Финансирование
Европейского Союза

ИНМИР

INSTITUTE FOR
WAR & PEACE REPORTING
IWPR
ИНСТИТУТ РЕПОРТАЖЕЙ ВОЙНЫ И МИРА

ПРАКТИЧЕСКОЕ **РУКОВОДСТВО**

ПО СБОРУ И АНАЛИЗУ ДАННЫХ ДЛЯ МИНИ-ИССЛЕДОВАНИЙ

Савия Хасанова

Алматы, Казахстан 2024

Авторка: Савия Хасанова
Корректор: Дамира Каженова
Дизайн и верстка: Диана Хигай
Под общей редакцией Азизы Мухаметовой

Данное руководство разработано специально для получателей финансирования в рамках проекта «Гражданское общество за Казахстан (CS4K)», финансируемого Европейским Союзом и реализуемого Институтом репортажей войны и мира (IWPR) совместно с ОФ «Институт Национальных и Международных Инициатив Развития (ИНМИР)».

Цель руководства — помочь организациям провести мини-исследования, которые станут основой для аналитических записок и адвокационных кампаний. Проект «CS4K» направлен на продвижение фундаментальных свобод и прав в Казахстане через активное участие гражданского общества. Основное внимание уделяется укреплению потенциала местных организаций гражданского общества (ОГО) в сфере адвокации и взаимодействия с сообществами для достижения положительных изменений, а также обмену опытом между организациями.

Проект реализуется международной неправительственной организацией IWPR в партнерстве с ОФ «ИНМИР» при финансовой поддержке Европейского союза.

ИНСТИТУТ РЕПОРТАЖЕЙ ВОЙНЫ И МИРА (IWPR)

– это международная организация, деятельность которой включает в себя проведение образовательных проектов для представителей государственных органов, экспертов, исследователей и журналистов. В Центральной Азии IWPR работает с 1999 года и имеет зарегистрированные офисы в Алматы, Бишкеке и Душанбе.

ЕВРОПЕЙСКИЙ СОЮЗ

– это экономический и политический союз 27 европейских стран. Он основан на ценностях уважения человеческого достоинства, свободы, демократии, равенства, верховенства закона и уважения прав человека, в том числе прав лиц, принадлежащих к меньшинствам. Он действует на глобальном уровне с целью продвижения устойчивого развития общества, окружающей среды и экономики во благо каждого.

Содержание

От авторки

Привет!

МЕНЯ ЗОВУТ САВИЯ ХАСАНОВА, И МЫ ВМЕСТЕ С КОМАНДОЙ ПРОЕКТА «ГРАЖДАНСКОЕ ОБЩЕСТВО ЗА КАЗАХСТАН (CS4K)», ПОДГОТОВИЛИ ДЛЯ ВАС ЭТО РУКОВОДСТВО.

Когда я только начинала делать свои первые исследования, мне приходилось собирать по кусочкам свои знания из занятий по статистике и аналитическим методам. Спустя более двадцати лет работы с данными, я оформила это в один простой документ, которого мне самой очень не хватало.

ЭТОТ ДОКУМЕНТ – не учебник, он не ставит перед собой цель предоставить вам фундаментальное обучение работе с данными. Это руководство очень практическое, в котором описаны шаги, даны советы и примеры (очень много примеров!) для проведения исследований в условиях ограниченных ресурсов.

ОСНОВНАЯ ЦЕЛЬ ДАННОГО РУКОВОДСТВА – дать исследователям и исследовательницам четкие и понятные инструкции по сбору, анализу и интерпретации данных.

В ПЕРВОМ РАЗДЕЛЕ «ЧТО ТАКОЕ ДАННЫЕ И ЗАЧЕМ ОНИ НУЖНЫ?» вы познакомитесь с основными понятиями, связанными с данными, и поймете, почему они важны для исследований. Мы рассмотрим различные типы данных, их свойства и применение. Также мы обсудим, как правильно начать их сбор с учетом дизайна исследования и его целей.

ВО ВТОРОМ РАЗДЕЛЕ «МЕТОДЫ СБОРА ДАННЫХ» мы подробно рассмотрим методы сбора данных, такие как опросы и анкетирование, интервью, получение данных из интернета. Поговорим мы и про выборку – можно ли обеспечить ее репрезентативность в ограниченных условиях?

В РАЗДЕЛЕ «АНАЛИЗ ДАННЫХ» мы немного поговорим о базовых методах статистического анализа: о средних величинах, о процентных соотношениях, о темпах прироста. Также в этом разделе мы попытаемся научиться избегать манипуляций с данными, чтобы результаты вашего исследования были объективными.

НАДЕЮСЬ, ЧТО ЭТО РУКОВОДСТВО СТАНЕТ ПОЛЕЗНЫМ ИНСТРУМЕНТОМ В ВАШЕЙ РАБОТЕ, ПОМОЖЕТ ВАМ ПОГРУЗИТЬСЯ В НАБОЛЕВШУЮ ИССЛЕДОВАТЕЛЬСКУЮ ТЕМУ И ВНЕСТИ ВКЛАД В РЕШЕНИЕ АКТУАЛЬНЫХ ПРОБЛЕМ В ВАШЕМ РЕГИОНЕ ИЛИ СФЕРЕ.

01.

ЧТО ТАКОЕ ДАННЫЕ И ЗАЧЕМ ОНИ НУЖНЫ?

В современном мире данных, когда вокруг нас вращаются терабайты информации, очень важно понимать, какие типы данных существуют и как их правильно использовать в исследованиях. От точных чисел и статистических расчётов до глубоких личных рассказов и мнений – каждый тип данных может удовлетворить различные исследовательские требования.

ТИПЫ ДАННЫХ

Данные – это зарегистрированная информация, которая может быть использована для обработки, анализа и принятия решений.

ДЛЯ ЧЕГО НУЖНЫ ДАННЫЕ?

Данные, при их качественной обработке и грамотном использовании, обладают огромной полезностью. На уровне организаций гражданского общества.

1) данные помогают принимать обоснованные решения, основываясь на фактах, а не на предположениях;

2) используя данные, можно проводить исследования, анализировать текущие тенденции и делать прогнозы как на уровне организации, так и в сфере деятельности;

3) при помощи данных проводится мониторинг и контроль, а также они помогают оптимизировать ваши финансовые ресурсы и управлять рисками.

ДАННЫЕ БЫВАЮТ:

- числовые – численность населения, уровень заболеваемости, рост в процентах;
- текстовые – результаты интервью, комментарии в социальных сетях, статьи и документы, книги;
- графические – схемы, графики, рисунки, фотографии;
- звуковые – человеческая речь, аудиозаписи, диктофонная запись интервью;
- видео – видеозаписи, фильмы, шортсы, рилсы.

В данном руководстве речь пойдет, в основном, о сборе числовых и текстовых данных.

[1] С английского языка shorts, reels – типы видео в YouTube и Instagram.

Данные, используемые в исследованиях, можно классифицировать по разным критериям. В Таблице 1 приведены основные виды данных, на которых и будет фокусироваться данное руководство.

