Government of the Russian Federation

Federal State Autonomous Educational Institution
for Higher Professional Education
«National Research University
«Higher School of Economics»

Sociology department

Discipline program
Sociological data analysis - 1

For course 040100.62 «Sociology», preparation for a Bachelor's degree.

Program developed by:
A.R. Bessudnov, D. Phil, abessudnov@hse.ru
A.V. Ryjova, anastasia.ryjova@yandex.ru

Approved at the meeting of the Department of sociological data collection and analysis methods
_____20.
Department chairman _____ I.M. Kozina

Recommended by the Board of Education for Sociology _____20.
Chairman _____ V.G. Ledyaev

Approved by the Academic Board of the Department of Sociology_____20.
Academic Secretary _____E.V. Nadezhdina (signature)

Moscow, 2011

Areas of use and reference codes.

This educational discipline program establishes the minimal standards attained for students' knowledge and skillset, and determines the content and types of academic studies and reporting.

The program is intended for professors teaching this particular discipline, teaching assistants, and students in the 040100.62 Sociology.

The program was developed in accordance with:

- The educational standards set forth by Federal State Autonomous Educational Institution for Higher Professional Education National Research University-Higher School of Economics.
- Educational program of 040100.62 Sociology.
- The University's working academic plan for the 040100.62 Sociology that was approved in 2011.

1. Goals for studying the discipline.

The goals for studying the "Sociological data analysis -1" discipline are:

– the study of and practical mastery of the basic methods of statistical data analysis on social sciences
– the study of and practical mastery of computer programs used for statistical data analysis (SPSS package)
– the acquired understanding of the specificity of working with quantitative data in social sciences, understanding types of tasks that can be resolved by using statistical methods.

2. Scope of student knowledge that will be acquired as a result of mastering the discipline.

As a result of mastering the discipline, the student will:
– know statistical data analysis methods that are covered in the course program
– be able to pose and understand sociological problems that can be solved with the help of statistical data analysis as well as understand the specifics of data that is used in statistical analysis
– have the skills for conducting independent statistical data analysis on a computer using SPSS.

As a result of learning the discipline, the student acquires the following expert skills:

Expert skill	Code in FSES/ SRU	Descriptions – main signifiers of acquired knowledge (markers of reaching results)	Forms and methods of teaching that aid in formation and development of skills
Ability to use the main laws of natural sciences in professional activities, use mathematical analysis, modeling, theoretical and experimental research methods (*partial development)*	OK-11	– used statistical analysis methods for social pattern research – interprets statistical data analysis results (statistical conclusion)	Lectures, practice sessions, reading recommended literature from the provided list
Mastery of main methods and ways of obtaining, keeping, processing data, skills in working on a computer as an information management method (*partial development)*	OK-13	- uses a computer for statistical data analysis	Practical sessions, independent work
Knowledge of a foreign language at a	OK-	– uses English-language	Reading English-

level sufficient for conversational communication, as well as for finding and analyzing foreign data sources (*partial development)*	15	information sources to search for information on statistical data analysis – demonstrates command of the English language sufficient for searching for information on statistical data analysis	language literature from the recommended literature list
Ability to independently formulate goals, set specific objectives of research in various fields of sociology and solve them with the help of modern research methods with the latest domestic and foreign expertise and using modern equipment, equipment, information technology *(partial development)*	PK-2	– demonstrates the ability to solve posed sociological problems by using statistical methods of data analysis (such as analysis of variance, regression analysis, factor analysis, etc) – evaluates the possibility for solving posed sociological problems by using statistical methods	Lectures, practice sessions, independent work
Ability and willingness to participate in outlining and executing scientific and technical documents, research reports, and presenting research results considering the potential audience (*partial development)*	PK-3	– demonstrates an ability to format the results of statistical data analysis in written work in accordance with academic publications standards	Independent work
Ability to process and analyze data and prepare analytical solutions, expert conclusions and recommendations (*partial development)*	PK-8	– mastered statistical data analysis methods studied in the course – uses statistical data analysis methods for sociological problem analysis	Practice sessions, independent work
Ability and willingness to plan and implement project activities in the field of public opinion research, organization of marketing services *(partial development)*	PK-9	– has mastered methods of analyzing data from public opinion polls – correctly interprets results from public opinion polls	Practice sessions, independent work

4. The discipline's place within the overall educational program structure.

This discipline is compulsory for the professional cycle of the Bachelor's 040100.62 Sociology program.

