Правительство Российской Федерации

Государственное образовательное бюджетное учреждение

высшего профессионального образования

Национальный исследовательский университет - Высшая школа экономики»

##### Факультет экономики

Программа дисциплины

Анализ и редактирование данных с пропусками и ошибками

для направления 080100.62 Экономика подготовки бакалавра

Автор: А.Б. Луппов

#### Рекомендована секцией УМС Одобрена на заседании кафедры

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ бизнес-статистики

Председатель Зав. кафедрой

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.Б. Луппов

«\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20 г. «\_31\_»\_\_августа\_\_\_\_\_\_\_\_\_\_\_\_ 2014 г.

Утверждена УС факультета

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ученый секретарь

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

« \_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20 г.

Москва 2014

**Тематический план учебной дисциплины**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Название темы | Всего часов по дисцип-лине | Аудиторные часы | Самосто-ятельная работа |
| Лекции | Сем. и практ. занятия |
| 1 | Концепция качества результатов проведения статистических обследований | 8 | 2 | 2 | 4 |
| 2 | Методы редактирования первичных данных наблюдения | 48 | 8 | 8 | 32 |
| 3 | Методы статистического анализа неполных данных | 52 | 10 | 6 | 36 |
|  | Итого: | 108 | 20 | 16 | 72 |

Базовые учебники

1. Суринов А.Е. и др. Статистика: Учебник. – М.: Изд-во РАГС, 2005. – 656 с. сс. 21-43.
2. Литтл Р. Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками / Пер. с англ. А. И. Никифорова / М.: «Финансы и статистика», 1991.
3. Кокрен У. Методы выборочного исследования. М.: «Статистика», 1976.
4. Carl-Erik Sarndal, Bengh Swensson, Jan Wretman (1995). Model Assisted Survey Sampling. Sprenger-Verlag New York, Inc.
5. Leslie Kish (1995). Survey Sampling. A Wiley- Interscience Publication. John Wiley&Sons Inc.
6. William G. Cochran. Sampling Techniques, third edition. John Wiley&Sons Inc., ISBN 0-471-16240-X.

Формы контроля:

**Итоговая оценка по учебной дисциплине складывается из следующих элементов:**

* Работа на практических занятиях (выполнение заданий);
* Домашняя работа;
* Зачет (контрольное мероприятие).

с весами соответственно 0,3; 0,3 и 0,4.

**Содержание программы**

Тема 1. Концепция качества результатов проведения статистических обследований

Основные понятия и положения концепции качества проведения и публикации результатов статистических обследований. Принципы планирования и управления проведением статистического наблюдения. Стандартные требования к качеству статистической информации на этапах проведения статистического наблюдения. Определение цели статистического исследования и выяснение потребностей пользователей в статистической информации. Требования к программе и дизайну вопросника. Оценка качества данных бизнес регистров и других административных баз данных. Критерии приемлемости применяемой выборки. Управление процессом сбора первичных данных. Автоматизированное редактирование первичных данных наблюдения. Идентификация и статистическая обработка данных нетипичных единиц наблюдения и пропусков, связанных с неответами респондентов. Содержание, структура и оформление статистической публикации. Компоненты метаданных, сопровождающих публикацию результатов статистического исследования.

**Основная литература:**

1. Суринов А.Е. и др. Статистика: Учебник. – М.: Изд-во РАГС, 2005. – 656 с. Глава 2.
2. Основные методологические положения по оценке качества статистической информации // официальный сайт Росстата. <http://www.gks.ru/bgd/free/meta_2010/Main.htm>.

**Дополнительная литература:**

1. Кокрен У. Методы выборочного исследования. / М.: «Статистика», 1976. Глава 13.
2. Carl-Erik Sarndal, Bengh Swensson, Jan Wretman. / «Model Assisted Survey Sampling» / Springer-Verlag, New York Berlin Heidelberg London 1991 1922, Sprenger-Verlag New York, chapter 17.

Тема 2. Методы редактирования первичных данных наблюдения

Стратегия повышения качества статистических данных. Процесс редактирования первичных данных. Качество статистики и эффективность редактирования. Процесс измерения и стоимость данных. Классификация наблюдений на приемлемые и подозрительные. Коррекция наблюдений, признанных ошибочными (импутация). Методы редактирования: обычные и типа Филледжи-Хольта. Спецификация правил редактирования. Проверка состоятельности правил редактирования. Обнаружение выделяющихся значений. Характеристика системы Филледжи-Хольта. Локализация ошибок. Использование весов. Импутация: детерминистический метод, методы донора (случайный выбор, ближайший сосед), метод оценок, множественная импутация. Импутация с применением нейронных сетей. Программные средства общего пользования.

**Основная литература:**

1. William E. Winkler / «Data Quality: Automated Edit/Imputation and Record Linkage» / RESEARCH REPORT SERIES (Statistics #2006-7) / Statistical Research Division, U.S. Census Bureau, Washington, DC 20233-9100.
2. Carl-Erik Sarndal, Bengh Swensson, Jan Wretman. / «Model Assisted Survey Sampling» / Springer-Verlag, New York Berlin Heidelberg London 1991 1922, Sprenger-Verlag New York, chapter 16.