ТАБЛИЦА 1. НЕКОТОРЫЕ КЛАССИФИКАЦИИ ДАННЫХ

КЛАССИФИКАЦИЯ	ТИПЫ	ОПИСАНИЕ	ПРИМЕР
ПО ФОРМЕ	Количественные	Представлены в виде чисел	Доход, возраст, число лет обучения
	Качественные	Описывают характеристики или свойства	Пол, национальность, образование
ПО ИСТОЧНИКУ ПРОИСХОЖДЕНИЯ	Первичные	Собраны непосредственно исследователями	Результаты опроса интервью, фокус-групп
	Вторичные	Данные, которые были собраны/сгенерированы ранее	Статистические данные, данные отчетов и исследований
ПО ДОСТУПНОСТИ	Открытые	Доступные данные для широкого круга пользователей (см. раздел «Открытые данные»)	Официальная статистика, опубликованная в интернете
	Закрытые	Ограничены в доступе, могут требовать специального разрешения	Персональные данные; данные, полученные по запросу
ПО СТРУКТУРЕ	Структурированные	Организованы в табличном виде, могут быть обработаны при помощи программ	Базы данных, анкеты с четкими вопросами
	Неструктурированные	Не имеют четкой структуры	Тексты, изображения, видео, аудио
ПО ДИНАМИКЕ ИЗМЕНЕНИЯ	Статичные	Отражают состояние явления на определенный момент времени	Число студентов вузов в 2024 году
	Динамические	Показывают изменение показателей во времени (по годам, месяцам, дням, часам и т.д.)	Динамика учащихся в вузах с 2010 по 2024 годы

Источник: компиляция различных источников, собственные примеры

Кроме того, есть еще номинальные и порядковые категории данных, интервальные, одномерные и многомерные, дискретные и непрерывные и пр. Подробнее о типах данных можно почитать здесь.

С ЧЕГО НАЧИНАТЬ СБОР ДАННЫХ: ДИЗАЙН ИССЛЕДОВАНИЯ

Прежде чем приступать к сбору данных, необходимо понять, о чем будет ваше исследование. Этот процесс очень важен, так как именно правильная постановка целей исследования и выбор его дизайна позволят обеспечить достоверность и надёжность результатов.

ЭТАП 1. ОПРЕДЕЛЕНИЕ ЦЕЛЕЙ И ЗАДАЧ ИССЛЕДОВАНИЯ

– четкое формулирование целей и задач исследования поможет определить, какую информацию нужно собрать и как ее потом использовать. Это помогает в выборе методов сбора данных и в их дальнейшем анализе.

Когда вы ставите цели, обратите внимание на то, что они должны быть:

ЧЕТКИМИ И КОНКРЕТНЫМИ

Цель исследования должна быть сформулирована ясно и однозначно. Она должна отражать то, что вы намерены выяснить в ходе исследования.

ПРИМЕР ПЛОХОЙ ПОСТАНОВКИ ЦЕЛИ:	ПРИМЕР ХОРОШЕЙ ПОСТАНОВКИ ЦЕЛИ:
«Проанализировать уровень занятости женщин в Казахстане».	«Изучить влияние гендерных стереотипов на занятость женщин в экономике Казахстана в последнее десятилетие».

РЕАЛИСТИЧНЫМИ

Цель должна быть достижима в рамках имеющихся у вас ресурсов. Если у вас небольшой бюджет, то замахиваться на обширное исследование не стоит. Трезво оцените, чего вы реально можете достичь в рамках бюджета и временных ресурсов и, при необходимости, сузьте вашу цель.

ПРИМЕР БОЛЕЕ ШИРОКОЙ ЦЕЛИ:	ПРИМЕР БОЛЕЕ УЗКОЙ ЦЕЛИ:
«Изучить влияние гендерных стереотипов на занятость женщин в экономике Казахстана в последнее десятилетие».	«Изучить влияние гендерных стереотипов на выбор женщинами специальности обучения в вузах Караганды».

СООТВЕТСТВОВАТЬ ПРОБЛЕМАТИКЕ

Большинство исследований проводятся для того, чтобы изучить причины и объяснить какую-то проблему и предложить рекомендации для ее решения. Поэтому цель должна быть тесно связана с исследовательской проблемой и соответствовать ей. Отталкиваясь на пример выше, проблемой может быть то, что в вузах Караганды *идет большой гендерный перекос среди студенток и студентов вузов – подавляющее число мальчиков учатся на технических специальностях, а девушек – на гуманитарных.*

ИЗМЕРИМЫМИ

Цели должны быть сформулированы таким образом, чтобы можно было оценить, достигнуты ли они. Когда вы формулируете цели, добавляйте в них или в гипотезы количественные измерения – например, «к 31 декабря опросить 300 респондентов».

ЭТАП 2. ВЫБОР ДИЗАЙНА ИССЛЕДОВАНИЯ

– вы определяете вид своего исследования, его временные рамки и формулируете гипотезы.

Исследования могут быть классифицированы по разным критериям, но основной из них – используемый тип данных (именно для этого мы о них и говорим!), методология их сбора и анализа. Существует три основных вида исследований: качественное, количественное и смешанное.

Качественные исследования нацелены на исследование глубинных причин, мотивов, контекстов поведения, мнений и восприятия людей. Для такого вида исследования в основном собираются качественные данные с помощью интервью, наблюдений и фокус-групп. Преимуществом этого вида исследования является полное погружение в исследуемую тематику, однако при этом его результаты гораздо сложнее генерализировать (распространить на генеральную совокупность).

Количественные исследования позволяют проводить количественную оценку и проверку гипотез, а также выявлять закономерности и тенденции. Этот вид исследования подразумевает сбор, обработку и анализ количественных данных, как первичных, так и вторичных. Большим преимуществом количественных исследований является объективность результатов, возможность обобщать выводы и переносить их на обширные группы людей.
При этом одним из недостатков такого вида исследований могут быть неправильные выводы из-за некорректных вычислений, расчётов или моделей.

СМЕШАННЫЕ ИССЛЕДОВАНИЯ

Смешанные исследования являются комбинацией качественных и количественных исследований, их методов сбора и анализа данных. Их преимуществом является более широкое понимание изучаемой проблемы, увеличение достоверности и точности полученных выводов. Однако такое исследование сложнее провести с точки зрения его планирования, временных и финансовых затрат.

ТАБЛИЦА 2. МЕТОДЫ СБОРА И АНАЛИЗА ДАННЫХ ДЛЯ РАЗНЫХ ВИДОВ ИССЛЕДОВАНИЙ

ИССЛЕДОВАНИЕ	МЕТОДЫ СБОРА ДАННЫХ	АНАЛИЗ ДАННЫХ
Качественное	Интервью Фокус-группы Наблюдение Анализ текстов	Тематический анализ Контент-анализ Анализ нарративов
Количественное	Опросы и анкеты Эксперименты Поиск вторичных данных	Статистический анализ Регрессионно-корреляционный анализ Описательные статистики
Смешанное	Проведение качественных интервью + Сбор и анализ количественных данных	Параллельный анализ данных: одновременная обработка качественных и количественных данных Последовательный анализ данных: один метод используется для предварительного исследования (например, качественное), а затем другой (количественное) для тестирования гипотез

После того, как вы поставили цели, задачи и определили, каким будет ваше исследование, необходимо сформулировать гипотезы (см. Вставку 1).

Источник: компиляция различных источников, собственные примеры

ВСТАВКА 1. ГИПОТЕЗА

Гипотеза – это утверждение, которое необходимо подтвердить или опровергнуть. В исследованиях под гипотезой, как правило, понимается предположение о существовании тех или иных особенностей процесса и влиянии определенных факторов на него.

Например, в известном исследовании о фемициде в Кыргызстане, гипотезой было предположение о том, что убийства женщин совершаются в основном близкими мужчинами и являются результатом систематического насилия.

Эта гипотеза, кстати, подтвердилась.

В зависимости от целей сферы вашей деятельности и целей исследования, гипотезы могут быть разными. Например, если ваше НПО занимается продвижением создания инклюзивной среды для всех, одной из целей вашего исследования может быть изучение инфраструктуры для людей с инвалидностью в государственных учреждениях Казахстана.

Для такого исследования вы будете выдвигать различные гипотезы:

- наличие пандусов не является/является достаточным;
- технические характеристики пандусов и лифтов не соответствуют нормативам об инклюзивной среде;
- в госучреждениях отсутствуют инклюзивная среда для незрячих (материалы шрифтом брайля, нет желтой плитки и пр.).

