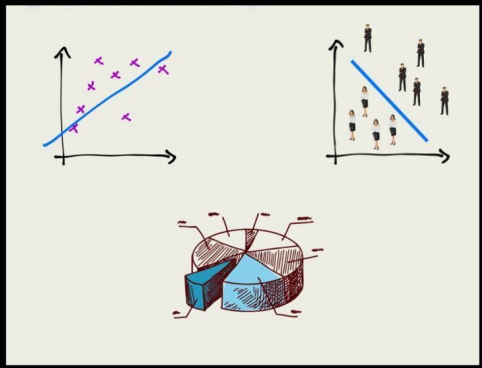
Никита Сергеев

Аналитика и Data Science

Для не-аналитиков и даже 100% гуманитариев...



Никита Сергеев

Аналитика и Data Science. Для не-аналитиков и даже 100% гуманитариев…

http://www.litres.ru/pages/biblio_book/?art=43114130
ISBN 9785005007346

Аннотация

Когда люди не инженерных специальностей слышат «аналитика и Data Science», то представляют разное. Кто-то видит таблицы и графики. Кто-то неподъемно сложные математические формулы. Кто-то программирование и искусственный интеллект…Но истоки этих понятий из области статистики, которая делится на описательную и аналитическую.И эта кажущаяся непостижимой аналитика – на самом деле нескучная, интересная и простая вещь. Чтобы ею пользоваться, не нужно ни изучение сложных формул, ни программирования…

Содержание

Аналитика и Data Science Для не-аналитиков и даже 100% гуманитариев…

Никита Сергеев

ISBN 978-5-0050-0734-6
Создано в интеллектуальной издательской системе Ridero

Большинство людей услышав о статистическом анализе представляют или технарей, или ученых, изучающих закономерности. Но статистика применяется далеко за пределами научных лабораторий: в рекламе, маркетинге, бизнесе, менеджменте, политике, образовании и т. д. А базовые знания анализа данных крайне полезны и в повседневной жизни.

И неважно какую должность Вы занимаете и какой род Вашей деятельности: в современном мире в любой профессии вероятность столкнуться с цифрами, большими объемами данных и поиском в них закономерностей с каждым днем стремительно мчится к 1 (или к 100%).

> Эта книга – это Ваша возможность попробовать «на вкус и ощупь» кажущийся таким невообразимо сложным и непонятным мир цифр и статистического анализа.

Книга вводит профессионалов из не-технических наук (менеджеры, гуманитарии, психологи, социологи, культурологи, экономисты, политтехнологи и т.д.) в захватывающий цифровой мир статистики и вероятностей – и поможет легко в нем ориентироваться, пользоваться и не бояться.

Она написана от простого к сложному так, что способствует погружению в аналитику и Data Science (наука о дан-

ных) совсем не «технически-инженерных» людей. Людей, казалось бы, совершенно далеких от этой очень прикладной дисциплины: менеджеров, гуманитариев и профессионалов социально-экономических дисциплин.

Невзирая на то, что сама книга о прикладной дисциплине и написана экспертом по анализу данных, владеющим всеми статистическими программами (от обычного Excel до SPSS) и языком программирования R, **– она совершенно не техническая**.

Книга НЕ содержит языка программирования R или Python.

НЕ пестрит запросами к базам данных.

В ней НЕТ теории вероятностей, невообразимых математических формул и матстатистики…

> Книга о прикладных практических инструментах, которые любой человек сможет сразу же использовать на рабочем месте, в жизни, в своих собственных исследованиях…

После ее изучения Вы сможете применять современные методы статистического анализа на практике, а также будете легко находить и видеть скрытые закономерности среди любых объемов данных, строить предсказывающие (прогностические) модели, добывать из цифр знания и информацию для принятия решений.

Автор книги – управленческий консультант с 20-летним опытом ведения консалтинговых проектов для крупней-

ших компаний по всему миру. Ведущий русскоязычный инструктор по инструментам ведения бизнеса и менеджменту на международной платформе UDEMY

https://www.udemy.com/user/nikita-sergeev-2/

В основу книги положены самые современные материалы, которые использовались в разных проектах (трансформация бизнес- и операционных моделей, M&A, реинжиниринг процессов, оптимизация численности, маркетинговые и социологические исследования, исследования персонала, разработка психодиагностического инструментария и тестов, анализ и прогнозирование и т.д.) и читались на специализированных MBA программах. В том числе, и в специализированных русскоязычных курсах по аналитике на международной платформе он-лайн образования UDEMY.

> Книга будет полезна любому, кто хочет научиться работать с данными – будь Вы жаждущий освоить статистику новичок или профессионал, желающий систематизировать знания или «освежить память».

Информация в книге в основном ориентирована на социально-экономические дисциплины, но рассматриваемые в ней методы анализа являются универсальными и подходят для компьютерных наук, промышленности, оценки качества, прогнозирования рисков, медицины, физики, химии, фармакологии, биомедицины, биотехнологий, генетики и т. д.

От автора

Почему я решил написать эту книгу? Наверное, по той же причине, по которой помимо основного рода деятельности и образования веду, казалось бы, довольно далекие от них курсы и мастер-классы по анализу данных как для сотрудников и менеджмента корпораций, так и в открытом доступе на международной образовательной платформе UDEMY для всех желающих.

> Современный мир, общество и компании – это данные, данные и данные. И их объемы на сегодня настолько обширны, что понять в них закономерности и строить прогнозы невооруженным глазом совершенно невозможно.

Я уже более 20 лет работаю с широким кругом менеджеров и профессионалов из разных стран, отраслей и организаций. И почему-то подавляющим большинством принято считать, что анализ данных – это нечто сакрально сложное и доступное только математикам, ИТшникам и инженерам. А менеджерам, гуманитариям и профессионалам социально-экономических наук это знание непостижимо.

Но это миф. Свой профессиональный путь я начинал именно с анализа данных будучи еще студентом-психологом – анализировал результаты социологических и маркетинговых исследований для международных компаний,

а также помогал академикам, кандидатам и докторам различных наук готовить практические части их диссертаций.

Я отчетливо помню, как в 90-х молодыми студентами мы все со страхом шли на первую лекцию страшнейшего для психологов предмета – «Математические методы в психологии». Но по факту предмет оказался совершенно несложным, а также поистине захватывающим и увлекательным.

С того времени уже много воды утекло… Я прослужил в вооруженных силах (помотался по ПВО, ВВС и ядерным войскам). Отработал в бизнесе на должностях старшего и высшего менеджмента от менеджера по маркетингу и оргразвитию до управляющего партнера по стратегии, слияниям и поглощениям. Сопровождал десятки одних из самых крупных в СНГ трансформационных проектов и реорганизаций. Обзавелся женой и 4 детьми. Набрал лишние 30 кило…. А также нашел то, что меня увлекает помимо научных исследований и инвестиций в области биотехнологий и медицины – я стал управленческим консультантом и занимаюсь трансформационными проектами для крупных корпораций.

Надеюсь, эта книга увлечет Вас анализом цифр и данных, выглядящих для многих не-технических профессионалов такими скучными, пресными, сложными и непонятными…

Я хочу, чтобы каждый читатель уловил: статистика и аналитика пронизывают как компании любого размера (будь то крупная транснациональная корпорация, небольшая фирма

или стартап), так и практически любую современную область знаний. С каждым днем все сложнее становится провести границу между любой современной профобластью (от биологии и медицины до управления организациями и персоналом) и аналитикой. А все социально-экономические исследования практически неотделимы от сравнений выборок, корреляционного, факторного и регрессионного анализа.

> Поэтому чем бы Вы ни планировали заниматься – вероятность необходимости использования статистики и анализа данных в современном мире с каждым днем становится все ближе и ближе к 1 или 100%.

Анализ данных у всех на слуху и на сегодня это один из самых востребованных навыков в любых сферах. Однако, как я наблюдаю, зачастую работа с данными не вызывает восторга ни у студентов, ни у сотрудников нетехнических специальностей, ни у менеджмента. Но в этой книге Вы увидите, что на самом деле аналитика и поиск закономерностей в данных – очень занимательная штука (да и не такая уж и сложная).