The study of this discipline is based on the following prerequisite disciplines:
--Probability theory and mathematical statistics

The main points/theses of this discipline must be subsequently used in studying the following disciplines:
– Sociological data analysis -2

– Introduction to measurement methods in sociology
– Log-linear analyses in sociology
– Latent character analysis methods
– Classification methods in sociology
– Non-parametric statistical methods in sociology
– Academic research seminar

4. Topic plan for the academic discipline

№	Name of section	Total hours	Lecture Hours	Seminar Hours	Practice sessions	Self-Study Hours
	Part 1					
1	Introduction to statistical data analysis in sociology	18	2	-	6	10
2	Descriptive statistics and statistical graphics	18	2	-	6	10
3	Variance analysis	18	2	-	6	10
4	Correlation and paired linear regression	18	2	-	6	10
5	Multiple linear regression	36	4	-	12	20
6	Regression models for binary and categorically dependent variables	36	4	-	12	20
7	Analysis of contingency tables	18	2	-	6	10
8	Identification of cause-and-effect relationships and regression analysis	18	2	-	6	10
	Total	**180**	**20**	**-**	**60**	**100**
	Part 2					
9	Factor analysis	28	2	-	6	20
10	Cluster analysis	28	2	-	6	20
11	Decision tree models	16	2	-	2	12
	Total	**72**	**6**		**14**	**52**

5. Forms of student scholastic attainment control.

Part 1.

Type of control	Form of control	1 year	Parameters			
		1	2	3	4	
In-progress (once every two weeks)	Homework			5	5	Written homework assignment (once every two weeks)
Midterm	Midterm			1	1	Independent work on a computer in

	exam					the lecture hall
Final	Final exam				1	Written examination

Part 2

Type of control	Form of control	1 year	Parameters			
		1	2	3	4	
In-progress	Homework, independent work	2				Written homework assignment (once every two weeks) Each seminar, students get an assignment for independent work, to be completed on a computer in the lecture hall
Final	Final exam	1				Written examination

5.1 Criteria for grading of skills and scholastic attainment

Assessment for all forms of current, intermediate and final control are on a 10-point scale.

Grades for homework, tests and exams are based on the following criteria:

- Statistical validity of problem solution
- Statistical completeness of problem solving
- Presence and correctness of the sociological interpretation of statistical conclusion,
- Correct formatting of statistical conclusion.

6. Discipline contents

Topic 1. Introduction to statistical data analysis in sociology.

Which social science problems are solved by using statistical data analysis? Examples of research studies built on statistical data analysis. Statistical data analysis in various social sciences (sociology, economics, political science, psychology). Sources of data for statistical analysis: polls, census, aggregated statistics. Main methods of analysis. Overview of course content.
SPSS basics. SPSS syntax.

Topic 2. Descriptive statistics and statistical graphics.

Measurement scale (interval, ordinal, categorical). Mean values, median, mode. Dispersion. Distributions and their characteristics.
Statistical visualization. Curve of frequency distribution. Scatter diagram. Histograms. Bar charts. Line charts. Box plot. Pie charts.

Topic 3. Variance analysis.

The concept of statistical hypothesis. Hypothesis testing. Tests for equality of the mean and proportion. Variance within and between groups. T-test and regression analysis. Univariate and multivariate analysis of variance. A graphical representation of the results of variance analysis.

Topic 4. Correlation and paired linear regression.

Pearson's correlation coefficient as a measure of the relationship between two interval variables. Paired linear regression. The relationship between correlation and regression. Assessment and interpretation of the regression coefficients and standard errors. The statistical significance of coefficients. Regression with dummy variables.

Topic 5. Multiple linear regression.

The concept of statistical control. Evaluation and interpretation of regression coefficients in multiple regression models. Interaction effects. R^2 coefficient of determination. Assumptions of multiple regression models and diagnosis models. Heteroscedasticity. Nonlinear relationships. Statistical outliers. Multicollinearity. Principles of regression model construction.

Topic 6. Regression models for binary and categorically dependent variables.

Linear probability model. Logistic regression. Interpretation of logistic regression coefficients. Odds and odds ratios. Predicted probabilities. Multinomial logistic regression. Ordinal logistic regression. Presentation of the effects of linear and logistic regression.

Topic 7. Analysis of contingency tables.

Cross tables. Chi-square test. Odds ratios. Observed and expected frequencies. Log-linear model. Saturated model. The use of log-linear models in social mobility research.

Topic 8. Identification of cause-and-effect relationships and regression analysis.

Identification of cause-and-effect relationships and regression analysis. Neumann-Rubin's counterfactual model. The problem of omitted variables. The problem of reverse causality. Possible solutions: a fixed-effects regression, instrumental variables, matching.

Topic 9. Factor analysis.

Factor analysis model as a model of latent variables. The different approaches to determining the number of factors. Percentage of explained variance as an indicator of the quality of the factor model. Individual values of factors. Saving factors as new variables. The rotation matrix of factor loadings. Orthogonal and non-orthogonal rotation methods.

Topic 10. Cluster analysis.

Agglomerative hierarchical cluster analysis. Cluster analysis of k-means. The problem of choosing a measure of distance and the shape of the cluster. The problem of stability of clustering. Methods for assessing sustainability. The description and interpretation of clustering results.