Для целей формулирования гипотез хорошо зарекомендовала себя практика мозгового штурма, в процессе которого члены вашей команды могут обсуждать и генерировать идеи для гипотез, проверять которые вы будете в своем исследовании.

После того, как вы выбрали каким будет ваше исследование, определили его цели и сформулировали гипотезы, можно приступать к выбору методов сбора данных, разработке инструментария и выборке. Об этом мы поговорим в следующей главе.

02.

МЕТОДЫ СБОРА ДАННЫХ

Как мы уже рассказывали выше, методы сбора данных зависят от того, каким будет ваше исследование. В данном руководстве мы фокусируемся на сборе данных для смешанных исследований, а именно: интервью, опросах и использовании вторичных данных.

ОПРОСЫ И АНКЕТЫ

Провести эффективный опрос и создать качественную анкету – то еще искусство. От того, какая ваша анкета – длинная или короткая, сложная или лаконичная, как расставлены вопросы и какие есть варианты ответов – зависит и заполняемость анкеты, и надежность результатов, и сложность обработки полученных данных.

ВОПРОСЫ В АНКЕТАХ БЫВАЮТ ЗАКРЫТЫМИ, ОТКРЫТЫМИ И ПОЛУЗАКРЫТЫМИ.

Закрытые – предлагают респондентам выбрать подходящий ответ из списка вариантов.

Открытые – позволяют респондентам самостоятельно написать ответ.

ЗАКРЫТЫЙ ВОПРОС:	ОТКРЫТЫЙ ВОПРОС:
Укажите ваш уровень образования: 1)Среднее 2)Профессиональное 3)Высшее	Напишите ваш уровень образования? _____

ВСТАВКА 2. ЧТО ТАКОЕ «ОПРОС», А ЧТО ТАКОЕ «АНКЕТА»?

Разница между опросом и анкетой заключается в том, что опрос – это сочетание вопросов, процесса сбора, обобщения и анализа ответов, в то время как анкета – это всего лишь набор вопросов.

Другими словами, анкета – это вопросник, а опрос – процесс, включающий в себя вопросник, методологию и анализ.

Полузакрытые вопросы – это комбинация открытого и закрытого вопросов, при которой респонденту предлагается сделать выбор из вариантов ответа или написать свой ответ, если ни один из предложенных не подходит (чаще всего это вариант ответа «другое»).

Например:

Укажите ваш уровень образования:
1)Среднее
2)Профессиональное
3)Высшее
4)Другое

Закрытые и полузакрытые вопросы могут принимать разные формы – от самых простейших вопросов с множественными вариантами ответа до более сложных –матричных вопросов.

Почему важно, какие вопросы бывают? В первую очередь потому, что чем проще вопросы и их форма, тем проще анализировать полученные результаты. Однако, если мы максимально упростим и опрос, и анкету, мы рискуем упустить какие-то важные моменты нашего исследования. Поэтому существует несколько правил, как создать хорошую сбалансированную анкету, которая привлечет респондентов и в то же время даст исследователям максимум нужной информации.

ОПРОС НЕ ДОЛЖЕН БЫТЬ СЛИШКОМ ДЛИННЫМ

Большинство респондентов идут вам навстречу, отвечая на анкету (конечно, если вы не мотивируете их денежным вознаграждением). Исследования социологических компаний показывают, что чем меньше вопросов в анкете, тем выше процент заполненных анкет.

СТАРАЙТЕСЬ ЗАДАВАТЬ БОЛЬШЕ ЗАКРЫТЫХ ВОПРОСОВ

Представьте, если каждый ваш вопрос будет открытым, в котором респонденту будет предложено самому или самой что-то написать? С одной стороны, вам будет это невероятно тяжело анализировать, ведь текстовые ответы труднее унифицировать. С другой стороны, такой опрос займет гораздо больше времени. Поэтому, если есть возможность избежать открытых и полузакрытых вопросов – смело делайте это.

НЕ ЗАДАВАЙТЕ СЛАБЫХ ВОПРОСОВ.

К слабым вопросам относятся наводящие, двусмысленные и предвзятые вопросы.

Наводящий вопрос – это вопрос, в котором уже есть какое-то мнение.
Например, представьте, что вы проводите опрос среди молодежи Казахстана о доступности аренды и покупки жилья. Вопрос «Считаете ли вы арендную плату за вашу квартиру завышенной?» будет наводящим, так как в нем вы уже даете мнение об арендной плате («завышенной»).

Вместо этого вопроса спросите:
«На ваш взгляд, арендная плата за вашу квартиру:
А) Завышена
Б) Соответствует моим возможностям и ожиданиям
В) Могу платить и больше».

Двусмысленный вопрос – это вопрос, когда в нем предлагается оценить две вещи. Например, вопрос **«Каково ваше мнение о ценах на аренду и покупку жилья в Казахстане?»** является двусмысленным, поскольку респондент должен выбирать на какую его часть – об аренде или покупке ему или ей отвечать.

Не бойтесь, двусмысленный вопрос легко переформулировать, либо разделив его на два, либо оставив лишь одну его часть.

Предвзятый вопрос – вопрос, в котором, вам фактически говорят, что нужно ответить. Например, вопрос **«Что вы думаете о государственной ипотечной программе, которая уже не раз доказала свою эффективность?»** является предвзятым, так как в нем уже транслируется вывод о программе, лишая респондента возможности сформулировать свое мнение. Этот вопрос можно задать корректно: **«Насколько эффективна, по вашему мнению, государственная ипотечная программа?».**

ПРЕДЛАГАЙТЕ СБАЛАНСИРОВАННЫЕ ВАРИАНТЫ ОТВЕТОВ

Варианты ответа должны быть обширными, иначе можно получить недостоверные результаты. Для примера вернемся к вопросу об эффективности ипотечной программы **(«Насколько эффективна, по вашему мнению, государственная ипотечная программа?»).**
Варианты ответов «Эффективна/Неэффективна» или «Высокоэффективна/Относительно эффективна/Не знаю» будут не сбалансированы. В первом случае возможности выбора респондента ограничены только двумя опциями, во втором – не предоставляется опция, говорящая о неэффективности программы, а вынуждающая респондентов давать положительный ответ. **В сбалансированном ответе будет дан более широкий диапазон опций для ответа:**

НЕСБАЛАНСИРОВАННЫЕ ОТВЕТЫ:

А) Эффективна
Б) Неэффективна

А) Высокоэффективна
Б) Относительно эффективна
В) Не знаю

СБАЛАНСИРОВАННЫЕ ОТВЕТЫ:

А) Высокоэффективна
Б) Относительно эффективна
В) Скорее не эффективна
Г) Совсем не эффективна
Д) Не знаю/Затрудняюсь

ЕСЛИ ВЫ ПРОВОДИТЕ ОНЛАЙН-ОПРОС, ОТНЕСИТЕСЬ ВНИМАТЕЛЬНО К ОБЯЗАТЕЛЬНЫМ И НЕОБЯЗАТЕЛЬНЫМ ВОПРОСАМ.

Конечно, как исследователи, мы стремимся получить наиболее полную анкету. Однако респонденты могут не знать ответов или не хотеть отвечать на некоторые вопросы. Если будет слишком много обязательных вопросов, это может заставить респондентов говорить неправду или перестать отвечать на наш опросник.

ПРОТЕСТИРУЙТЕ АНКЕТУ

Проведите пилотный опрос. Это поможет выявить возможные ошибки в формулировке вопросов и их логике, а также оценить время, необходимое на заполнение анкеты.

ОБРАЩАЙТЕ ВНИМАНИЕ НА ЛОГИКУ АНКЕТЫ

Анкета, как правило, состоит из трех частей:

— вводная часть (краткое описание анкеты – для кого она предназначена и какая ваша цель);
— основной блок (вопросы для исследования);
—заключительный блок (демографические вопросы).

