# Biostatistics

**Lead Instructor(s)**  
Pervouchine, Dmitri

---

## Is this syllabus complete, or do you plan to edit it again before sending it to the Education Office?

The syllabus is a final draft waiting for approval (once approved the syllabus will be published on the public web-site and other systems)

## Contact Person

Dmitri Pervouchine

## Contact Person’s E-mail

pervouchine@gmail.com

---

## 1. Annotation

### Course Description

This introductory course to statistics and probability theory is modeled as an extension of a traditional university Statistics course and Advanced Placement Course in Statistics to a broader spectrum of topics, while keeping the spirit of quantitative discourse applied to real-life problems. The material is offered in 5 consecutive modules (see Course Outline below), each containing a lecture, a discussion section, and a practicum. For practical exercises we use R programming language and R-Studio software. However, this course is focused on statistics rather than R; therefore, each practicum is designed with the purpose to demonstrate and reinforce understanding of concepts introduced in the lecture rather than to provide a training in R.

### Course Prerequisites / Recommendations

Formally speaking, there are no prerequisites. Nonetheless, most concepts of probability theory and combinatorics will be reintroduced only briefly. Therefore, Biostatistics is recommended for the audience who are already familiar with the basic concepts of probability.

---

## 2. Structure and Content
<table>
<thead>
<tr>
<th>Topic</th>
<th>Summary of Topic</th>
<th>Lectures (# of hours)</th>
<th>Seminars (# of hours)</th>
<th>Labs (# of hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploratory Data Analysis.</td>
<td>Exploratory data analysis: bar-plot, histogram, cumulative distribution function, box-plot, scatter-plot, pie charts etc. Population, sampling, measures of center and spread, percentiles, outliers. Experiments versus observational studies, confounding factors, simple random sample, other types of sampling, biases in sampling techniques. Introduction to R programming language and R Studio: Data types, variables, packages, functions, handling files/scripts/projects. Object-oriented programming in R statistics. Practicum: Basic plots in R using the ggplot2 package.</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Statistical Inference, part 1.</td>
<td>The concept of hypothesis testing, type I and type II error, false discovery rate. Significance and confidence level, p-value. Confidence intervals. One-sided and two-sided tests and confidence intervals. Sampling distribution, estimators, standard error. Normal probabilities in application to p-value. One-sample and two-sample tests for independent and matched samples with known variance. The case of unknown variance and Student t-distribution, assumption of normality. Pooled variance and equal variances assumption. Practicum: One- and two-sample tests with known and unknown variance, test for proportions, simulation involving confidence intervals and t-distribution.</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
3. Assignments

<table>
<thead>
<tr>
<th>Assignment Type</th>
<th>Assignment Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Set</td>
<td>Exploratory data analysis and experiment design</td>
</tr>
<tr>
<td>Problem Set</td>
<td>Probability theory</td>
</tr>
<tr>
<td>Problem Set</td>
<td>Hypothesis testing and confidence intervals</td>
</tr>
<tr>
<td>Problem Set</td>
<td>Non-parametric tests</td>
</tr>
<tr>
<td>Problem Set</td>
<td>Regression and ANOVA</td>
</tr>
<tr>
<td>Project</td>
<td>Final project: statistical report on a dataset</td>
</tr>
<tr>
<td>Test/Quiz</td>
<td>Midterm exam</td>
</tr>
<tr>
<td>Test/Quiz</td>
<td>Final Exam</td>
</tr>
</tbody>
</table>

4. Grading

<table>
<thead>
<tr>
<th>Type of Assessment</th>
<th>Graded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade Structure</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity Type</th>
<th>Activity weight, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework Assignments</td>
<td>15</td>
</tr>
<tr>
<td>Midterm Exam</td>
<td>35</td>
</tr>
<tr>
<td>Final Exam</td>
<td>40</td>
</tr>
<tr>
<td>Projects</td>
<td>5</td>
</tr>
<tr>
<td>Attendance</td>
<td>5</td>
</tr>
</tbody>
</table>

Grading Scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>61</td>
</tr>
<tr>
<td>C</td>
<td>41</td>
</tr>
<tr>
<td>D</td>
<td>32</td>
</tr>
<tr>
<td>E</td>
<td>31</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
</tr>
</tbody>
</table>

Attendance Requirements: Mandatory with Exceptions

5. Basic Information

Maximum Number of Students
Maximum Number of Students

<table>
<thead>
<tr>
<th></th>
<th>Overall:</th>
<th>Per Group (for seminars and labs):</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

Course Stream: Science, Technology and Engineering (STE)

Course Term (in context of Academic Year): Term 3

Course Delivery Frequency: Every year

Students of Which Programs do You Recommend to Consider this Course as an Elective?