Начнется книга с довольно широкого и немного философского контекста – вначале я вкратце расскажу важность моделей исследуемых объектов для правильного построения гипотез, анализа и объяснения результатов. Также остановлюсь на разграничении того, что является, а что не является аналитикой. И пройдусь по основным понятиям статистики.

Далее мы с Вами сфокусируемся на анализе данных и по-

иске в них скрытых закономерностей. Мы рассмотрим те методы, которые Вы после каждой главы сможете сразу же применять в работе. Этому, по сути, и будет посвящена основная часть книги.

А поскольку сейчас понятие Data Science (наука о данных) и анализ данных плавно вплетены в такую область как машинное обучение (Machine Learning – ML) и искусственный интеллект (Artificial Intelligence – AI) – то напоследок я расскажу и обо всем этом новоязе.

В основной части книге я отобрал современные наиболее ходовые в социально-экономических направлениях методы анализа данных. К ним привел конкретные примеры использования в моей практике. Но, помимо этого, написал немного о подготовке массивов к анализу, а также об основных функциях Excel, которыми покрываются 90% бизнес-задач.

Оговорюсь, что написать об Excel – это скорее вынужденная мера. Просто часто после курсов и тренингов менеджеры и специалисты не-технических дисциплин задают мне вопросы как решить ту или иную «аналитическую» задачу – а большинство этих «аналитических» задач решается условно 5 основными функциональностями Excel.

> Книгу я старался написать так, чтобы любой читатель, независимо от уровня подготовки в части аналитики, и уловил основные концепции, и освоил прикладные методы.

Каждый раздел книги структурирован таким образом,

чтобы Вы не только ориентировались в методах, а и легко соотносили их с решаемыми аналитическими задачами. В книге в практическом русле рассматриваются те методы и инструментарий, которые покрывают львиную долю аналитических бизнес-задач и которыми Вы самостоятельно сможете пользоваться в работе.

Но тем, кто хочет всерьез освоить тему, а не просто прочесть «еще одну умную книгу», настоятельно рекомендую **сразу же после каждого раздела отрабатывать все методы на практике**. Для этого у Вас под рукой будет Excel и программа PSPP (распространяется в открытом доступе официальная статистическая программа). А также массивы данных (считай таблички и выгрузки с данными в Excel) из Вашей профессиональной деятельности – отрабатывайте методы сразу прямо на них. Ну и эта книга сожержит инструкции по работе как с Excel, так и с PSPP для каждого метода – так что по сути является одновременно и самоучителем.

О, подумал кто-то, обещали простоту – а только начали читать, и уже появилась какая-то страшная аббревиатура … PSPP… Многие пугаются, что надо будет изучать дополнительное программное обеспечение – «Давай Excel, он есть у всех!».

Да, можно реализовывать всю аналитику и в офисном приложении Excel. Но, боюсь, после этого Вы возненавидите аналитику (а аналитика – это не таблички-диаграммы или

дашборды со средними и %: мы об этом еще отдельно поговорим). Особенно после того, как будете 99% времени тратить на написание скриптов и формул в Excel, которые никто кроме Вас неспособен будет прочесть. Или от безысходности найдете выход в покупке недешевых специальных надстроек к Excel.

PSPP не страшнее Excel (даже на порядок проще). А кроме того, эта программа аналогична такому коммерческому IBM'овскому программному продукту как SPSS, который широко используется аналитиками крупных корпораций и международных исследовательских агентств. Научившись работать в PSPP – Вы считай умеете работать и в SPSS. А это очень ценный прикладной навык для не-технических профессий.

Возможно, после прочтения книги кто-то захочет послушать лекции и посмотреть как аналитика работает «вживую» для решения разных задач (от маркетинга и сегментации клиентов до вопросов управления персоналом), а также выполнить практические упражнения на «живых» примерах. Приходите на он-лайн курс «Аналитика и Data Science для менеджеров и гуманитариев» на крупнейшей образовательной платформе UDEMY:

https://www.udemy.com/analytics-and-data-science/?couponCode=BOOK_READER

Даже если Вы просто взяли полистать эту книгу любопытства ради, но аналитика, невзирая на все доводы, пока совер-

шенно не из области Вашего интереса – то книга все-равно попала в Ваши руки не зря. Наверняка у Вас есть знакомые, которым книга станет полезной – поделитесь с ними информацией о ней.

ОКОЛО-АНАЛИТИЧЕСКИЕ РАЗГОВОРЫ

Бизнес-жаргон: статистика, метрики, $Dashbord_s$, KPI_s… и аналитика

Для не-технических специалистов аналитика – понятие обычно обширное и часто включающее то, что является «совсем не очень аналитикой». Дам небольшое разъяснение понятий (по крайней мере, как их следует трактовать исходя из предмета данной книги).

Хочу внести ясность, поскольку время от времени наблюдаю как нахватавшиеся фраз сотрудники компаний путают одно с другим и часто, имея ввиду одно, говорят совершенно о другом. Хотелось бы дополнительно расставить точки над «Ё» в части одинакового понимания и ожиданий читателей того, что они найдут (или не найдут) в этой книге.

Сначала пройдемся по четырем моментам, которые в бизнесе порою жестко ассоциированы с аналитикой. Но таковой они не являются. Они все отражены на ***рис. 1***.

Рис. 1. Важные вещи: но это – не аналитика…

В бизнесе слово **статистика** используется повсеместно. Часто можно услышать при постановке задачи сотруднику от руководителя – «Приготовь статистику». Речь в таком случае идет не о науке, а о том, чтобы приготовить какие-то отчеты с определённым набором **количественных данных** за период.

Объем продаж, количество клиентов, численность предприятия, число визитов на сайт, количество лайков в соцсети…. Т.е., это любые **данные**, накопленные за период времени.

Еще одно избитое в менеджменте слово **метрики**. Это определенные показатели, которые являются производны-

ми от данных. Обычно их получают простыми формулами путем вывода %, суммирования, отнимания, деления или умножения одного статистического показателя на другой. Но иногда бывают более сложные формулы. Метрики уже могут отражать эффективность процессов, активностей, управления, предприятия и т. д.

Например, «3 основные бизнес-метрики нашего стартапа», или «наши HR-метрики показывают неэффективное использование бюджета на персонал». Примерами метрик могут служить такие показатели как конверсия, HR ROI, отток / текучесть клиентов или персонала, % лайков от просмотров, количество ошибок на 1000 транзакций и т. д.

Метрика позволяет отвечать на вопросы «хорошо или плохо», «эффективно или неэффективно».

Дашборд (Dashboard) – это дословно панель приборов, т.е. интерфейсное представление или форма, в которую выводится набор метрик или данных, важных для отслеживания хода операционной деятельности или эффективности бизнеса.

Сюда отбираются те метрики и данные главного процесса (value chain), изменение которых требует вмешательства и принятия управленческих решений.

KPIs (Key Performance Indicators) – они же ключевые показатели эффективности. Все хотят, чтобы они были коли-

чественными в виде метрик или «статистик». Но на практике часто используют и качественные. Каждый количественный KPIs – по сути метрика. Но не каждая метрика является KPI. Т.е., в KPIs попадают только именно ключевые для определённого периода (обычно года) метрики или данные.

Аналитика – это слово во многих организациях используют, зачастую подразумевая данные за период или метрики.

Но **аналитика – это совсем другого рода вещь.** Это поиск скрытых закономерностей и построения прогностических (предсказывающих, предиктивных) алгоритмов посредством конкретного набора аналитических инструментов. Аналитика проверяет модели на прочность или позволяет находить новые модели исследуемых объектов или процессов.

В книге мы **не будем говорить о метриках**. Кто решил ее прочесть с ожиданием разобраться как правильно подобрать метрики под компанию, процесс, продукт, систему… – Вам не сюда.

И в книге мы вообще никаким образом **не будем касаться ни KPIs, ни построения Dashboard-ов**. Потому что эти вопросы вообще к анализу данных и аналитике не имеют отношения. Это чистой воды вопросы систем управления.