Вопросы должны логично следовать один за другим, а не «перескакивать» с одной темы на другую – это может запутать респондента.

ИНТЕРВЬЮ

Интервью – один из распространенных методов сбора данных для проведения качественных и смешанных исследований.

Метод интервью заключается в прямом диалоге между исследователем и респондентом с целью получения глубинной информации по определённой теме. Интервью позволяет глубже понять мнения, чувства, мотивации и поведение людей.

В зависимости от степени формализации вопросов и гибкости в ходе беседы, интервью может быть:

Структурированным – когда вопросы заранее определены, задаются в строго установленном порядке, а ответы записываются в виде заранее обозначенных вариантов. Этот вид интервью по своей формализации похож на опрос;

Полуструктурированным – когда список вопросов подготовлен заранее, но порядок и формулировки вопросов могут быть гибкими в зависимости от ответов респондентов;

Неструктурированным – когда диалог с респондентом ведется в форме свободной беседы, вопросы возникают по мере развития диалога.

Вопросы для интервью, как и для опросов, должны быть составлены с учетом определенных правил, которые позволят провести его эффективно.

В первую очередь, определите список тем, которые вы хотите обсудить (эти темы напрямую зависят от целей и гипотез вашего исследования).

Для составления вопросов для структурированного интервью, обратитесь к памятке в разделе **«Опросы».**

Если вы проводите полуструктурированное интервью, то составьте открытые вопросы так, чтобы **они побуждали респондента к развернутым ответам.**

Например, вы проводите интервью среди представительниц молодежных экологических организаций с целью узнать, какие экоинициативы популярны среди молодых людей Казахстана.

Вы можете начать с легкого вопроса о деятельности самой организации: **«Расскажите о своем опыте (об опыте вашей организации) в экоактивизме?»,** продолжив углубляться в тему: **«Как вы считаете, инициативы, которые вы продвигаете, находят отклик среди молодежи?», «А какие находят?» или «Какие не находят?».** Далее вы можете спросить о каких-то конкретных инициативах или проблемах, которые вы исследуете.

Не забывайте, вопросы должны быть:

ПРОСТЫМИ, ИНТЕРЕСНЫМИ И КОНКРЕТНЫМИ

Например: **«Считаете ли вы, что глобальное потепление – это плохо?»** – это слишком общий и неинтересный вопрос (в принципе в мире на него уже получен однозначный ответ). Его можно заменить на: **«Как вы считаете, какие последствия глобального потепления ожидают Казахстан уже в ближайшие пять лет?».**

ДОЛЖНЫ УЧИТЫВАТЬ ОПЫТ, ОБРАЗОВАНИЕИ ИНТЕРЕСЫ РЕСПОНДЕНТА

Не стоит задавать специфичные вопросы людям, чьей сферы интересов они не касаются;

ПРЕДУСМАТРИВАЙТЕ УТОЧНЯЮЩИЕ ВОПРОСЫ

«Можете привести пример?», «Что именно вы имеете в виду под этим?».

ИЗБЕГАЙТЕ НАВОДЯЩИХ ВОПРОСОВ (см. раздел «Опросы и анкеты»).

ИЗБЕГАЙТЕ ВОПРОСОВ С ДВОЙНЫМ ОТРИЦАНИЕМ

(и такое бывает!), которые могут запутать респондента.
Вместо «Разве вы не против изменений?» спросите: **«Как вы относитесь к изменениям?».**

После того, как вы составили вопросы, проведите несколько пробных интервью. Это позволит вам выявить возможные проблемы с вопросами, понять, насколько четко они сформулированы, а также оценить, сколько времени понадобится на проведение интервью.

ВСТАВКА 3. ЭТИКА ИНТЕРВЬЮ

Интервью – вещь чувствительная. Во время интервью люди будут делиться с вами своим личным опытом, профессиональными лайфхаками, собственным мнением, которое может быть сенситивным.

- При выборе места и времени, отдайте право выбора респонденту, но убедитесь, что место располагает к комфортному разговору;
- Получите оформленное согласие респондента (в некоторых случаях респондентов просят подписать письменное согласие);
- Записывайте интервью, только предварительно получив согласие респондента на аудио- или видеозапись;
- Гарантируйте вашим респондентам конфиденциальность и анонимность и всячески обеспечьте их.

Правильная подготовка и структура вопросов помогают получить максимально полезную информацию, обеспечивая успешное проведение интервью и достоверные результаты исследования.

НЕМНОГО ПРО ВЫБОРКУ

Выборка – это часть генеральной совокупности, которая охватывается исследованием, наблюдением, опросом.

Для определения размера и характеристик выборки при проведении исследований существует ряд статистических формул и подходов.

Размер выборки

Методы определения размера выборки – то есть числа респондентов, которых необходимо опросить – бывают разные, но все они в основном зависят от размера генеральной совокупности, уровня значимости и погрешности. Базовыми параметрами для расчета размера выборки являются **уровень значимости в 95%** (что означает, что с вероятностью 95% выборка отражает генеральную совокупность) и **погрешность в 5%** (ошибка, в пределах которой ответы генеральной совокупности могут отличаться от ответов выборки).

Подробнее с формулами расчета и примерами размера выборки можно ознакомиться здесь и здесь. Однако, если для вас эти формулы слишком сложны и непонятны, существует большое количество различных онлайн-калькуляторов расчета размера выборки. Например, если размер вашей генеральной совокупности составляет 1000 человек, то при уровне значимости в 95% и погрешности в 5% вам нужно опросить 278 человек. При этом существует правило, что, начиная с определенного размера генеральной совокупности, выборка все равно будет ее представлять вне зависимости от размера. То есть для генеральной совокупности в 10 тысяч человек достаточно опросить 385, а для генеральной совокупности в 100 тысяч человек – 398.

Обязательно ли, чтобы размер выборки был статистически значимым? Как правило, обеспечить достаточный размер выборки, чтобы генерализировать результаты на уровень страны, можно только в условиях крупных исследований и обладая значительными ресурсами.

В случае, когда вы проводите мини-исследование с ограниченными финансовыми и временными ресурсами, возможности исследователей по обеспечению охвата и репрезентативности выборки значительно сокращаются. Но даже если выборка имеет недостаточный размер, чтобы представлять совокупность, вы все равно можете получить ценные данные.

СБАЛАНСИРОВАННОСТЬ ВЫБОРКИ

Правила минимального размера выборки действуют при предположении, что субъекты генеральной совокупности имеют схожие характеристики. Но в подавляющем большинстве случаев это не так, и помимо обеспечения достаточного объема выборки, нужно обеспечить репрезентативность ее характеристик.

Предположим, вы проводите исследование о возможностях трудоустройства людей с инвалидностью в г. Алматы. Чтобы обеспечить репрезентативность характеристик выборки, нужно понять, сколько в генеральной совокупности женщин и мужчин, какие есть возрастные группы, какой у людей уровень образования, сколько процентов людей пытались искать работу или находятся в ее поиске, или уже имеют работу и т.д.

Сбалансированная выборка будет отражать в себе все эти характеристики.

Вот еще несколько шагов, которые помогут вам в обеспечении репрезентативности выборки:

ПОПРОБУЙТЕ ПРИМЕНИТЬ МЕТОД СЛУЧАЙНОЙ ВЫБОРКИ

Случайная выборка из списка: если у вас есть список потенциальных респондентов, вы можете случайным образом выбирать из этого списка. Это может быть выполнено с помощью генератора случайных чисел.

Стратифицированная выборка: разделите вашу аудиторию на группы (страты) по определённым признакам: например, на возрастные группы, или по критериям поиска работы, или по их географическому положению, а затем проводите случайную выборку внутри каждой группы.

