Biometry. Lecture 4

Alexey Shipunov

Minot State University

January 20, 2012

Outline

Outline

Outline

Minot State UNIVERSITY

Previous final question: the answer

What is a most important difference between observation and experiment?

Previous final question: the answer

What is a most important difference between observation and experiment?

- Minimal influence *vs*. planned influence

R break: generate 30 random numbers

```
> sample(1:100, 30, replace=TRUE)
```

R break: rolling dice

```
> dice <- as.vector(outer(1:6, 1:6, paste))
> sample(dice, 2, replace=TRUE)
```

R break: new commands

```
as.vector()
outer()
paste()
sample(..., replace=TRUE)
TRUE (T) and FALSE (F) keywords
```

Data
Principles of sampling

Randomization

- Randomization: every object should have equal chances to be in the sample
- One of the best ways: introduce order which is knowingly absent in nature

Weevil experiment: comparing poisons

- Poison on filter paper
- Take the first weevil from a jar; put it on paper; count time; change chemical
- Why the most effective chemical is always first?

What to search in data

- Generalities
- Comparisons
- Associations: correspondences, correlations and relations
- Structure

Generalities

- Center
- Range
- Descriptive and inferential methods

R break: inferential and descriptive methods

```
# Internal data: precipitation in main US cities
> precip
# Average precipitation
> mean(precip)
# Using Student's (t) test for confidence interval
> t.test(precip)
```

Comparisons

- Are two samples equal (taken from one population)?
- Greater or less?
- Most of comparisons are double, multiple comparisons are statistically dangerous.

Correspondences

- Samples are somehow connected.
- Correspondence does not show neither the strength nor direction of connection.

Correlations

- Samples are somehow related.
- Correlation shows the strength but not the direction of relation.

Regressions

- One sample is a response on the other (response ~ factor model)
- Regression measures both the strength and direction of relation
- Complex regressions are often called **models**

Structure

- Dataset contains groups, subgroups etc. (internal structure)
- Descriptive and inferential

Processing of data

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 and MiniTab
- 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

Processing of data

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 3,500 extension packages

R pros and cons

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

Final question (2 points)

Final question (2 points)

What is a difference between inferential and descriptive methods?

Summary

- **Replication** and **randomization** are two basic principles of research
- Descriptive methods *show*, inferential methods *prove*

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.
Appendix A, Preface, Section 1.1.

Minot State UNIVERSITY