В общем, если даже прочитав аннотацию и предыдущие разделы Вы все еще надеетесь узнать в книге как подбирать

эффективные метрики, формировать KPIs и дашборды для компании, функции, процесса или продукта – оставьте Вашу надежду, ибо в этих вопросах данная книга никак не поможет.

В части **данных** – мы обзорно коснемся формирования правильных массивов данных, с которыми можно «по-человечески» работать. Но перечислять какие данные обычно собираются для тех или иных направлений (продажи, маркетинг, производство, HR, социология и т.д.), для чего их использовать и в каких расчетах применять, как организовать хранилища данных – эти вопросы также не из тематики книги.

Книга также почти **не касается вопросов визуализации данных** (хотя даже эту тему многие считают аналитикой) – это вопросы обработки и представления данных / информации, но не аналитики.

А вот, собственно говоря, **аналитике**, **набору современных инструментов для поиска скрытых закономерностей и прогностического анализа** и будет посвящена книга.

Книга поможет тем, кто хочет, к примеру, научиться с определенной долей вероятности отвечать на такие вопросы:

· Будет ли соискатель эффективен на должности продавца?

· Как долго будет клиент пользоваться услугами компании?

· Кто из клиентов в ближайшее время перестанет пользоваться услугами?

· Насколько понизится мотивация персонала при снижении удовлетворенности возможностями карьерного роста?

· Что повлияло на выбор того или иного кандидата в президенты?

· Вернет ли потенциальный заемщик кредит?

· И т. д.

Глава с двумя оговорками для высшего менеджмента

В этом разделе речь все о том же, что не входит в предмет данной книги, но сквозь «другие очки» – «***вид сверху» глазами высшего руководства компании***.

Этот раздел в дополнение к предыдущему написан специально для представителей высшего менеджмента («злые языки» говорят, что для отпугивания нежелающих делать своими руками).

Книга не покрывает такие вопросы менеджмента как:

· устройство и построение корпоративных систем аналитики (построение аналитических функций в компаниях)

· оценка уровня зрелости аналитической функции компании

УСТРОЙСТВО И ПОСТРОЕНИЕ КОРПОРАТИВНЫХ СИСТЕМ АНАЛИТИКИ (ПОСТРОЕНИЕ АНАЛИТИЧЕСКИХ ФУНКЦИЙ В КОМПАНИЯХ).

Многие компании путают аналитику с тем, как внедрить и управлять аналитической функцией по всему предприятию. Путать корпоративную систему аналитики с непосредственно аналитикой – то же самое, что путать *корпоративную систему управления проектами* с *непосредственным управлением проектом*.

Корпоративная аналитическая система – это и корпоративная методология, и аналитические спецподразделения (офисы), и процессы, и оборудование с программным обеспечением и т. д. И тема эта вообще из области проектирования организаций, а не аналитических методов и инструментария.

Но в рамках данной книги будут наборы методов прогностической аналитики и поиск инсайтов с применением простых описательных статистик. Это то, что **отдельно взятый человек может своими руками использовать на своем рабочем месте или в жизни**. Эти методы могут внедряться в корпоративных системах аналитики как отдельные компоненты, но они никак **не заменитель всей системы или ее элементов**.

В общем, **книга не о корпоративных системах аналитики**.

УРОВЕНЬ ЗРЕЛОСТИ АНАЛИТИЧЕСКОЙ ФУНКЦИИ КОМПАНИИ.

В бизнес-структурах аналитикой, как я упоминал в предыдущей главе, называют все что угодно: от просто данных и до KPIs с Dashboard'ами. И «ноги растут» от того же понимания ***уровня развития/зрелости аналитических функций в организациях,*** который **не предмет данной книги**.

Об уровнях зрелости упомяну только здесь и один раз.

Когда я анализирую **уровень зрелости аналитической функции в компании**, то базируюсь на используемых уровных PWC (Price Waterhouse Coopers):

Уровни зрелости аналитической функции

Это на самом деле достаточно общий подход, но PWC активно с ним работают, потому приписываю его им.

Здесь первый уровень – *уровень данных* – обозначает способность предприятия извлекать данные и иметь отчеты с констатацией и описанием того «**что есть на сегодня и уже случилось**». Здесь вовсю фигурируют всем из-

вестные отчеты с накопленными данными за периоды (в них не особо заморачиваясь могут также накладывать линейные линии трендов).

Два следующих – *метрики с отчетами и диагностика* (сюда же относятся дашборды и бенчмарки) – обозначают, что компания может осуществить диагностику и понять **«почему случилось и насколько все плохо\хорошо».** Эти два уровня, кстати, в более ранних версиях были объединены в один уровень. Вот здесь уже вовсю работают *описательные статистики*, в том числе процентили, квартили, моды, медианы, средние и т. д. В книге мы рассмотрим методы описательной статистики, которые читатель сможет использовать, но не будем рассматривать как их визуализировать, строить дашборды или «нарезать» KPIs.

Следующий уровень – *инсайты* – это не отдельные методы, а способность организации собирать данные из разных систем и источников в *едином информационном поле.* По сути, наличие корпоративного хранилища данных, из которого можно извлекать данные и используя все те же *описательные статистики* обнаруживать находки/инсайты не всегда видны в рамках одной системы с данными одной направленности. В книге я покажу как с использованием прикладных функций Excel соединить данные из разных источников, а также приведу менеджмент-кейсы с инсайтами при использовании простых описательных статистик. Но в книге не будет о том, как отстроить этот уровень зрелости в организа-

ции.

И последний уровень – прогностическая *аналитика* – это способность компании строить предиктивные (предсказательные) модели, базирующиеся на скрытых закономерностях и неочевидных взаимосвязях во всех имеющихся у нее данных. Это уже применение новомодных систем искусственного интеллекта (AI). В данной книге будут изложены методы аналитической статистики (корреляции, регрессии, факторный и кластерный анализ и т.д.), которые прочитавший профессионал сможет сразу использовать в своей работе. Но здесь не будет о том, как и с помощью каких систем вывести компанию на такой уровень зрелости.

Но в последнее время многие консультанты говорят, что **есть еще один некий уровень для организации**, который интересует именно высшее руководство компаний – *прескриптивная аналитика* (еще Вы могли слышать на конференциях или от консультантов «нормативная» или «предписательная» аналитика).

Чем интересен ТОР'ам этот уровень и чем же он отличается от тех уровней, на которых работает описательная статистика и прогоностическая аналитика? Если описательная статистика отвечает на вопрос «что было?», а прогностическая аналитика «что будет?» – то прескриптивная аналитика пытается ответить на вопрос «а что кому и где делать?» + «к чему приведут те или иные действия?».

Но, в отличие от описательной и аналитической статистики, прескриптивная аналитика – это *не отдельная область знаний*, со своей методологией, специфическими методами или понятиями. Это смесь прогностических методов (базируется на них), автоматизации процессов, бизнес-правил и автоматизированных управленческих предписаний к исполнению.

Прескриптивная аналитика: рассматривать ли как уровень?

Т.е, это скорее попытка автоматизации управленческих решений и воздействий. Повторю: *прескриптивная ана-*

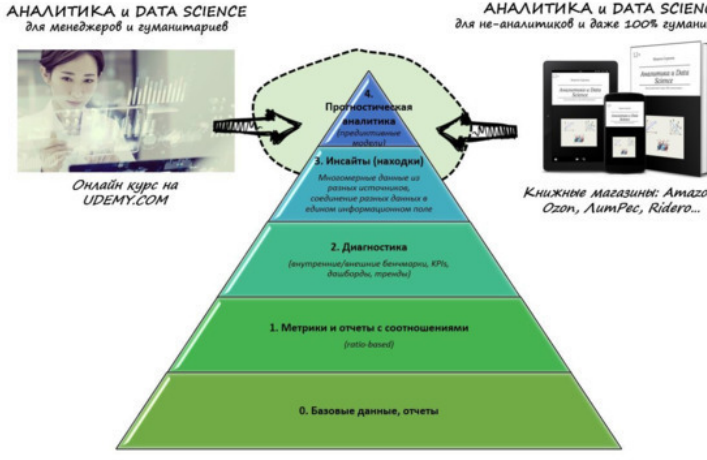
литика – это «смесь» из использования методов прогностической аналитики, математических бизнес-моделей, бизнес-правил, алгоритмов, автоматизированных процессов и управленческих решений и т.д., чтобы оценить возможные будущие исходы (последствия) действий компании. Это искусство конкретной компании использовать вышеперечисленное для моделирования возможных вариантов будущего и автоматического принятия управленческих решений и воздействий.

