Подготовка собранных данных к анализу

Лекция 10

Звоновский, к.с.н.

Основные используемые понятия

После окончания полевых работ собранные данные никогда не находятся в виде, приемлемом для анализа. Подготовка данных к анализу состоит из двух этапов: редактирование данных и формирование массива для обработки.

Редактирование данных – проверка, коррекция и фильтрация собранных в результате полевых работ данных, расположенных на бумажных и электронных носителях.

Формирование массива представляет собой доведение массива данных до состояния, в котором возможна обработка первичных данных согласно программе исследования и поставленных в нем целей и задач

Редактирование данных

Полевое редактирование – редактирование, выполненное в отношении части собранных работ (собранных одним интервьюером, на одной территории, под руководством одного супервайзера)

Офисное редактирование – редактирование, выполненное в отношении всей массы собранных первичных документов

Редактирование чаще всего состоит из двух частей – контроль выполнения полевым персоналом **инструкции по сбору** данных (методики и правил), а также – проверку **полноты заполнения** бланков первичных документов.

В современных методах сбора данных в полевой инструмент заложена возможность контроля

Редактирование данных

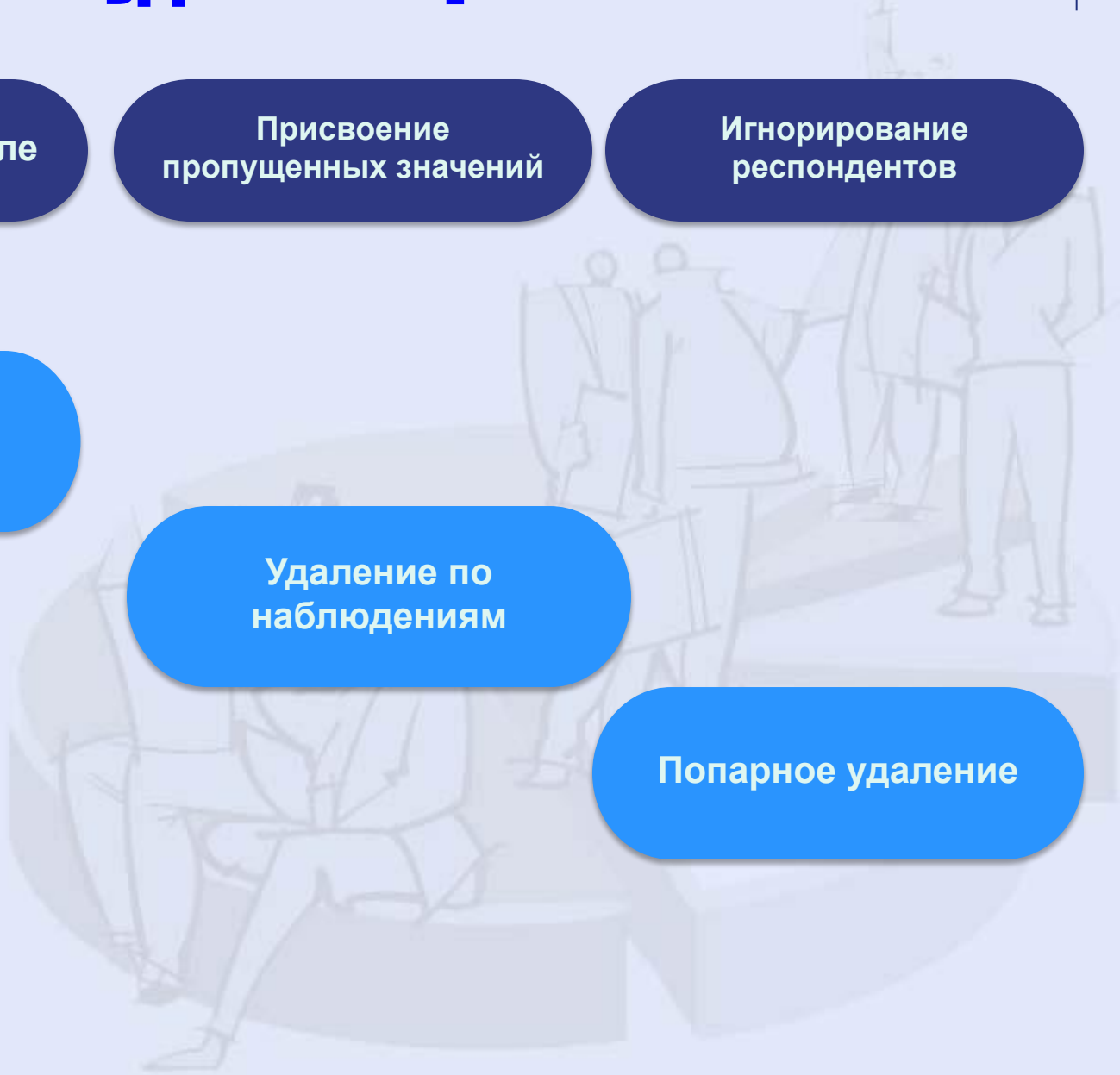
Невыполнение всех или части требований по методу и правилам **сбора всех или части данных** может привести к существенным искажениям результатов и невозможности достичь поставленные цели и задачи исследования. Чаще всего, внесенные искажения невозможно исправить.