Это поможет гарантировать, что все ключевые группы представлены в выборке.

Используйте случайные номера и случайные генераторы: если вы работаете с большим количеством данных, можно использовать случайные генераторы, чтобы выбрать респондентов из базы данных. Например, вы можете назначить каждому пользователю уникальный номер и затем случайным образом выбрать номера;

ИСПОЛЬЗУЙТЕ ВОЗМОЖНОСТИ ИНТЕРНЕТА И СОЦСЕТЕЙ

Если генеральная совокупность не известна, вы можете проанализировать социальные сети вашей организации и использовать рекламные инструменты этих сетей для таргетирования определенных групп пользователей. Это может помочь вам привлечь нужных респондентов и обеспечить более широкий охват целевой аудитории.

Проводите опросы на разных платформах социальных сетей, чтобы охватить более широкую аудиторию. Например, Facebook, Instagram, X имеют различные пользовательские базы и могут дать разнообразные данные;

ОТСЛЕЖИВАЙТЕ И КОРРЕКТИРУЙТЕ ВЫБОРКУ

Постоянно отслеживайте процесс сбора данных и при необходимости корректируйте методы сбора, чтобы улучшить репрезентативность;

ИСПОЛЬЗУЙТЕ КОРРЕКТИРОВКУ ВЕСОВ

Если ваша выборка не совсем репрезентативна, используйте корректировку весов для учета диспропорций. Это можно сделать, например, увеличив вес данных от недопредставленных групп или уменьшив вес данных от перепредставленных групп.

Предположим, в вашу выборку попали большинство женщин одной возрастной группы, тогда вы можете: либо скорректировать возрастные группы, либо повысить веса для тех групп, которых у вас мало;

УЧИТЫВАЙТЕ ВОЗМОЖНЫЕ ИСКАЖЕНИЯ ПРИ ОПРОСАХ В ИНТЕРНЕТЕ

В интернете могут возникать специфические искажения, например, такие как самоотбор (респонденты сами выбирают участвовать в опросе), поэтому учитывайте это при интерпретации результатов.

В целом, если вы проводите опрос, очень важно изначально прописать в методологии все возможные ограничения и потенциальные искажения, чтобы избежать недостоверных результатов.

ПОЛУЧЕНИЕ ДАННЫХ ИЗ ИНТЕРНЕТА

Открытые данные – это те данные, которые могут быть свободно использованы, распространены и модифицированы, то есть:

— вы можете их скачать в интернете;

— они не защищены авторской лицензией;

— они имеют машиночитаемый формат (формат электронных таблиц, Excel формат, форматы .csv, .tsv, json).

ВСТАВКА 4. ДЛЯ ЧЕГО ВАЖНО ЗНАТЬ ПРО ФОРМАТ?

Данные в немашиночитаемых форматах невозможно сразу обрабатывать и анализировать.

Самым известным немашиночитаемым форматом является формат. **PDF (Portable Document Format).** Для того, чтобы начать работать с данными в .PDF формате, необходимо для начала их конвертировать, например, в формат Excel. Для этого вы можете использовать любые доступные онлайн-конвертеры, в частности ILovePDF, ADOBE, SmallPDF и многие другие.

Кроме того, отсканированные копии тоже можно конвертировать. Для этого применяется технология OCR – Optical Character Recognition. Достаточно написать в гугле «OCR конвертация» и подобрать подходящий для себя инструмент. Один из них – конвертер OnlineOCR.

В эпоху развития открытых данных все больше и больше источников в интернете стараются предоставлять данные в машиночитаемых форматах. Так же поступают и большинство государственных органов – они цифровизируют свои ведомства, публикуют все больше данных, создают отдельные порталы открытых данных. В Таблице 3 даны ссылки на некоторые порталы данных госорганов Казахстана, в Таблице 4 – интернет-ресурсы международных организаций с открытыми данными.

ТАБЛИЦА 3. НЕКОТОРЫЕ ИНТЕРНЕТ-САЙТЫ ГОСУДАРСТВЕННЫХ ОРГАНОВ РК С ДАННЫМИ

НАЗВАНИЕ	ВЛАДЕЛЕЦ	URL	СКАЧИВАНИЕ
НОРМАТИВНО-ПРАВОВЫЕ АКТЫ КАЗАХСТАНА	Министерство юстиции	http://adilet.zan.kz/rus	Да (pdf, doc)
ПОРТАЛ ОТКРЫТЫХ ДАННЫХ КЗ	**Правительство**	**https://data.egov.kz/**	**Да**
НОРМАТИВНО-ПРАВОВЫЕ АКТЫ ПО КАТЕГОРИЯМ, ВЕДОМСТВАМ И ОБЛАСТЯМ	Правительство	https://legalacts.egov.kz/	Да (pdf, doc)
ОТКРЫТЫЙ БЮДЖЕТ	Правительство	https://budget.egov.kz/	Да (xls)
ПОРТАЛ ГОСЗАКУПОК КЗ	Министерство финансов	https://goszakup.gov.kz/?setstyle=normal	Да (api)
БЮРО НАЦИОНАЛЬНОЙ СТАТИСТИКИ АГЕНТСТВА ПО СТРАТЕГИЧЕСКОМУ ПЛАНИРОВАНИЮ И РЕФОРМАМ РК	**Правительство**	https://stat.gov.kz/	**Да**

Источник: база ресурсов собрана контрибьюторами Школы Данных Кыргызстан

ТАБЛИЦА 4. МЕЖДУНАРОДНЫЕ ПОРТАЛЫ ОТКРЫТЫХ ДАННЫХ

НАЗВАНИЕ	ВЛАДЕЛЕЦ	URL
CORRUPTION PERCEPTION INDEX	Transparency International	http://www.transparency.org/
CRIME AND CRIMINAL JUSTICE STATISTICS	UNODC	United Nations Office of Drugs and Crime (UNODC) Crime and Criminal Justice Statistics
INTERNATIONAL BUDGET SURVEY	International Budget Partnership	International Budget Survey
TRADE STATISTICS	UN Comtrade	http://comtrade.un.org/
ENTERPRISE SURVEY	World Bank	http://www.enterprisesurveys.org/
ATTITUDE SURVEYS	Pew Research Centre	http://www.pewglobal.org/category/datasets/
UN DATA	UN	http://data.un.org/
OECD DATABASE	OECD	https://www.oecd.org/statistics/datalab/
GLOBAL HEALTH DATA EXCHANGE	GHDE	http://ghdx.healthdata.org/
UNIVERSAL HUMAN RIGHTS INDEX DATABASE	UNHR	http://uhri.ohchr.org/
CHILDREN'S DATABASE	UNICEF	https://data.unicef.org/
FOREST DATABASE	Global Forest Watch	http://data.globalforestwatch.org/
FOOD AND AGRICULTURE DATABASE	UNFOA	http://www.fao.org/faostat/en/#data/FS
GENDER DATA PORTAL	World Bank	http://datatopics.worldbank.org/gender/
SDG DATA	SDGs Global Dashboard	https://www.sdgsdashboard.org/
HUMAN DEVELOPMENT REPORTS	UNDP	http://hdr.undp.org/en/data
WORLD BANK DATABANK	World Bank	http://databank.worldbank.org/data/home.aspx
UNFPA DATABASE	UNFPA	http://kyrgyzstan.unfpa.org/topics/family-planning
HUMANITARIAN DATA EXCHANGE	UNOCHA	https://data.humdata.org/
GLOBAL COMPETITIVENESS INDEX	World Economic Forum	http://reports.weforum.org/global-competitiveness-report-2014-2015/rankings/
OPEN SANCTIONS	Open Sanctions	http://www.opensanctions.org/
CARBON ATLAS	Global Carbon Project	http://globalcarbonatlas.org/

Источник: база ресурсов собрана контрибьюторами Школы Данных Кыргызстан

Закрытые данные – это данные, доступные узкому кругу пользователей. К ним относятся, например, персональные данные, данные в немашиночитаемых форматах, данные под лицензией или которые нужно покупать, а также данные, получаемые по запросу.