**Дополнительная литература:**

1. Yancey, W.E., and Winkler, W. E. (2004), «BigMatch Software,” computer system, documentation (http://www.census.gov/srd/www/byyear.html).
2. Winkler, W. E. (2006), “Overview of Record Linkage and Current Research Directions,” U.S. Bureau of the Census, Statistical Research Division Report (http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf ).

Тема 3. Методы статистического анализа данных с ошибками и пропусками

Структуры ошибок в данных. Механизмы возникновения ошибок в данных. Планирование экспериментов и ошибки в данных. Быстрые методы обработки многомерных данных с ошибками. Обработка ошибок в данных выборочных обследований. Теория выводов, основанных на применении функции правдоподобия. факторизация правдоподобия для методов, когда механизм порождения ошибок игнорируется. Метод максимального правдоподобия для структур ошибок в данных общего вида: механизм порождения ошибок игнорируется. МП-оценивание в задачах, связанных с многомерным нормальным распределением. Анализ частично классифицированных таблиц сопряженности без учета механизма порождения ошибок. Смешанная модель для нормально и ненормально распределенных неполных данных с игнорированием механизма порождения ошибок. Модели с неигнорируемым механизмом порождения ошибок. Модельный подход к обработке неполных данных из-за ошибок при проведении выборочных обследований: Байесовские модели и многократное заполнение пропусков.

**Основная литература:**

1. Литтл Р. Дж.А., Рубин Д.Б. / «Статистический анализ данных с пропусками» / М.: «Финансы и статистика», 1991. Главы 1-12.

**Дополнительная литература:**

1. Кокрен У. «Методы выборочного исследования». / М.: «Статистика», 1976.
2. Carl-Erik Sarndal, Bengh Swensson, Jan Wretman. / «Model Assisted Survey Sampling» / Springer-Verlag, New York Berlin Heidelberg London 1991 1922, Sprenger-Verlag New York, Inc. 1995.
3. Leslie Kish / « Survey Sampling » / A Wiley- Interscience Publication. John Wiley&Sons Inc. New York, Chickhester, Brisbane, Toronto, Singapure, 1995.

**Вопросы для оценки качества освоения дисциплины**

1. Что понимается под качеством статистической информации?
2. Перечислите основные компоненты системы контроля качества статистических данных на этапах их формирования.
3. Определите аспекты качества статистических данных, выражаемые в терминах целостности, востребованности, точности и достоверности, своевременности, доступности, интерпретируемости и сопоставимости.
4. Назовите основные этапы процесса организации и проведения статистических обследований, получения и публикации их результатов.
5. Какие потенциальные источники возникновения ошибок при проведении обследований вы знаете?
6. Какие основные требования качества предъявляются к процессу проектирования и использования статистических вопросников?
7. Объясните понятие актуальности данных бизнес регистров и других административных баз данных, а также, что означает термин «ложная активность» единиц.
8. Какие характеристики точности оцененных по данным выборки показателей программы обследования вы знаете?
9. Объясните разницу между точностью и достоверностью оцененных по выборке итогов обследования.
10. Объясните используемые при сборе данных термины: недостижимость респондента и неответ респондента.
11. Для чего нужен этап редактирования первичных данных наблюдения?
12. Можете ли вы указать какой-либо формальный критерий различия между нетипичными и экстремальными значениями в данных наблюдения?
13. Чем определяется качество публикации со статистическими данными с точки зрения потребителей?
14. Что такое метаданные статистической публикации и какие ее компоненты вы можете назвать?
15. Что понимается под редактированием первичных данных наблюдения?
16. Укажите два основных этапа процесса редактирования первичных данных.
17. Назовите базовые индикаторы, которые обычно отслеживаются в процессе редактирования для управления системой обработки первичных данных.
18. Укажите критерии выбора оптимальной стратегии редактирования первичных данных наблюдения.
19. Объясните разницу между обычными методами редактирования и методами типа Филледжи-Хольта.
20. Перечислите основные типы правил редактирования, которые обычно применяются на практике.
21. Назовите три вида уравнений баланса правил редактирования.
22. В чем заключается проверка состоятельности правил редактирования?
23. Опишите метод Хидироглы-Бертелота обнаружения выделяющихся значений.
24. Опишите метод редактирования данных Филледжи-Хольта.
25. Каковы особенности метода детерминистской импутации данных наблюдения?
26. В чем заключается метод донора для импутации данных наблюдения и какие его основные разновидности вы знаете?
27. В чем заключается метод оценок для импутации данных наблюдения и какие его основные разновидности вы знаете?
28. Опишите выделяемые структуры пропусков в данных.
29. Какие механизмы порождения пропусков в данных вы знаете?
30. Опишите метод наименьших квадратов для обработки неполных данных.
31. Какие преимущества и недостатки имеют методы взвешивания для обработки пропусков в выборочных данных.
32. Отметьте основные особенности теории выводов по неполным данным, основанных на применении функции правдоподобия.
33. В чем состоит метод максимального правдоподобия для обработки структур пропусков в данных?
34. Какие критерии для подбора моделей по частично классифицированным данным вы можете предложить?
35. Опишите различия в теории правдоподобия для моделей с неигнорируемыми и игнорируемыми механизмами порождения пропусков в данных.
36. В чем состоит Байесовский подход (модель) для обработки данных обследования с пропусками?
37. Отметьте достоинства и недостатки метода многократного заполнения пропусков в данных обследований.

**Автор программы:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ А.Б. Луппов/