Но я персонально не расцениваю этот уровень как часть уровня зрелости аналитической функции. Не потому, что тут нет отдельного предмета, методологии, методов и т. д. Ведь на уровне «Инсайтов» их также нет. Но уровень инсайтов/находок базируется на описательной статистике, со своим предметом, задачами, методологией и методами – т.е., все еще лежит в границах аналитической дисциплины. А на уровне прескриптивной аналитики переплетается и автоматизация, и системы управления, и собственно аналитическая функция. Т.е., это более широкая и мультифункциональная область.

Ну и еще мне на сегодня прескриптивная аналитика выглядит (пока что) созданной консультантами «упаковкой под продажу» аналитических систем в крупные корпорации.

Оговорки сказал. А если подытожить предмет книги, то данная книга (как и одноименный онлайн курс

на UDEMY) – это то, что сфокусировано на **методах поиска инсайтов и прогностической аналитики**, но **не сборник рассказов о том, как «подтягивать» уровень зрелости аналитических функций компаний**.



Книга о поиска инсайтов и методах прогностической аналитики

Но в любом случае, если Вы хотите разобраться в методах и попробовать как аналитика работает «вживую» для решения бизнес-задач независимо от уровня Вашей должности – данная книга безусловно будет Вам полезна.

Особенности социально-экономической реальности

В последнее время везде пишут о том, как важно нести гуманитарные и социально-экономические знания (бизнес, коммуникации, менеджмент, предпринимательство и т.д.) в технические направления.

Мне, наряду с необходимостью нести «гуманитарно-социально-экономический свет» инженерам-технарям, не менее важным видится нести технические навыки гуманитариям. Чтобы последние могли более системно принимать решения и опираться в своих концепциях на более твердый фундамент, а не собственные размышления и суждения, подкрепленные только навыками убеждения и лидерско-харизматическими приемами.

Отдельная интересная тема для русской науки и ее масштабирования в век капитализма – это «нести» навыки бизнеса и менеджмента непосредственно в научную среду. Неимоверное количество знаний и открытий умирают в стенах НИИ только потому, что их создатели ограничиваются в лучшем случае разговорами с такими же учеными-экспертами или публикацией в журнале, который читают такие же ученые-эксперты.

Одни не считают нужным (да и ниже их уровня) популяризировать свои открытия. Другие может и хотели бы

добиться практического использования продукта их труда (знаний и открытий), но понятия не имеют какими методами и как этим управлять в эпоху капитализма. Но на этой теме я останавливаться в книге не буду.

К социально-экономическим наукам относятся науки, которые оперируют не естественными физическими законами и закономерностями (гравитация, время, пространство, масса, рост, вес, скорость света, давление и т.д.), а такими вещами как восприятие, поведение, мнения, отношения, качества, установки и все порождаемые ими социально-экономические явления.

> Любая организация, общество, рынок… – это в первую очередь социально-экономические системы. Для анализа данных в этих системах используются те же методы, что и в технических науках, но есть несколько главных особенностей, которые необходимо помнить.

Аналитика в социально-экономических науках (в противовес с естественно-инженерными) сталкивается с пятью главными особенностями – ***рис. 2***.

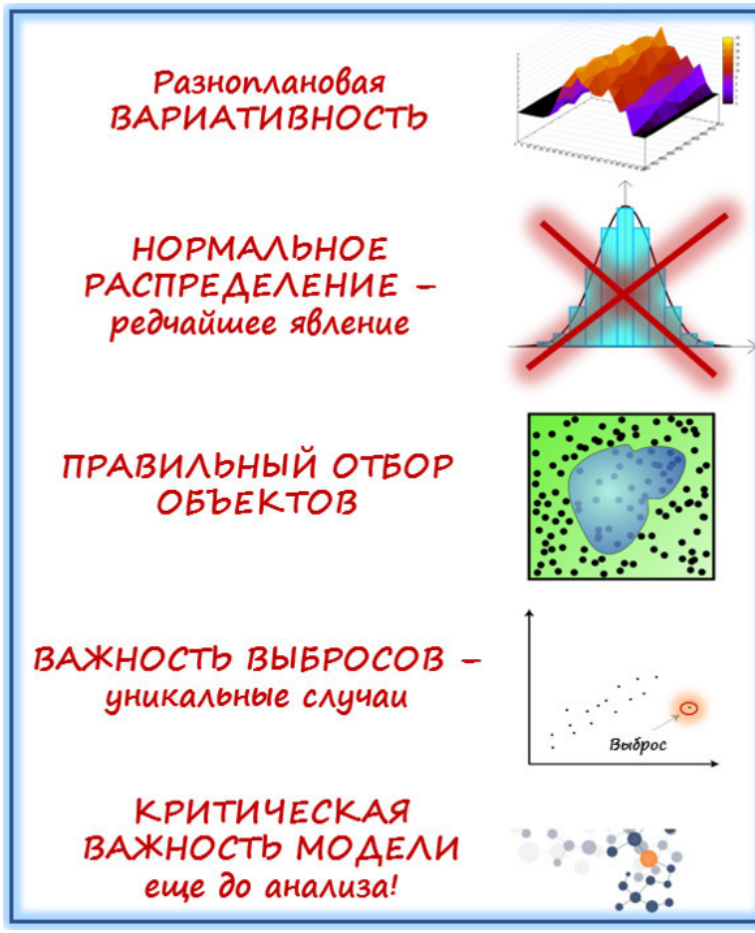


Рис. 2. Особенности аналитики в социально-экономической реальности

Теперь разберем этот рисунок.

Во-первых, социально-экономическая система – это очень изменчивая система.

Скорость падения яблока прогнозируема – сколько и где бы Вы это не повторяли. А деньги, трафик, усилия для результата или популярность (то, что изучается в социально-экономических системах) – совершенно нет.

Т.е., если переменные имеют физические ограничения, препятствующие большому разбросу или смещению размеров – и вероятность случая, кардинально отличающегося от основной массы, крайне низка: это одно. Но измерьте, например, корреляции на фондовом рынке за разные периоды – и коэффициенты будут резко меняться от периода к периоду.

А я часто встречаю, как гуманитарии выдают обнаруженные в социально-экономической реальности корреляции как некие реальные «материальные» зависимости (еще и позиционируют эти статистические взаимосвязи как причинно-следственные). Но вот что-то никто ни разу не предсказал по ним поведение фондового рынка…

Или возьмите компанию – измерьте удовлетворенность персонала, внедрите программу улучшений (даже сделайте что-то небольшое) – и у Вас эффект! Но через год Вы заметите как удовлетворенность сползает вниз… Что повлияло? Почему? Новые люди пришли? Старые привыкли?

Во-вторых, здесь не работает закон нормального

распределения.

В социально-экономических дисциплинах закон нормального распределения – это непозволительная роскошь. Но многим менеджерам и гуманитариям он почему-то кем-то крепко «вбит в головы»…

Если мерять рост или вес – да, будет работать закон нормального распределения. Но в социально-экономических системах чаще всего наоборот – мы не будем наблюдать красивую симметрию нормальной кривой. Скорее будет обратная картинка: смещение в одну или в другую сторону.

Так, в конкретно взятой стране 2% людей могут владеть 60—90% капитала.

На любом рынке есть несколько игроков, занимающих 60—90% доли рынка.

Несколько рок-исполнителей или авторов книг забирают на себя 90% популярности и продаж.

Из 100 кандидатов в президенты 5% заберут 95% голосов. И т. д.

Да та же удовлетворенность сотрудников работой в компании будет давать смещение или в одну, или во вторую сторону – и в придачу влиять на другие аспекты работы (это так проявляется способность удовлетворенности, как базовой эмоции, к генерализации).

В-третьих, важность выборки случаев / объектов / наблюдений для применения их ко всей популяции (вся популяция объектов называется «генеральная совокупность»),

которую Вы исследуете.