ВСТАВКА 5. ПАРУ СЛОВ О ЗАПРОСЕ

От того, как вы составите запрос на данные, зависят быстрота и полнота получения данных. Несколько подсказок, которые вам помогут составить качественный запрос:

- делайте его максимально детальным;
- пишите четко, какие именно данные нужны;
- за какой период нужны данные;
- на какой территории;
- сразу попросите данные в машиночитаемом виде;
- нарисуйте шаблон таблицы для заполнения;
- сошлитесь на закон о доступе к информации;
- напишите контакты для оперативной обратной связи (например, ваш телефон или email).

ВЕБ-СКРЕЙПИНГ

Веб-скрейпинг – от английского слова «scrape» – «скрести», «соскабливать» – это технология получения данных и информации из веб-страниц. Скрейпинг может выполняться как программистами при помощи кода, так и обычными пользователями с помощью онлайн-инструментов скрейпинга.

Когда нам нужен скрейпинг? Когда на веб-странице есть какие-то таблицы или списки, недоступные для скачивания. В случае, если таких таблиц или списков много, мы можем потратить важное для нас время на копирование таблиц, тогда как при помощи скрейпинга их будет извлечь гораздо быстрее. Кроме того, не все данные и списки легко вставляются в табличный формат при копировании.

Существует большое количество различных онлайн веб-скрейперов. Одними из легких, удобных и понятных являются Data Miner и Instant Data Scraper. Оба эти скрейпера представляют из себя расширения для Google Chrome, и легко устанавливаются на ваш браузер.

В интернете можно найти множество хороших пользовательских инструкций о том, как эти скрейперы работают и в каких случаях их полезно использовать. Для начала можно посмотреть видео от самих разработчиков: по этой ссылке для Data Miner, а по этой ссылке для Instant Data Scraper.

Кроме того, существуют скрейперы для скачивания комментариев из социальных сетей. Иногда для наших исследований мы можем изучать нарративы, которые публикуются в соцсетях, пабликах или на форумах, и анализ комментариев может быть очень показательным.

Один из доступных инструментов скрейпинга комментариев – ExportComments, очень легкий, удобный и не дорогой (да, большинство инструментов в интернете имеют хороший бесплатный функционал, но, если вы захотите расширить его, нужно будет оформить платную подписку).

03.

АНАЛИЗ ДАННЫХ

СТАТИСТИЧЕСКИЙ АНАЛИЗ

Существует огромное число различных методов анализа качественных и количественных данных.

Для анализа данных интервью, фокус-групп используется тематический анализ, контент-анализ, дискурс-анализ или анализ нарративов. Данные виды анализа основаны на глубинном изучении текстов интервью, определении ключевых или повторяющихся тем в них, на изучении частотности, важности и связей.

Для анализа количественных данных используются различные методы, включая статистический анализ, регрессионно-корреляционный анализ, тренд-анализ и анализ описательных статистик.

Регрессионно-корреляционный анализ основан на построении моделей и позволяет выявлять причинно-следственные связи между переменными, тренд-анализ фокусируется на изучении паттернов, изменений в данных в определенный тип времени, анализ описательных статистик позволяет получить информацию о массиве данных, чтобы впоследствии, например, использовать эти данные в машинном анализе.

Статистический анализ — наиболее распространенный и простой вид анализа, который применяется в исследованиях. Он используется для проверки гипотез и основан на выявлении закономерностей, тенденций и соотношений между переменными.

В данном руководстве мы расскажем о формулах, используемых в базовом статистическом анализе, которые позволят вам сделать основные выводы на основе собранных вами данных.

ВСТАВКА 6. ЧИСТЫЕ ТАБЛИЦЫ

Перед тем как анализировать данные, необходимо их подготовить. Убедитесь, что ваши таблицы **«чистые»**: не содержат ненужного текста, пустых и ненужных строк и столбцов, все столбцы имеют заголовки, а формат данных соответствуют самим данным.

Есть также два важных правила, которым мы рекомендуем всегда следовать:

- никогда не работайте в оригинальной таблице. Создавайте копию таблицы и работайте в ней, а оригинальные данные сохраняйте и не трогайте, чтобы при необходимости к ним вернуться;
- создавайте лист с метаданными – данными о ваших данных. На этом листе мы записываем всю важную информацию: названия таблиц, их источники, дату скачивания и ссылки на оригинальные данные.

СРЕДНИЕ ЗНАЧЕНИЯ

Среднее арифметическое – это сумма всех значений выборки, делённая на количество этих значений.

Медиана – это значение, которое находится в середине упорядоченного (отсортированного по убыванию или возрастанию) набора данных. Если количество данных чётное, медиана — это среднее арифметическое двух центральных значений.

Мода – это значение, которое встречается наиболее часто в наборе данных.

Например, в рамках опроса вы собрали данные о доходах ваших респондентов:

РЕСПОНДЕНТ	ДОХОД В МЕСЯЦ, ТЕНГЕ	СРЕДНИЕ	
Респондент 1	100880	Мода – **100880** – чаще всего встречается	Среднее арифметическое – **468004** – сумма всех доходов, разделенная на 10
Респондент 2	100880		
Респондент 3	100880		
Респондент 4	100880		
Респондент 5	273000	Медиана – **461500** – среднее арифметическое середины отсортированного ряда	
Респондент 6	650000		
Респондент 7	658522		
Респондент 8	750000		
Респондент 9	945000		
Респондент 10	1000000		

РАСЧЕТ ДОЛИ (УДЕЛЬНЫЙ ВЕС)

Доля – это часть целого, например, одна из трех или одна третья или 0.33. Доля в процентах – это то же самое, но умноженное на 100%, например, 5%, 25%, 33%.

Доля – это также соотношение части одного показателя к целому числу.
Например, процент респондентов, ответивших «да»; процент казахстанцев, имеющих автомобили; процент женщин, не имеющих доступа к высшему образованию.

Для того, чтобы высчитать долю в процентах, можно использовать принцип пропорции.

Например, в 2021 году всего выявили 5753 человека с онкологией. Из них 2621 — это мужчины. Сколько процентов от всех, кто заболел раком в 2021 году, составили мужчины? А сколько женщины?

Используя пропорцию, получим:

5753 человек = 100%
2621 мужчин = ?%
?% = (2621*100%)/5753

Или логичнее представить эту запись так: 2621/5753*100% = 45,6%

Чтобы высчитать долю женщин, имеющих онкологические заболевания, достаточно из 100% вычесть 45,6%. Получим: 54,4%.

ПОКАЗАТЕЛИ ДИНАМИКИ

Процентное изменение показывает, на сколько процентов изменилось значение переменной по сравнению с её исходным значением.

$$\text{Формула расчета: } = \frac{\text{Текущее значение} — \text{Предыдущее значение}}{\text{Предыдущее значение}}$$

$$100\% = \left(\frac{\text{Текущее значение}}{\text{Предыдущее значение}} — 1 \right) \times 100\%$$

Например, за январь-июнь 2024 года в Казахстане умерло 66.097 человек, а за тот же период 2023 года – 63.601. На сколько процентов людей умерло больше в 2024 году?

$$\text{Процентное изменение} = \frac{66097 — 63061}{63061} \times 100\% = 4{,}8\%$$

По аналогии, можно рассчитывать и сокращение/падение показателей:
Например: за январь-июнь 2024 года в Казахстане родилось 184.673 человек, а за тот же период 2023 года – 189.515. На сколько процентов меньше родилось младенцев в 2024 году?

$$\text{Процентное изменение} = \frac{184673 — 189515}{189515} \times 100\% = — 2{,}6\%$$

В данном разделе мы рассмотрели лишь базовые формулы для анализа данных, которые могут помочь начать работу с результатами исследований. Однако в реальной практике существует гораздо больше разнообразных методов анализа — от простых до сложных, каждый из которых может применяться в зависимости от типа данных, целей исследования и специфики ваших задач. Изучение и освоение этих методов позволит проводить более глубокие и точные исследования.

