistics. Use R!. DMK Press, 2012. [In Russian]. English translation of the book will be partly available on the course Web site.

Course in general

Grading

Exams

- Four **equal** exams are given during the semester.
- Only the **three** best exams contribute to the final grade.
- Missed exams count zero points. There are **no make-up** exams.

Labs

- Receiving zero points for **more than one** laboratory results in a failed course.
- Grading of laboratories is based on reports.
- Written reports are prepared and finished during laboratory sessions and sent via e-mail or passed to the instructor right after the particular laboratory session.
- It is expected that you have reviewed the lecture contents before you come to lab.

Absence

There are five legitimate reasons for absence from labs:

1. emergency situations,
2. attested medical conditions,
3. military duty,
4. participation in MSU sports events,
5. dependent sick leave.

Absence from laboratories ust be announced to me via e-mail in advance. I strongly recommend attending lectures regularly. Statistically, students who achieved best grades are **always attend lectures**.

Lectures

- Every lecture will have a computer-based, practical part.
- On the lecture, I may give short test question(s) to answer.

Points

Points are distributed as follows:

- Three best exams: $\leq$ 300 points (assuming 100 points per exam)
- Lecture tests: $\leq$ 30 points (1–3 points per question)
- Laboratories: $\leq$ 240 points (20 points per lab $\times$ 12 labs)

Grading points may vary between exams, tests, and labs.

Grades

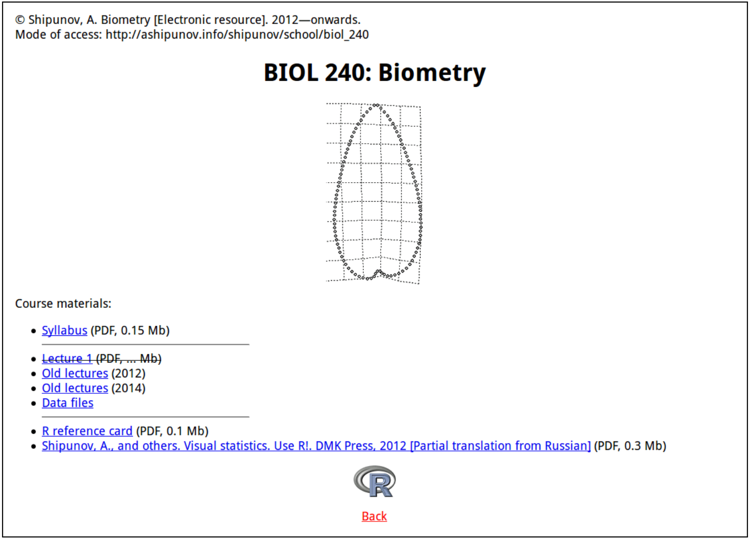
- $A \geq 90\%$
- $B \geq 80\%$
- $C \geq 70\%$
- $D \geq 60\%$
- $F < 60\%$

A minimum of one letter grade will be deducted from the grade for academic dishonesty / plagiarism.

Tentative course sequence

- Data and data processing
- How to process data: R basics
- Types of data
- One-dimensional data, descriptive statistics
- Two-dimensional data, contingency tables
- Correlation
- Regression
- ANOVA
- Multidimensional data, data mining

Course Web site



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

BIOL 240: Biometry

Course materials:

- Syllabus (PDF, 0.15 Mb)

- ~~Lecture 1 (PDF, ... Mb)~~
- Old lectures (2012)
- Old lectures (2014)
- Data files

- R reference card (PDF, 0.1 Mb)
- Shipunov, A., and others. Visual statistics. Use R!. DMK Press, 2012 [Partial translation from Russian] (PDF, 0.3 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 Gnumeric 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 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 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

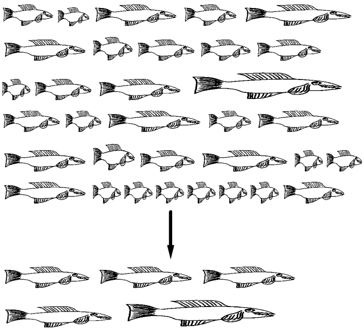
```
88 22 52 31 51 63 32 57 68 27 15 20 26 3 33 7 35 17
28 32 8 19 60 18 30 104 0 72 51 66 22 44 75 87 95 65
77 34 47 108 9 105 24 29 31 65 12 82
```

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

Statistics

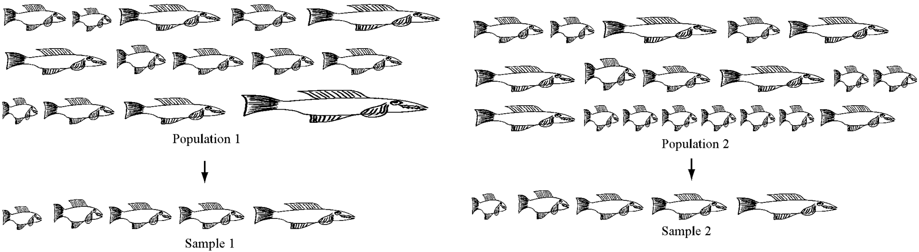
Samples

Sampling



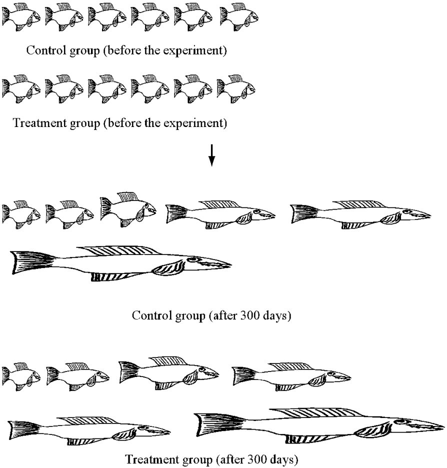
- Biologists often work with large numbers of objects and therefore need to sample (subset) initial population.
- However, the sample may not necessary be a good representative of a population.
- Only statistical tools will help to determine the reliability of the sample.

Comparing two populations



- Even samples chosen at random from two different populations may not necessary be different.
- Only statistics will help to recognize “true” difference from “false” difference.

Experiments


Control group (before the experiment)
Treatment group (before the experiment)
Control group (after 300 days)
Treatment group (after 300 days)

- Biologists often conduct experiments. However, natural variation among individuals within a sample may obscure any effect of an experimental treatment.
- Again, only careful examination of samples with appropriate tools will make results of experiment robust.

Final question (2 points)

Final question (2 points)

What is sampling?

Together with name and answer, supply your 6-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 others.
Visual statistics. Use R!
DMK Press, 2012. Translated from Russian.