Измерив какие-то физические величины в одном месте, Вы скорее всего получите ± те же самые в другом – ну или с минимальной вариативностью.

Но измерив, например, отношение к кандидату в президенты или расовым вопросам в регионе, Вы точно не получите их ± такими же в другом. Или, замерив удовлетворенность работой в одной компании, Вы не получите тот же результат в другой компании.

И, **в-четвертых, важно понимать, что одно-единственное социально-экономическое явление может перевернуть все Ваши представления и закономерности вверх дном.** В естественно-технических системах каждый один уникальный случай не ведет к глобальным изменениям.

И **пятое – наличие модели для анализа в социально-экономических дисциплинах критически важно.**

Модель (Ваше представление, набор предположений об исследуемом объекте) должна предшествовать анализу (кроме случаев, когда у Вас поисковый анализ, цель которого изобрести новые или уточнить существующие модели – но в бизнесе таким вряд ли Вы будете заниматься).

Только по модели Вы можете описать, измерить и прогнозировать поведение / развитие какого-то события или объекта. О важности моделей поговорим отдельно в следующей главе.

Модель

Раздел обязателен к прочтению, даже тем, кому он кажется философским и далеким от аналитики.

Под моделью не имеются ввиду статистические алгоритмы и методы обработки данных.

Словом «модель» обозначается некое представление исследуемого объекта, процесса, явления.

> **Модель** – это набор увязанных между собой предположений и понятий, выстраивающий определенный взгляд на объективную реальность.

На ***рис. 3*** изображены несколько наиболее известных моделей – Солнечная система, ДНК, молекула…

Рис. 3. Несколько наиболее известных моделей

Например, элементы ДНК – пары нуклеотидов имеют

4 компонента АТГЦ (аденин, тимин, гуанин и цитозин), которые имеют взаимосвязь А с Т и Г с Ц.

Конечно же, модель строится на основании ограниченного множества известных нам данных (элементов, компонентов, свойств и взаимосвязей) об оригинале (реальном объекте объективной реальности).

Самим оригиналом (объектом объективной реальности) модель не является и на объективную реальность (окружающий мир, явление, протекающие процессы и т.д.) она никоим образом не влияет.

Зато она влияет на наше понимание и отношение к этой реальности.

Только модель любого объекта позволяет нам:

· формально его описать

· делать измерения и интерпретацию полученных результатов

· спрогнозировать его поведение / развитие в будущем

· а также понять его историю в прошлом.

Кроме того, модель позволяет постоянно обучаться, уточнять и добавлять взаимосвязи между ее элементами и компонентами – и, возможно даже, накопленные знания со временем изменят само наше представление о модели. Схематически это все изображено на ***рис. 4***.

Рис. 4. Динамика взаимосвязей модели и реальности

Вспомните, как развивались представления (модели) о Земле по мере накопления знаний и установления новых взаимосвязей: от плоскости на китах и черепахах до Земли-центра и до того, что она крутится вокруг Солнца (***рис. 5***).

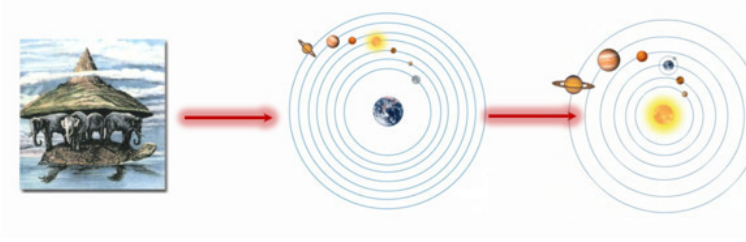


Рис. 5. Изменение представлений о модели Земли по мере накопления данных и знаний

С моей т.з. наличие некой общей модели особенно важно для социальных, экономических и бизнес-дисциплин, где представление о реальности (модель) на порядок важнее чем для той же биологии, геологии, физики, астрономии и т.д., базирующихся на фундаментальных естественных законах.

А люди часто брезгуют моделями, считая их уделом ученых-теоретиков, отдавая предпочтение инструментам / методам… Но эффективность применения инструмента крайне зависит от того, для чего и применительно к какой реальности (объекту, событию, процессу и т.д.) мы его используем.

Я сам не раз наблюдал как менеджеры, профессионалы и даже ученые использовали аналитический инструментарий для прогнозов, но без понимания модели результаты этих попыток предсказаний были аналогичны гаданию на картах Таро.

Даже если рассматривать бизнес и организацию, кото-

рые являются социально-экономическими системами. Любой бизнес, любая организация внутри себя также может быть представлена простой операционной моделью как набором элементов и компонентов со взаимосвязями (на ***рис. 6*** авторское представление).

Рис. 6. Базовое представление операционной модели

предприятия

Если посмотреть шире (***рис. 7***) – то организация является открытой системой и неразрывно связана с внутренней и внешней средой.

Если посмотреть еще шире, детализируя окружение компании: клиенты, конкуренты, продукт, процессы, структура, культура и сотрудники компании, ее поставщики и вся экономика – все это уже элементы большой бизнес-модели.

Соответственно на базе моделей аналитику можно очень успешно применять в бизнесе для принятия более взвешенных бизнес-решений, особенно в условиях неопределенности.

Рис. 7. Связь операционной модели с внешней и внутренней средой

Модель – одна из важнейших вещей в аналитике. Именно модель исследуемого объекта / явления / процесса позволяет правильно осуществить анализ: от того какие данные собирать и до того как правильно интерпретировать полученные данные.

Интуиция или аналитика?

Среди людей есть те, кто верит цифрам, а есть те, кто полагается на «чуйку» и интуицию. И это также выражено в бизнесе и менеджменте.

Многие полагают, что достаточно только чутья, бизнес-интуиции и имеющегося опыта – и приводят в пример ряд успешных проектов или решений, принятых вопреки статистике, исследованиям и аналитике.

Например, некоторые приводят Генри Форда, который когда-то сказал, что если бы он полагался на исследование мнений клиентов, то ему бы пришлось заниматься выведением более быстрых пород лошадей, а не автомобилями.

Лукавят, потому что с одной стороны речь тут о технологии, а с другой стороны Г. Форд на самом деле никогда не брезговал аналитикой в управлении предприятием.

Более того, только аналитика позволяет накапливать знания, наращивать и объяснять опыт, усиливать практическую интуицию, а в самом идеальном варианте – возвести к пониманию неких концептуальных моделей.

Я говорю об интуиции и опыте в связке, потому что для меня интуиция – не что иное как «свернутый опыт» человека. Например, говорят, что опытный механик «по звуку машины» может определить проблемы. На самом деле он улавливает ряд мельчайших моментов (данных) в работе авто,

но просто уже делает их интерпретацию на таком уровне автоматизма, что не способен объяснить на что именно он обращал внимание, когда поставил «точный диагноз».

> Дискуссия о том, что важнее – опыт / интуиция или аналитика несостоятельна в принципе. Вообще ИЛИ здесь неуместно – более целесообразно использовать **И**.

Ведь сама по себе ни статистическая информация, ни ее анализ, ни обнаруженные статистические значимые взаимосвязи действительно не дают автоматических ответов на вопросы – поэтому модель, интуиция, размышления и воображение (творческий подход) имеют очень большое значение.