Отсутствие части информации в собранных анкетах (бланках интервью и пр.) может быть устранено после окончания полевых работ путем обработки неудовлетворительных ответов.

Неудовлетворительные ответы – зафиксированные или незафиксированные ответы отдельных респондентов, делающие невозможными их обработку вместе с другими единицами наблюдений, а также их перекодирование.

Обработка неудовлетворительных ответов

- Возвращение в поле
- Присвоение пропущенных значений
- Игнорирование респондентов
- Замена на нейтральное значение
- Удаление по наблюдениям
- Попарное удаление



Кодирование открытых вопросов

Кодирование открытых вопросов – присвоение кода, чаще всего, численного, для представления ответа на конкретный вопрос, предполагавший только собственную формулировку респондента.

Проблема – респондент может отвечать в любой лексике, так, как он понял вопрос, и так, как он хочет ответить.

Чем Вам понравился главный герой фильма? – Он крутой.

Что Вы не едите за завтрак? – Обед и ужин.

Как Вы оцениваете деятельность Президента? – Я не довольна работой нашего ТСЖ

Кодирование открытых вопросов

57. Скажите, пожалуйста, почему Вы не удовлетворены своей работой? Чем именно Вы не удовлетворены?	**260**
очень высокая загруженность	1
низкая заработная плата	1
zarplata i neracionalnoe ispolzovanie dnya	1
адекватной оплатой	1
большая загруженность бумагами	1
большая загруженность и ответственность	1
большая загруженность из-за нехватки врачей	1
…..	..
хочется больше общаться с пациентами, че с бумагами	1

1	Зарплата
2	Материально-техническое обеспечение (недостаток оборудования, устаревшее оборудование, нехватка материалов, реактивов и прочее, слабое финансирование)
3	Негативное отношение населения
4	Непрофессионализм начальства (больше интересуются правильностью заполнения бумаг, нежели качеством лечения, неудовлетворительная организация процесса работы, отношение начальство и отсутсвие с его стороны поддержки)
5	Большая нагрузка (большой поток пациентов)

Кодирование переменных

Кодирование переменной с единственным возможным численным значением – создание одного поля одного из цифровых форматов для данной переменной.

Кодирование переменной с несколькими возможными численными значениями – создание нескольких полей одного из цифровых форматов для данной переменной.

Кодирование переменной с одним или несколькими возможными нечисленными значениями – создание одного или нескольких полей текстового формата для данной переменной.

Кодировочная книга (codebook)

Кодировальная книга – таблица соответствий между собранными данными и переменными электронного массива данных. Описывает правила преобразования информации, имеющейся в полевых документах в коды, используемые при анализе данных.

Корме полевой информации, включает в себя служебные данные – например, номер проекта, номер оператора, время переноса данных и пр.