Topic 11. Decision tree models.

Features of the algorithms CHAID and CRT. Classification errors, the definition of risk. Misclassification (Misclassification Cost). Ability to set different values of misclassification costs for different types of errors, the interpretation of relevant models. Checking the quality of model solutions to the problem of excessive tree trimming: cross-check, V-fold cross-validation, pruning.

7. Educational technologies.

Lectures and practice sessions are held in a computer lab.

8. Grading means of student progress and overall student assessment.

8.1 Topics for in-progress control assignments.

Every two weeks the students will be given a written homework assignment, that will include 3-4 problems to be solved in SPSS. The problems usually include independent data analysis, sociological interpretation of analysis results and written report on the results.

9. Discipline grading.

Grades for the first part of the course are composed of grades for practice session attendance (20%), homework (30%), two written independent works at the end of each module (15% each), and a written examination at the end of the fourth module (20%). Therefore, the cumulative grade will be 80% of the final grade for the course, whereas the grade for the final written examination – 20% of the final grade for the course. The grade for the final written examination is considered a "blocking" grade, i.e. a student who fails the examination will not be able to get a passing grade for the course.

Written homework is to be submitted every two weeks. Homework will not be accepted late, except for cases where they cannot be submitted on time with good reason (these cases should be discussed individually with the instructor). In case of illness, the student must make arrangements with the instructor teaching practice sessions about the deadlines for homework submission (providing medical documents proving illness). The deadline for missed homework in this case cannot exceed two weeks from the day of recovery. Homework assignments cannot be submitted after the final written examination has been taken under any circumstances.

Written independent work is to be done in the computer lab. They consist of solving a number of problems using SPSS and take two academic hours to complete. It is allowed to use any materials except for conferring with other students or third parties. In case of illness or absence (or other reasonable circumstances) on the day of the written independent work is due, the student must notify the instructor teaching practice sessions in advance (prior to the beginning of independent work session). Allowing for repeated opportunities to make-up independent written work for the students who have missed it on regularly scheduled days is entirely up to the instructor teaching practice sessions.

Written examination consists of two parts: 1) answer a number of theoretical questions 2) interpreting SPSS tables with statistical analysis results that were obtained using one of the methods studied in the course. The use of all materials is allowed during written examination. Written examination lasts 2 academic hours.

The cumulative grade is calculated by using a weighed average grade for the control items listed above (practice session attendance, homework, independent work). Weighted average grades are rounded up to the nearest whole number (for example, 7.3 is rounded to 7, 7.6 is rounded to 8, 7.5 is rounded to 8). An exception to this is the rounding of grades less than 4, which are always rounded down (3.99 is rounded to 3).

The final grade is calculated based on the weighted average grade for the cumulative control (80%) and the written examination (20%). In calculating the average, rounded grades for cumulative control and written examination are used. In rounding the final grade the same rules apply as when rounding the cumulative control grade. Therefore, if a student received a grade of 3 for cumulative control, he must get an 8 or higher on the final examination in other to pass the course (3*0.8*0.2=4). A student who receives a grade of 2 or lower for cumulative control cannot receive a passing grade for the course, other than through commission review (2*0.8+10*0.2=3.6, rounded to 3).

If a student is re-taking the final examination, he is not allowed to gain extra points to compensate his grade for cumulative control.

Assignments during commission reviews are composed of two parts: 1) independently solving problems in SPSS and 2) answering a number of theoretical questions and interpreting

statistical tables. Each part is 60 minutes long with a 15 minute break between the parts. Answers are to be in written form. It is forbidden to use any materials during commission review. Cumulative grade counts during commission review as follows: if a student has a grade of 3 or lower for the cumulative grade, he cannot get a final grade higher than 4 for the course overall, regardless of knowledge displayed for the commission. For students who have a cumulative grade of 4 or higher, the normal scheme of grade weighing is used (80% - cumulative grade, 20% - grade for commission review).

Grades for the second part of the course are formed as follows: grades for work during the semester including seminar work and homework (deadlines determined by the instructor, assignments submitted late without a reasonable excuse are not accepted; if such an excuse is present, the instructor may give a later deadline for the assignment) account for 30%; grades for the report on data analysis results, collected during course work on sociological research methods, account for 30% of the final grade; and the written examination at the end of the second part of the course amounts to 40% of the final grade. Grades for the examination are blocking.

If a student receives a grade of 2 or less for course work during the semester, then the accumulated points, including the grade for the course project report, do not count towards the final grade, so the cumulative grade becomes zero, and the final grade is composed only of the results of the final examination, and therefore cannot be higher than 4.

The final grade is originally formed on a 100-point scale and is rounded as follows:

Grade on 100-point scale	**Grade on 10-point scale**
0-14	1
14-24	2
24-39	3
40-47	4
48-57	5
58-67	6
68-77	7
78-87	8
88-97	9
98-100	10