Схематически дополняемость аналитики и опыта друг-другом можно представить так (***рис. 8***):

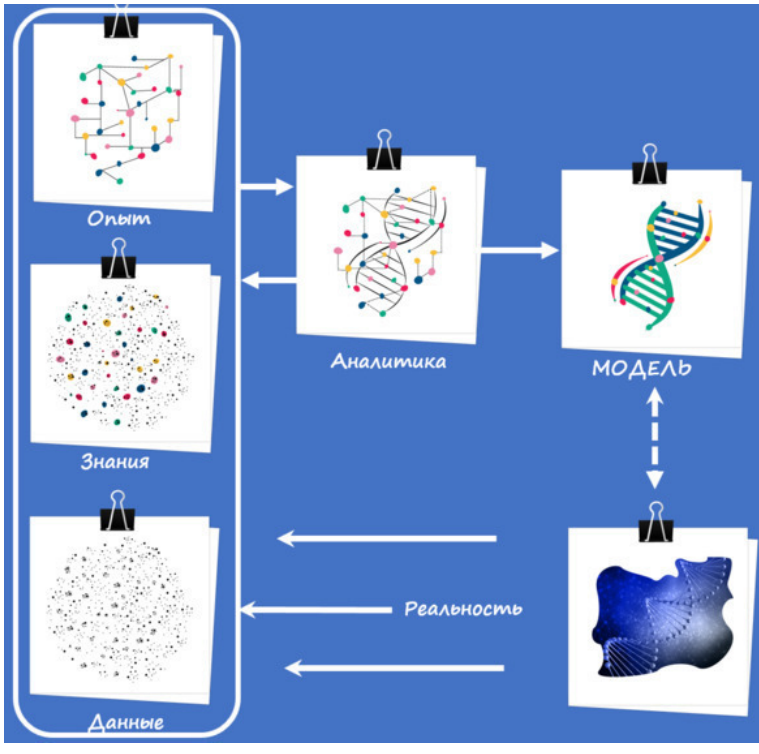


Рис. 8. Дополняемость опыта и интуиции аналитикой

Немного объяснений к картинке. Сначала мы снимаем / регистрируем / собираем / получаем из реальности некие **данные** (причем данные в широком смысле слова и в любом виде).

Далее данные превращаются в **знания**, которые потом

объединяются какими-то связями (вот это событие произошло потому, что было вот то-то и то-то) на основании нашего взаимодействия с реальностью. Знаниями и опытом мы уже можем делиться с другими.

Аналитика может нам помочь уточнить наши взаимосвязи: как опровергнуть их наличие в реальности, так и обрисовать скрытые взаимосвязи, которых мы сами не замечали. Это формирует более целостную картину.

В итоге при взаимодействии данных, знаний, опыта и аналитической проверки у нас может родиться некое концептуальное представление реальности (какого-то объекта, процесса, явления, случая и т.д.) – **модель**.

Это не сама реальность – это только ее модель, наше представление о ней. Но на базе этой модели мы уже можем более эффективно обмениваться пониманием реальности с другими людьми, а также постоянно его уточнять, приращивая новые знания и устраняя пробелы.

Есть еще, конечно, **креативная отсебятина** (кстати, очень часто встречаемая в менеджменте, социально-экономических и гуманитарных направлениях). Когда человек что-то увидел, чего-то нахватался – и из этого породил в голове какую-то ерунду и, уверовав в нее, обозвал некой моделью (***рис. 9***).

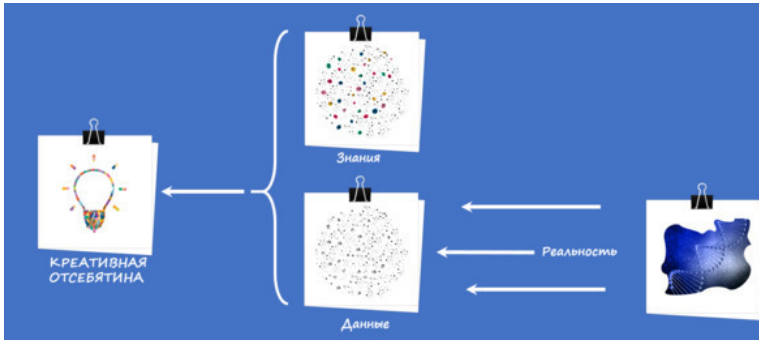


Рис. 9. Модели без опыта и аналитики зачастую имеют очень отдаленные связи с реальностью

Иногда, конечно, бывает, что из такого креатива рождаются ± верные модели. Но они все равно проверяются только опытом, аналитикой и самой реальностью.

Какая лучшая программа для анализа данных?

Существует ряд программ для анализа данных. От всем уже привычного Excel, до коммерческих продуктов типа SPSS, Statistica, OCA и вплоть до отдельного языка программирования R, созданного специально под аналитику. Есть и бесплатные аналоги дорогостоящего коммерческого программного обеспечения – например, программа PSPP как аналог SPSS.

В интернете есть ряд официальных инструкций, курсов, книг и самоучителей по той или иной аналитической программной среде (какие кнопки нажимать, где находится та или иная функция, где смотреть вывод результатов и т.д.).

> Но главное – понимать, что все эти программы **не заменители «головы» аналитика**.

Это всего лишь инструментарий. Но, невзирая вроде на эту понятную истину, постоянно разворачиваются баталии на тему «какая программа лучше». Всегда хочется спросить о критерии «лучшести» – ведь каждая программа имеет свои плюсы и минусы, возможности и ограничения.

> Решение об использовании той или иной программной среды – это на самом деле исключительно вопрос профессиональных и личных предпочтений.

Я, например, в своей практике использую несколько инструментов: подавляющая часть того, что я делаю, сделана в SPSS, OCA и Excel.

SPSS и OCA – поскольку привык ими пользоваться. Excel – потому, что удобен для бизнеса и его может открыть, просмотреть и отследить логику формул любой бизнес-пользователь.

Для некоторых задач использую R. Но с языков программирования я бы не рекомендовал начинать не-техническим профессионалам. Это дольше, сложнее, да и вряд ли Вы в своей работе столкнетесь с настолько емкими задачами, чтобы не решить их более простым способом.

Потому, что использовать – больше будет зависеть от того, что Вы решите и осилите освоить. Однозначно в бизнесе (за исключением, если Вы профессиональный аналитик и это Ваша ежедневная работа) самым ходовым инструментом является Excel. Бизнес – это клеточки Excel.

Потому и в данной книге вначале будет показана реализация описательных статистик в Excel, чтобы Вы могли применять эти навыки в знакомом офисном приложении. Но по мере усложнения методов и уровня аналитики мы перейдем на PSPP (аналог-заменитель SPSS).

При обучении прикладному инструментарию для нас с Вами критерием «лучшести» является простота и привычность. Чтобы читатели тратили время не на изучение программы, а фокусировались на сути решаемых задач.

И мой выбор для начинающих и не-инженерных профессий – однозначно Excel и PSPP. Но не просто читайте разделы и главы, а после прочтения сходу отрабатывайте методы в этих программах на Ваших массивах.

Упоминая Excel, не хочу сформировать неправильные ожидания к книге, потому сделаю ударение: в книге не будет обучения базовым навыкам работы с Excel. Изложение книги предполагает, что читатель уже на минимальном базовом уровне знаком с Excel.

Очень краткие итоги раздела

Что я хотел, чтобы читатель вынес из раздела:

1. Никогда не ставьте ИЛИ между аналитикой и интуицией. Всегда И. Не умаляйте роль творчества и случайностей.

2. Пять особенностей социально-экономической реальности:

· Изменчивость

· Редкость нормального распределения

· Репрезентативность выборки

· Пристальное внимание к выбивающимся из общего массива случаям / объектам / наблюдениям

· Важность модели

3. Модель должна предшествовать анализу, чтобы иметь возможность объяснить и проинтерпретировать данные.

4. Разницу между данными, метриками, КПД, дашбордами и собственно аналитикой как поиском скрытых закономерностей и построения прогнозов посредством специального набора инструментов.

5. Неважно какой программный продукт / инструмент Вы используете – используйте то, что знаете. Программы / инструменты дополняют и повышают эффективность, но не заменяют человека.

ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЙ АНАЛИЗ

О статистическом анализе

Нас повсюду окружают данные. В соцсетях, в магазинах, рекламе, метро… даже в авиалайнере. Весь мир – это цифры.

Нам может казаться, что собирая данные (при чем все больше и больше), мы контролируем большое количество важных вещей и держим ситуацию под контролем.

Но на самом деле важно уметь отбирать именно те данные, которые помогают понять ситуацию и принять решения, даже располагая неполной информацией. Какие именно данные важны помогает понять модель, о которой мы уже говорили.

С данными помогает работать такая наука как статистика. Именно она позволяет придать понятный вид и смысл огроменным массивам данных, состоящим даже из миллиардов или триллионов значений.

> Статистика делится на описательную и аналитическую. Мы в книге рассмотрим оба эти ответвления.