Кодировочная книга (codebook)

Номер переменной	Имя переменной	Номер вопроса	Инструкции по кодированию
1	Идентификатор респондента		От 001 до 890, добавляя первые нули, где нужно
2	Код проекта		31 (одинаков для всех респондентов)
3	Код интервьюера		Кодировать так же, как в анкете
4	Код данных		Кодировать так же, как в анкете
5	Код времени		Кодировать так же, как в анкете
6	Код верификации		Кодировать так же, как в анкете
7	Кто совершает покупки	Д1	Мужчина-глава семьи = 1 Женщина-глава семьи = 2 Другой = 3 Введите указанный номер Пропущенные значения = 9
8	Осведомленность о магазине 1	Д2	Для вопроса II частей от a до j Введите указанный номер Не очень осведомлен = 1 Очень осведомлен = 6 Пропущенные значения = 9
9	Пол	Д3	Кодировать так же, как в анкете Пропущенные значения = 9
10	Возраст	Д4	Кодировать так же, как в анкете Пропущенные значения = 9
11	Возрастная группа		Макет пересчитает переменную 11 в возрастные группы

Перенос данных в электронный массив

CAPI/CAWI/CATI

Ввод с клавиатуры

Цифровое сканирование (сканеры специальных кодов)

Оптическое сканирование

Перенос данных в электронный массив

Ввод с клавиатуры

При переносе данных из бумажного вида в электронный возникают ошибки, связанные с тем, что этот **перенос** выполняет **человек**. При этом чаще всего данных так много, что к переносу привлекается большое число неквалифицированных сотрудников.
Для борьбы с этими ошибками используют несколько техник.

1. **Выборочный контроль введенных данных.** Супервайзер сравнивает какую-то долю – обычно 10% - бумажных и электронных вариантов записей. Если число ошибок в этой доле превышает заранее уговоренный уровень, проверяются уже 20% ввода и т.д.
2. **Повторный ввод**. Данные переносятся из бумажный формы в электронную дважды, желательно, различными сотрудниками. Для анализа используется объединенный массив.

Распределения данных как контроль ввода

В какой отрасли Вы работаете, какова сфера Вашей деятельности?	
промышленное производство (в т.ч. добывающие отрасли)	155
сельское, лесное, рыболовное и т.д. хозяйство	17
строительство	35
сфера услуг, сервиса, бытового обслуживания	66
общественное питание, ресторанный бизнес	18
…	…
судебные органы, юриспруденция	4
транспорт, складское хозяйство	30
другое	18
ВСЕГО	**490**
Missing	*310*

Получение и обзор первичных таблиц линейного (частотного) распределения значений измеряемых переменных позволяют увидеть возможные ошибки при сборе и переносе данных.

Перекрестные таблицы

	Если бы в следующее воскресенье в нашей области проходили выборы в законодательное собрание, за какую партию Вы бы, скорее всего, проголосовали?										
	Справедливая Россия	ЛДПР	КПРФ	Единая Россия	Патриоты России	Правое дело	Яблоко	другие	испортят бюллетень	не голосуют	не определились
Самара	10	25	25	114	3	2	2	8	4	48	46
Тольятти	9	23	28	64	0	1	3	5	1	22	7
малые города	0	17	7	56	0	1	4	0	6	34	30
села	4	8	36	37	0	1	0	2	5	24	47
ВСЯ область	**38**	**73**	**82**	**281**	**3**	**5**	**9**	**15**	**16**	**128**	**150**

Перекрестный анализ первичных данных позволяет обнаружить наиболее заметные ошибки, возникшие при сборе данных

Проверка гипотез

Нулевая (null) гипотеза – H_o - гипотеза о том, что полученные результаты не показывают никакого значимого различия между группами генеральной совокупности.

Альтернативная гипотеза – H, - гипотеза, утверждающая, что существуют значимые различия между отдельными группами генеральной совокупности.

Нулевая гипотеза может отвергнута, но она не может быть принята на основании лишь одной проверки.