<table>
<thead>
<tr>
<th>Masters Programs</th>
<th>PhD Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy Systems</td>
<td>Life Sciences</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>Materials Science and Engineering</td>
</tr>
<tr>
<td>Materials Science</td>
<td></td>
</tr>
<tr>
<td>Photonics and Quantum Materials</td>
<td></td>
</tr>
</tbody>
</table>

Course Tags: Math, Biotechnology, Arts

6. Textbooks and Internet Resources

<table>
<thead>
<tr>
<th>Web-resources (links)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://apcentral.collegeboard.com/">http://apcentral.collegeboard.com/</a></td>
<td>AP statistics</td>
</tr>
<tr>
<td><a href="https://www.statsoft.com/">https://www.statsoft.com/</a></td>
<td>Statistica</td>
</tr>
<tr>
<td><a href="http://www.statmethods.net/">http://www.statmethods.net/</a></td>
<td>Quick-R</td>
</tr>
<tr>
<td><a href="http://ggplot2.org/">http://ggplot2.org/</a></td>
<td>ggplot2 cookbook</td>
</tr>
</tbody>
</table>

7. Facilities

Software

R-studio, ggplot2

8. Learning Outcomes
### Knowledge

By the end of the course students will be able to read and understand statistical problems, carry out statistical tests and interpret their results, choose appropriate statistical tests for their particular problems, summarize and graphically represent the data, design and conduct statistical experiments. Special attention is paid to the interpretation of statistical tests and writing conclusions. The very basic skill that is an absolute must to get a passing grade for the course is the ability to compute normal probabilities, which will be tested in all exams.

### Skill

Students are expected to develop an understanding of basic statistical concepts such as sampling from populations, hypothesis testing, confidence intervals, point estimates, standard error, decomposition of variance etc. On the practical side among other skills students are expected to be able to solve real-life statistical problems and write coherent reports (in any language) regarding their analyses and conclusions. The latter objective is not limited to biology, medicine or any other subject area.

### Experience

Students are expected be able to read and understand statistical problems, carry out statistical tests and interpret their results, choose appropriate statistical tests for their particular problems, summarize and graphically represent the data, design and conduct statistical experiments.

### 9. Assessment Criteria

**Input or Upload Example(s) of Assignment 1:**

<table>
<thead>
<tr>
<th>Select Assignment 1 Type</th>
<th>Homework Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input Example(s) of Assignment 1 (preferable)</strong></td>
<td></td>
</tr>
<tr>
<td>1. Given a normal distribution with a mean of 25, what is the standard deviation if 18% of the values are above 29? Explain.</td>
<td></td>
</tr>
<tr>
<td>2. A company claims that its aspirin tablets will relieve headaches faster than any other aspirin on the market. To determine whether the claim is valid, random samples of size 15 are chosen from aspirins made by the company and by another drug producer. An aspirin is given to each of the 30 randomly selected persons suffering from headaches and the number of minutes required for each to recover from the headache is recorded. The sample means are 8.5 and 9.1, while the standard deviations are 4.2 and 4.9.</td>
<td></td>
</tr>
<tr>
<td>a. Explain which sources of bias are associated with this experiment, and how you would reduce them.</td>
<td></td>
</tr>
<tr>
<td>b. Test at the 5% significance level that the company’s aspirin cures headaches significantly faster than the other pill. Explain your solution</td>
<td></td>
</tr>
<tr>
<td>3. Design an experiment to test the efficacy of a new leather-protecting compound to be used as a shoe protector during winter season. Name your variables, confounding factors, how you would account for their impact, whether you use blinding etc.</td>
<td></td>
</tr>
</tbody>
</table>

**Assessment Criteria for Assignment 1**

1. Statistical reasoning and proper conclusion
2. Correct method choice
3. Numerically correct answer

**Input or Upload Example(s) of Assignment 2:**
Problem Set

1. The CO2 data frame contains the results of an experiment on the cold tolerance of the grass species Echinochloa crusgalli. Use ggplot2 to present the time series graphically. Do not carry out statistical tests.

2. Demonstrate by simulation that a 2-sample confidence interval for the difference of two means indeed has the desired level of significance. Consider the following context. Two pharmaceutical pipelines are supposed to produce a food supplement. The food supplement is supposed to contain certain amount of active compound A. In order to assess the amount of this compound, two samples of size 12 were drawn from these pipelines. Sample means were found to be 127.5 mg for the first pipeline and 120.1 mg for the second pipeline, respectively. Assume the standard deviation of 8.1mg for each pipeline.
   a. Construct and interpret a 95% confidence interval for the difference of means. Make sure you interpret the confidence level. State the assumptions that are required for your interval to be valid.
   b. Simulate N=1000 pairs of samples drawn from the appropriate distribution with the mean 125 mg and standard deviation 8.1mg, construct a 95% confidence interval for the difference of means for each sample and count how many times your confidence interval covers the hypothetical value of 0. c. What number of times would you expect your confidence interval to cover the actual value of difference? Perform an appropriate statistical test to support your opinion.