Подробнее о методах анализа данных можно почитать здесь или пройти курсы на Медиашколе CABAR.asia по анализу данных в гугл-таблицах, методам социальных исследований, статистическому анализу опросов и многому другому.

КАК ДЕЛАТЬ КОРРЕКТНЫЕ ВЫВОДЫ ИЗ ДАННЫХ

Часто, сами того не осознавая, мы допускаем ошибки при интерпретации данных или манипулируем выводами, сделанными на основе анализа этих данных.

Существует несколько важных правил, которым нужно следовать еще на этапе сбора данных, чтобы избежать ошибок впоследствии.

В случае, если вы проводите опрос или интервью, старайтесь следовать правилам, которые мы обсуждали ранее – уделите внимание подготовке корректных и качественных вопросов, попробуйте, насколько это возможно, определить репрезентативную выборку, контролируйте процесс проведения опроса и интервью.

В случае же, если вы пользуетесь вторичными данными – данными из интернета, исследований и статистики – обратите внимание на:

Источник. Убедитесь, что источнику можно доверять. В случае использования официальной статистики, лучше опираться на данные статистических агентств, признанных международных организаций, запросов в государственные органы. При использовании вторичных данных всегда ищите первоисточник. Записывайте все источники в метаданных (см. Вставку 7);

Фиксируйте изменения в данных. В первую очередь, не забывайте, что работать рекомендуется только в копии, не трогая оригинал (см. Вставку 7). Если во время сбора данных вы внесли изменения в методологию, поменяли какие-то данные – обязательно документируйте эти изменения. Если данные в оригинальном источнике поменялись, старайтесь использовать самые последние;

Контролируйте доступ к вашим таблицам. Предоставьте доступ к данным только тем людям, кто непосредственно с ними работает;

Проверяйте данные. Просите коллег проводить перекрестную или двойную проверку. Применяйте контрольные суммы или иные расчеты для валидации ваших данных.

Мы рекомендуем, чтобы эти правила стали неотъемлемой составляющей любого вашего ресерча – назовем это первыми правилами соблюдения «гигиены данных».

После того, как все данные собраны, и вы приступаете к их анализу и интерпретации, важно помнить о потенциальных ограничениях и ошибках.

Давайте рассмотрим несколько распространенных примеров.

ОШИБКА МАЛОГО ОБЪЁМА ВЫБОРКИ

Недостаточный объём данных может привести к неверным выводам, так как результаты могут не быть репрезентативными. Чем больше объём выборки, тем надёжнее результаты анализа.

Как мы уже обсуждали ранее в разделе про типы исследований и выборку, не всегда у нас есть достаточно ресурсов для того, чтобы провести масштабное репрезентативное исследование. В этом случае внимательно

отнеситесь к своим выводам: 1) не утверждайте, что результаты вашего исследования репрезентативны и их можно генерализировать на всю страну; 2) везде указывайте размер выборки, чтобы ваши читатели знали о ваших ограничениях;

ЛОЖНЫЕ КОРРЕЛЯЦИИ

Ложные корреляции – это ситуации, когда два показателя коррелируют между собой, но не имеют причинно-следственной связи.
Например, известная в мире Теория подолов, которая утверждает, что длина юбок коррелирует с направлением фондового рынка. Якобы короткие юбки предвещают рост (бычий рынок), а длинные юбки – спад (медвежий рынок). Это яркий пример ложной корреляции, так как реальные причинно-следственные связи этих двух явления отсутствуют. Многие вещи в мире оказываются просто совпадениями, несмотря на близкие к 100% коэффициенты корреляции.

Правило «correlation does not imply causation» («корреляция не означает причинно-следственную связь») применимо не только в таких достаточно очевидных кейсах, но и в неочевидных.
Например, вы проводите исследование об успеваемости учеников и проверяете гипотезу о том, что доступность книг приводит к более высокой успеваемости. Ваши данные показали, что в семьях, где есть много книг, оценки учеников выше, что в принципе подтверждает вашу гипотезу. Но так ли это?
Влияют ли именно книги на успеваемость учеников или в семьях, где успеваемость детей высокая, больше внимания уделяют образованию и поэтому там много книг?

Для того, чтобы говорить о наличии достоверной «причинно-следственной связи», нужно учесть множество факторов, что зачастую не является тривиальным и требует отдельного научного подхода.

ВЫБОРОЧНАЯ ИНТЕРПРЕТАЦИЯ И CHERRY-PICKING

Выборочная интерпретация – это использование данных, которые подтверждают уже существующую гипотезу, в то время как другие данные, опровергающие её, игнорируются.

«Cherry-picking» (с англ. «сбор вишенок») – выборочное представление данных, ситуация, когда информация подбирается таким образом, чтобы она демонстрировала желательные нам результаты.

Ярким примером cherry-picking являются всякие рекламные ролики, в которых говорят о 100% эффективности того или иного средства или продукта, умалчивая (или приписывая мелким шрифтом) об ограничениях методологии тестирования этих продуктов. Например, что выборка очень маленькая, или что в тестировании лекарства участвовали люди только с легкими симптомами и т.д.

К выборочной интерпретации и cherry-picking можно отнести и заявления выборного штаба президента Кыргызстана Садыра Жапарова о том, что президента «поддержали 80% избирателей». В этом заявлении игнорируются данные о том, сколько избирателей пришло на выборы, то есть явка. На самом деле, явка составила 39,16%, то есть Жапарова поддержали 31% избирателей, а не 80%.

К cherry-picking относится и анализ выборочных периодов во времени. Например, если в последние несколько периодов мы видим быстрый рост какого-то показателя, при этом 10 лет назад он был еще быстрее, но мы этот факт игнорируем. Всегда подходите внимательно к выбору временного периода для вашего анализа. Почему вы берете только 5 лет? Или почему 10 лет? На эти вопросы должна давать ответ ваша методология;

НЕПРАВИЛЬНОЕ ИСПОЛЬЗОВАНИЕ СРЕДНИХ

Использование среднего значения без учёта разброса данных может привести к искажённым выводам. Среднее арифметическое не всегда отражает реальную картину, если данные имеют сильные отклонения.

Самый простой пример – разброс в данных о доходах. Предположим, в группе из 10 респондентов семь человек имеют доход в 100 тысяч тенге, два человека – в 1 миллион и 1 человек – в 10 миллионов тенге. Если мы посчитаем среднее арифметическое дохода в этой группе, то оно составит 1 млн 270 тысяч, что более чем в 12 раз выше дохода большинства респондентов в группе. В этом случае использование среднего арифметического является некорректным, и лучше применять моду или медиану (см. раздел «Анализ данных»);

НЕВЕРНАЯ ИНТЕРПРЕТАЦИЯ ПРОЦЕНТНЫХ ИЗМЕНЕНИЙ

Некоторые интерпретации процентов могут привести к недоразумениям. Например, рост с 1 до 2 – это 100%, а с 100 до 101 – только 1%, хотя абсолютное изменение в обоих случаях очень мало.

Кроме того, когда мы считаем процентные доли на малых выборках, иногда могут возникнуть ложные представления о размерах выборки. Если вы опросили пять человек, четыре из которых ответили утвердительно, то использование процентов и игнорирование размера выборки могут ввести в заблуждение.