Результатов проверки нулевой (пустой) гипотезы может быть два – принята нулевая гипотеза (т.е. различий нет) и принята альтернативная гипотеза (т.е. различия есть)

Проверка гипотез

Варианты гипотез:

Среднее количество кинотеатров, которые посещают жители города, составляет 3,0

Более 10% домохозяйств постоянно делают покупки в одних и тех же универмагах

Сторонники двух различных кандидатов отличаются по своим социально-демографическим характеристикам

Одна гостиница имеет более привлекательный образ, чем ее ближайший конкурент

Большая осведомленность об авторе книги приводит к более позитивному отношению к его произведениям

Проверка гипотез

Односторонний критерий (тест) – проверка нулевой гипотезы, когда альтернативная гипотеза выражена направленно.

Например, мы предполагаем, что доля рынка, занятая данным сортом пива, превышает 20%.

Значит, H_o: $\rho \leq 0{,}20$, а $H_,$: $\rho \geq 0{,}20$

Двусторонний критерий (тест) – проверка нулевой гипотезы, когда альтернативная гипотеза выражена ненаправленно.

Например, мы предполагаем, что за нашего кандидата проголосует 10% избирателей округа.

Значит, H_o: $\rho \neq 0{,}20$, а $H_,$: $\rho = 0{,}20$

Проверка гипотез

Предположим, что мы должны вывести на рынок новый бренд пива, в случае, если в целевой группе он будет занимать не менее 20%

Тогда H_o: $\rho \leq 0{,}20$, а $H_,$: $\rho \geq 0{,}20$

Мы можем принять **верное решение** в двух случаях**:**

1. Нулевая гипотеза отвергнута, и действительно доля потребителей больше 20%.
2. Нулевая гипотеза принята, и действительно доля потребителей меньше 20%.

Мы можем совершить ошибку, если

1. Нулевая гипотеза отвергнута, но в действительности доля потребителей меньше 20% (первого рода)
2. Нулевая гипотеза принята, и действительно доля потребителей больше 20% (второго рода).

Перенос данных в электронный массив

- Сформулировать H_0 и H,
- Выбрать подходящую статистику
- Выбрать уровень значимости
- Собрать данные и рассчитать проверочную статистику
- Определить вероятность выбранной статистики и сравнить с выбранным значением значимости
- Отклонить или принять H_0
- Сделать вывод и принять решение

Проверка гипотез

Выбор статистики – выбор способа измерения отклонения измеряемого значения от тестируемого уровня.

Если тестируется превышает ли доля рынка уровень в 10%, значит выбирается способ измерения значимости отличия измеренного значения от 20%.

Чаще всего используются нормальное (z), биноминальное распределение, распределение Стъюдента или хи-квадрат.

В данном случае мы будем использовать z-распределение для доли:

$$Z = (\bar{\rho} - \rho) / \sigma$$

$$\sigma = \sqrt{\rho(1-\rho)} / n$$

$$\sigma = \sqrt{0,2(1-0,8)} / 500 = 0,018$$

Проверка гипотез

Выбор уровня значимости – это выбор при котором может произойти ошибка первого рода.

Традиционно выбирается 95%. Это позволяет после получения выборочных значений как увеличить, так и уменьшить уровень значимости.

Предположим, что из 500 респондентов, 110 сообщили, что являются потребителями изучаемой марки пива. Тогда

$$\rho = 0{,}22$$

$$\sigma = \sqrt{0{,}2(1-0{,}8) / 500} = 0{,}018$$

$$Z = (0{,}22 - 0{,}20) / 0{,}018 = 1{,}111$$

Площадь под кривой – 0,733

Таким образом, нулевая гипотеза не отвергается.

Проверка гипотез

Мощность критерия – это вероятность отклонения нулевой гипотезы β (ошибка второго рода) , когда она ложна и должна быть отвергнута. Чем ниже α, тем выше β.

Критическое значение (выбранной) статистики – значение, при котором вероятности ошибки первого и второго рода равны.