Assessment Criteria for Assignment 2

1. Statistical reasoning and proper conclusion
2. Implementation
3. Clarity of display items

Input or Upload Example(s) of Assignment 3:

Select Assignment 3 Type

Projects

Your assignment is: cabbages dataset from MASS library

The results of the project should include (a) 1-page report and (b) one-slide presentation for no longer that 5 minutes. The report *must* include graphical displays to represent the data, at least one statistical test carried out at the 5% significance level, and interpretation of the result in the context of the data. If the description of the data frame is not sufficient, you may want to look up the publications cited in the description page.

Assessment Criteria for Assignment 3

1. Problem statement and motivation
2. Correct method choice
3. Clarity of display items
4. Statistical reasoning and proper conclusion

Input or Upload Example(s) of Assignment 4:

Select Assignment 4 Type

Final Exam

MULTIPLE CHOICE PART

1. The language school teaches English at four different levels. Over past years, approximately 65% of students who were enrolled at the beginning of the year completed all four levels by the end of the year. In 2019, 124 students completed all four levels by the end of the year out of 169 students who were enrolled. Is there enough evidence to claim that the
proportion of enrolled students who completed all four levels by the end of the year has increased in 2019?

A. Yes, at the 1% significance level  
B. Yes, at the 2.5%, but not at the 1% significance level  
C. Yes, at the 5%, but not at the 2.5% significance level  
D. Yes, at the 10%, but not at the 5% significance level  
E. Not even at the 10% significance level  

2. A gender equality team conducts a study, in which it compares the salaries of men and women in a large biomedical company with 5000 male and 5000 female employees. How large a sample of n men and n women should be surveyed to estimate at the 95% confidence level the difference between average monthly salaries of men and women to within ±$50? Assume standard deviation of $460 for each group.

A. n = 560  
B. n = 504  
C. n = 651  
D. n = 916  
E. n = 1120  

3. Which of the following is NOT an important assumption in two-way ANOVA?
A. Independence of observations  
B. Homogeneity of variances  
C. Normal distribution of observations  
D. Balanced number of levels of the two factors  
E. Additive contribution of the two factors  

FREE RESPONSE PART

4. Caffeine (particularly in the form of coffee) is one of the most widely consumed stimulants in the world, with 90% of American adults consuming caffeine-infused beverages almost daily (O’Callaghan et al, Risk Manag Healthcare Policy 2018, 11:263–271). While there is clear evidence that caffeine enhances performance, it can have a disruptive effect on human sleep, and hence some people avoid taking coffee in the afternoon. A study of coffee drinking habits was carried out among male and female employees of a small IT company. The respondents entering the building were asked the question: "What is your favorite time to drink a cup of coffee?"

Breakfast After Lunch Dinner  
Women 15 9 19  
Men 21 2 10

(a) A statistician decided to use $\chi^2$ -test to assess the association between gender and coffee drinking habits. Explain why $\chi^2$ -test cannot be used here.  
(b) Carry out the statistical test that can be used instead (you may leave a symbolic representation for the P-value).  
(c) Another statistician suggested to merge ‘After Lunch’ and ‘Dinner’ into a single ‘Afternoon’ category. Can $\chi^2$ -test be used now? If so, carry out the test at the 5% significance level. If not, explain why.  
(d) Yet another statistician noticed that according to this data, men tend to drink coffee in the morning, while women tend to drink coffee in the afternoon. Do your results from part (c) support this observation and, if so, at which significance level?  
(e) Do you think that the results of this study can be generalized to the entire population of coffee drinkers? Explain why or why not.
The course is offered at two levels: MS and PhD. The distinction is in the weights.

MS: 15% HOMEWORKS + 35% MIDTERM + 40% FINAL + 5% PROJECT + 5% ATTENDANCE

PhD: 5% HOMEWORKS + 35% MIDTERM + 40% FINAL + 15% PROJECT + 5% ATTENDANCE

MS students are normally offered a data frame for the final project, but exceptionally they can bring their own data. PhD students normally bring their own data, but they can be exceptionally offered a data frame for the final project. MS students are graded on the basis of their final presentation only. PhD students are graded on the basis of both final presentation and report.