В случае маленьких выборок данные лучше представлять в абсолютных показателях, а не в относительных.

НЕКОРРЕКТНЫЙ ПРИМЕР ПРЕДСТАВЛЕНИЯ ДАННЫХ:

80% респондентов ответили, что они сталкивались с коррупцией в госучреждениях

КОРРЕКТНЫЙ ПРИМЕР ПРЕДСТАВЛЕНИЯ ДАННЫХ:

Четыре из пяти опрошенных респондентов столкнулись с коррупцией в госучреждениях

СРАВНЕНИЕ НЕСОПОСТАВИМЫХ ПОКАЗАТЕЛЕЙ ИЛИ НЕСОПОСТАВИМЫХ ПЕРИОДОВ

Проблема сравнения «яблок с апельсинами» легко проецируется на проблемы интерпретации данных.

Когда мы сравниваем разные регионы, страны, выборки и пр., важно учитывать численность популяции.

Например, сравним между собой число молодежи в городах Алматы и Астана. По данным Бюро национальной статистики Казахстана, численность молодого населения в Алматы в начале 2024 года составила 695,6 тысяч человек, в Астане – 455,7 тысяч человек.

Где молодежи больше? Если мы говорим об абсолютных показателях, то в Алматы живет больше молодого населения. Но если мы хотим сопоставить, в каком из городов, скажем, выше представленность молодого населения, то лучше рассчитать процентную долю молодого населения среди всей численности населения городов. Используя опять же данные Бюро нацстатистики, получим, что:

% доля молодых людей в Алматы = (695 625/ 2 228 677)*100 = **31.2%**

% доля молодых людей в Астане = (455 712/1 430 117)*100 = **31.9%**

Получим, что доля молодежи в обоих городах примерно одинаковая.

Еще одним примером, когда важно учитывать численность населения, являются такие показатели, как уровень преступности, заболеваемости, доступность той или иной инфраструктуры, ВВП – их можно рассчитывать на 100 тысяч, 10 тысяч, 1 тысячу или на душу населения.

Самый простой пример – сравнение стран по ВВП, когда для того, чтобы корректно сравнить богатство этих стран, мы используем показатели ВВП на душу населения. Например, по данным Всемирного Банка, ВВП Казахстана в 2023 году составил 261 млрд долларов США, а ВВП Люксембурга – 86 млрд долларов США, что в три раза ниже, чем казахский ВВП. Однако, если мы посмотрим на показатели ВВП на душу населения, то увидим, что в Люксембурге он гораздо выше – 128 тысяч долларов на душу населения против 13,1 тысячи долларов в Казахстане.

Этим же принципом можно воспользоваться, чтобы сравнить доступность больничных организаций в регионах Казахстана.

Бюро нацстатистики Казахстана публикует данные по числу больничных организаций по регионам и годам, а также численность населения регионов Казахстана в динамике. И в Алматинской, и в Костанайской областях число больничных организаций примерно одинаково – 44 и 42. Но где доступность этих организаций выше? Для этого нам нужно рассчитать число больниц на 100 тысяч населения каждой области. Для этого воспользуемся следующей формулой:

$$\text{Число больниц на 100 000 населения} = \frac{\text{Число больниц}}{\text{Численность населения}} * 100000$$

(Эту формулу можно использовать для нормализации любых показателей – просто меняйте ее числитель и знаменатель. Вы также можете поменять множитель, если вы хотите рассчитать показатель на 1000 населения или на 1 млн населения).

Итак:

$$\text{Доступность больничных организаций в Алматинской обл.} = \frac{44}{1531167} * 100000 = \mathbf{2,9}$$

$$\text{Доступность больничных организаций в Костанайской обл.} = \frac{42}{029984} * 100000 = \mathbf{5,1}$$

За счет того, что в Костанайской области численность населения ниже, на каждые 100 тыс. ее жителей приходится больше больниц.

Кроме того, нельзя сравнивать несопоставимые временные периоды – показатели целого года и 6 месяцев другого года. В случае, если вы проводите динамический анализ – например, сравниваете как изменились инфраструктурные условия в школах за несколько лет, важно обеспечить одинаковую выборку в эти годы – вы не можете сравнивать в динамике разные школы.

ИГНОРИРОВАНИЕ МАСШТАБА ПРОБЛЕМЫ

При проведении исследований очень важно учитывать масштаб проблемы. Он может влиять на то, как вы интерпретируете данные и какие решения предлагаете. Если проблема затрагивает небольшую группу людей или ограниченный временной период, ее решение может быть проще и быстрее.

Но если масштабы проблемы большие или она носит системный характер, может потребоваться более комплексный подход и длительный анализ.

Например, данные говорят о том, что мужчины тоже становятся жертвами домашнего насилия. При том, что это правда, масштаб у этого явления в разы ниже, чем у насилия в отношении женщин. В среднем 95% пострадавших от семейного насилия – это женщины и только 5% мужчины, кроме того, только в 3% случаев женщины это насилие совершают.

Не умаляя эмпатии к пострадавшим мужчинам, нужно помнить, что системность и масштабность проблемы находится именно среди насилия в отношении женщин, и именно эта проблема нуждается в оперативном решении.

Методы, которые мы рассмотрели, помогут избежать манипуляций с данными и некорректных выводов, помогая вам проводить объективный анализ. Однако ключевым фактором остаётся подход исследователей и исследовательниц к проблеме с самого начала.

Важно правильно формулировать вопросы, грамотно собирать данные и тщательно выбирать методы анализа. Такой осознанный подход на всех этапах исследования гарантирует вам достоверные результаты и обоснованные выводы.

ОБ АВТОРКЕ

САВИЯ ХАСАНОВА – ДАТА-ЖУРНАЛИСТКА, ИССЛЕДОВАТЕЛЬНИЦА, МЕНТОРКА ИЗ КЫРГЫЗСТАНА. ИМЕЕТ ДИПЛОМ КРСУ ПО СПЕЦИАЛЬНОСТИ «МАТЕМАТИЧЕСКИЕ МЕТОДЫ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ В ЭКОНОМИКЕ», А ТАКЖЕ СТЕПЕНЬ МАГИСТРА ПО ЭКОНОМИКЕ УНИВЕРСИТЕТА ГУМБОЛЬДТА В БЕРЛИНЕ.

С 2008 года принимала участие в различных исследовательских проектах и написании аналитических отчетов для международных организаций в странах Центральной Азии, включая ПРООН, ООН-Женщины, ООН ЭСКАТО и др. С 2017 года работает в сфере аналитической и дата-журналистики. Являлась редакторкой аналитического портала «Central Asian Analytical Network» при Университета Джорджа Вашингтона. Вместе с коллегами из ОФ «Школа данных Кыргызстан» проводила тренинги по анализу и визуализации данных в Центральной Азии и Монголии, имеет несколько авторских курсов по анализу данных, в том числе для IWPR. В 2021 году совместно с Анной Капушенко была удостоена мировой премии по дата-журналистике «The Sigma Awards» за исследование о фемициде в Кыргызстане. В настоящее время является менторкой программы по дата-журналистике Internews in Kyrgyzstan, программы исследований ООН-Женщины по фемициду, редакторкой правозащитных исследований.

Финансирование Европейского Союза

ИНМИР

INSTITUTE FOR WAR & PEACE REPORTING
IWPR
ИНСТИТУТ РЕПОРТАЖЕЙ ВОЙНЫ И МИРА

Эта публикация финансируется Европейским Союзом. Её содержание является исключительной ответственностью IWPR и не обязательно отражает точку зрения Европейского Союза.

© Все права сохраняются за IWPR CA. Материал предназначен для личного изучения или использования в образовательных целях, некоммерческих продуктах или услугах при условии, что IWPR CA указан как источник и правообладатель.