Задача описательной статистики только описать объ-

ект, процесс, явление – используя среднее значение, % распределения, количество и т. д.

Аналитическая статистика использует более сложные методы, которые позволяют рассчитать взаимосвязи между переменными, а также понять, являются ли эти взаимосвязи просто случайными совпадениями или реальными закономерностями.

Анализ данных является ключевым этапом, в ходе которого происходит непосредственная проверка соответствия собранной информации нашим моделям явлений, процессов или объектов.

И более того: в ходе анализа формулируются и проверяются / уточняются существующие или рождаются новые модели, отражающие те закономерности, которые мы нашли в собранных данных.

Исследователь, ученый, менеджер или работник выдвигает определенную модель явления / процесса / объекта, демонстрирует соответствие (либо противоречие) данных и содержащихся в них закономерностей этой модели – и только потом может опираться на модель, отвлекаясь уже от самих данных. Нам, к примеру, уже не нужно постоянно опираться на данные, чтобы понимать, что Земля вращается вокруг Солнца.

> Именно статистический анализ позволяет нам находить скрытые закономерности, которые дают нам больше понимания о реальности и уточняют как она

работает.

Но, прежде чем искать закономерности, надо рассмотреть несколько важных вещей из области статистики – и мы их далее рассмотрим в рамках этого раздела.

Выборка и генеральная совокупность

Реальность обычно представлена невероятно большим количеством случаев / наблюдений / объектов. Людей, жителей, клиентов, компаний, растений или животных и т. д. И вся их популяция представляет собой **генеральную совокупность**.

Например, если объектом нашего интереса (за кем мы желаем понаблюдать и изучить) являются жители конкретного города, то все они и есть наша генеральная совокупность. Но если объектом интереса были бы, к примеру, только люди трудоспособного возраста (или имеющие право голоса на выборах) в этом городе, то наша генеральная совокупность уменьшилась бы.

При решении отдельных задач вполне легко можно исследовать всю генеральную совокупность.

Например, у Вас есть текущая база подписчиков он-лайн журнала – и необходимо предсказать кто из них с высокой долей вероятности не продлит подписку со следующего года.

Для этого у Вас, по сути, есть доступ к базе данных по всей генеральной совокупности – и Вы можете сделать аналитику, используя данные всей базы. Посмотреть, люди с каким профилем демографии, поведения, предпочитаемых рубрик чтения и т. д. не продлевали подписку в прошлом и, наложив обнаруженные закономерности на существующую базу,

получить условно доверительный прогноз кто не продлит ее сейчас.

Также с генеральной совокупностью могут иметь дело специалисты кадровых служб, проводящие анализ сотрудников предприятия.

Другое дело, когда Вы решите изучить всех потенциальных клиентов, рынок кандидатов на вакансии или избирателей. Вот тут Вы столкнетесь с тем, что всех их изучить невозможно и дорого. Поэтому Вы будете исследовать только некоторых, а полученные результаты распространять на всю генеральную совокупность.

> Вот те некоторые выбранные из **генеральной совокупности** объекты / образцы / люди / события и будут называться **выборкой**.

Но с выборкой не все так просто. Основная сложность в формировании выборки – это понимание того, какие именно объекты / образцы в нее включить так, чтобы иметь полную картину. Ведь она должна быть **репрезентативной** – т.е., **полученные по ней результаты должны с высокой долей точности отражать генеральную совокупность.**

Иллюстративно генеральная совокупность, выборка и вопрос ее репрезентативности изображены на ***рис. 10***.

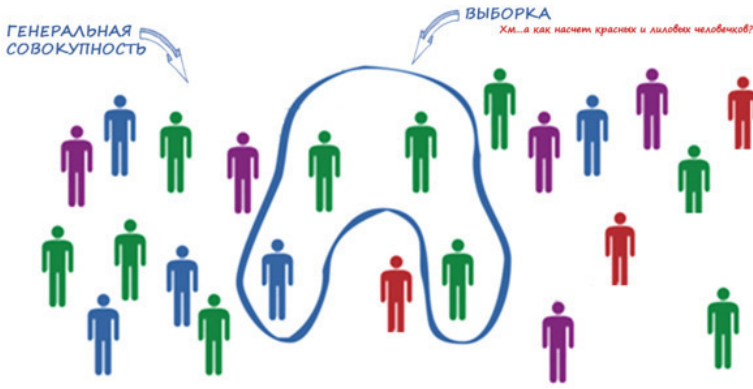


Рис. 10. Генеральная совокупность, выборка и вопрос ее репрезентативности

Неужели это настолько важно – какая будет выборка? Приведу такой пример (надеюсь, не обижу чувства верующих). Например, Вы выберете всех, кто участвовал в военных действиях. Эти люди выжили – и Вы обнаружите статистически значимую зависимость с молитвой перед боем. Вы будете впечатлены – неужели молитва реально помогает выжить? Можно ли заявить об этом?

Нет, нельзя. Во-первых, возможно Вы просто путаете причину и следствие (статистические взаимосвязи не означают причинно-следственные связи, о которой мы поговорим позже) – просто во время боевых и критических для жизни моментов люди начинают чаще молиться и надеяться на высшие силы. Поэтому правильная интерпретация – это

опять же вопрос модели (элементов и их взаимосвязей) объекта / явления / процесса, который Вы исследуете.

А во-вторых, есть главная проблема в Вашем исследовании – Вы не знаете, сколько также молились, но погибли. Потому что не можете их опросить – они мертвы. Т.е., Вы отобрали нерепрезентативную выборку: она не представляет собой генеральную совокупность.

Для того чтобы выборка отражала генеральную совокупность, чаще всего используют три основных подхода:

1. Случайный: когда объекты для изучения отбираются из генеральной совокупности случайным образом.

2. Стратифицированный: когда генеральную совокупность разбивают на группы (страты) по важным для модели признакам (например, пол, возраст, отрасли, поведение, использование продукта с определенной частотой, частота посещения церкви и т.д.). Объём (%) каждой группы задает то количество объектов / наблюдений, которые надо отобрать из каждой группы. Получаются квоты на отбор тех или иных объектов.

3. Серийный: когда изымают партию товара, выбирают людей, проживающих в многоквартирном доме на конкретной улице, или берут целиком отдельные отделы в компании и т. д.

Соответсвенно, генеральная совокупность и выборка связаны напрямую: чтобы отобрать репрезентативную выборку, главное иметь правильное представление о всей генеральной

совокупности.

А как рассчитать, сколько же объектов / случаев / наблюдений из генеральной совокупности необходимо включить в выборку?

Для этого есть специальная формула расчета (*спокойствие: книга, как и обещано, без формул*), которая для расчета размера выборки использует «размер генеральной совокупности», «допустимую вероятность» и «доверительный интервал»:

· **Размер генеральной совокупности** – это количество **всех** объектов / наблюдений / случаев в генеральной совокупности.

· **Доверительная вероятность** – это считайте показателем точности / достоверности (о сути вероятности как таковой мы поговорим чуть позже). В практике обычно принимается 95%. Можно брать, конечно, значение выше или ниже – например, от 85% до 99,9%. Но тогда число объектов / случаев / наблюдений в выборке будет уменьшаться или увеличиваться соответственно.

· **Доверительный интервал** – это допускаемый Вами диапазон реальных значений при применении полученных на выборке результатов к генеральной совокупности. Задается в % и говорит о том, насколько ± % (в каком «коридоре») может отличаться истинное значение в генеральной совокупности от полученного в выборке. Например, если то-

варом по какому-то параметру в выборке клиентов довольны только половина (50%), то при доверительном интервале ±5% с вероятностью 95% истинное значение будет лежать в диапазоне от 47,5% до 52,5% (это ±5% от полученных в выборке 50% довольных).

Для сравнения: например, мы хотим узнать мнение 100.000 клиентов (генеральная совокупность).

Если нас устроит 95% вероятность с ±5% доверительным интервалом – то достаточно опросить 383 клиента.

Если Вас устроит ±10% – то хватит мнений всего 96 клиентов.

Ну а если нам «кровь из носа» необходима почти 100% вероятность (например, 99,7%) и чтоб почти без интервала (скажем, ±0,1%) – то готовьтесь исследовать почти всех клиентов, а именно 95.745.

> Стандартно для социально-экономической реальности достаточно надежным считается использовать вероятность 95% и доверительный интервал ±5%.

По большому счету, **чем выше Вы укажете вероятность и меньший доверительный интервал – тем больше объектов из генеральной совокупности должно попасть в выборку**.

Сколько объектов брать в выборку – решать Вам ис-

ходя из допускаемых Вами погрешностей (все равно 100% достичь не получится) и экономичности (сколько затрат готовы понести на извлечение данных из выборки).

Сама **формула расчета размера / объема выборки по большому счету Вам не нужна**, так как расчет выборки автоматизирован как в спецпрограммах, так и в ряде онлайн калькуляторов.

> **Онлайн калькуляторы** можно найти через *любой поисковик в интернет* (задайте запрос **«онлайн калькулятор выборки»**).

В калькулятор останется внести размер генсовокупности, а также устраивающую Вас вероятность и доверительный интервал – **и калькулятор рассчитает сколько образцов (объектов / наблюдений / случаев) Вам необходимо исследовать в генеральной совокупности** .

А ЕСЛИ ВЫ РАБОТАЛИ СО ВСЕЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТЬЮ И ОТ НЕЕ ПОЛУЧИЛИ ТОЛЬКО НЕКИЙ % ОТВЕТОВ?

Такое часто бывает, к примеру, когда Вы опрашиваете всех сотрудников предприятия. Вы раздали анкеты всем, а получили только некое количество ответов.

Тогда задача сводится к обратному: зная размер генсовокупности и выборки – рассчитать доверительный интервал, чтобы понимать, насколько полученные в выборке данные ± могут отличаться в генеральной совокупности.

Имея *размер генеральной совокупности* и *количество заполненных анкет* (выборка) можно рассчитать доверительный интервал (те наши ±%) для того или иного % ответов определенной категории.

Например, если в компании 5.000 сотрудников, а сдали анкеты только 3.250, то при доверительной вероятности 95% доверительный интервал будет ±1,02%. Считается это также в онлайн калькуляторах. Пример, как выглядят такие онлайн калькуляторы в сети Интернет на ***рис.10.1***:

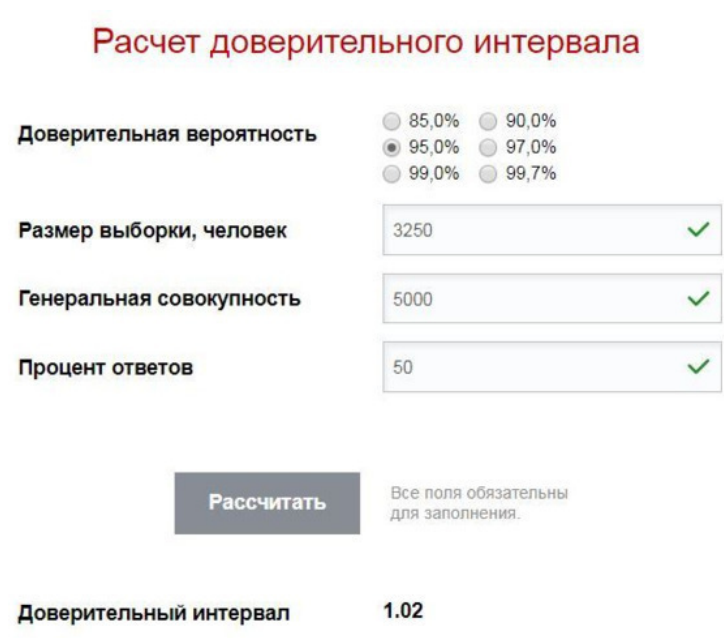


Рис.10.1. Пример онлайн калькулятора расчета доверительного интервала

Но в расчете доверительного интервала есть один нюанс по поводу **поля «Процент ответов».**

Внимательно читаем: рассчитанный доверительный интервал будет справедливым для альтернатив ответов сотрудников, которые набрали 50%. *Для альтернатив, которые набрали другие % – доверительный интервал будет другим.*

Например, Вы задали сотрудникам такой компании вопрос «Довольны ли Вы стилем менеджмента в компании?» с тремя вариантами ответа и такими % ответов персонала:

· Доволен – 50%

· Насколько доволен, настолько недоволен – 15%

· Недоволен – 35%

В данном случае, доверительный интервал (или ошибка выборки) будет ±1,02% будет справедлива только для «довольных» – т.е. доля довольных будет в диапазоне 50±1,02% (от 48,98 до 51,02).

Но для средней альтернативы доверительный интервал (или ошибка выборки) будет ±0,73%.

А для «недовольных» ±0,97%.

Т.е, подставляя в поле «Процент ответов» разные значения альтернатив в зависимости от % отметивших их сотрудников, мы будем получать разные значения доверительного

интервала для альтернатив.

На практике, если в целом *ошибка выборки (значения доверительно интервала) Вас устраивает в целом для «Процент ответов» 50,* то далее просто смотрят полученные % ответов.

Переменные

Данные обычно состоят из большого количества отдельных показателей, которые называют переменными. Это, например, доход, количество клиентов, город или страна, отдел, род войск, зарплата, пол, частота курения, количество посещений или часов порносайтов, частота занятия сексом в неделю, количество детей, социальный статус и т. д.

> Переменная имеет свое значение для того или иного объекта /случая / наблюдения.

По большому счету переменная – это характеристика объекта / случая / наблюдения. Например, цвет глаз у каждого человека будет свой.

Т.о., каждый случай, объект или наблюдение имеют свои характеристики, т.е., имеет свое значение той или иной переменной. Переменные описывают объект.

Например, на ***рис. 11*** в качестве примера приведены Валя и Иван – это **объекты / случаи / наблюдения**.

Рис. 11. Объекты и переменные

А их рост, цвет глаз, доход, место проживания, частота путешествий и другие характеристики – это **переменные**.

Например,

· Валя -женщина, Иван – мужчина.

· Рост Вали = 1,7 метра, а Ивана 1,82.

· У Вали глаза голубые, у Ивана зеленые.

· Валя живет в Омске, Иван в Москве.

· Месячный доход Вали – 80.000 руб, а Ивана – 200.000 руб.

· Валя ездит на отдых за границу редко – раз в несколько лет, Иван часто – несколько раз в год.

Шкалы для измерения переменных

Каждая переменная может принимать различные значения. Значения переменных варьируются и отличаются от случая к случаю, от объекта к объекту.

Ну и Вы уже наверняка заметили, что они могут быть измерены в различных шкалах.

Например, пол – 0 и 1 или 1 и 0. Т.е, мужчина или женщина.

Доход, который выражается в рублях и может принимать большое количество разных значений, хоть до копеек.

Или частота поездок за границу, курения, использования интернета…

Разные шкалы имеют разную информативность. От того, какая шкала используется, зависят также и методы анализа, которые к ней можно применять.

Статисты придумали разные типы шкал ***(см. рис.)*** но их в целом можно объединить в три основных типа, которые в книге приводятся в порядке возрастания информативности:

Рис. Типы шкал – и их 3 основные вида

Номинальная шкала (***рис. 12***) – например, пол, город, страна, семейное положение, политическая партия, ФИО кандидата в президенты.

Рис. 12. Номинальная шкала

По сути, это шкала наименований и классификаций. С ней бессмысленно проводить какие-либо математические

операции. Цифры в ней ничего не значат (не имеют эмпирического значения). Если, например, мы поставим 1 Уфе, а 2 – Самаре, это не означает, что Уфа на ступеньку ниже Самары. Мы можем даже поменять цифры между городами – это ничего не изменит.

Т.е., эта шкала всего лишь определяет принадлежность наблюдения, случая или объекта к какой-то группе и позволяет классифицировать объекты. Тут мы можем посчитать только количество объектов в группе (количество или % мужчин и женщин в нашей выборке; количество людей из разных стран или профессий).

Отдельно при рассмотрении номинальных шкал стоит выделить ***дихотомии*** – переменные с двумя значениями. Пол, прошёл / не прошёл тест, выжил / погиб, любой вопрос с вариантами ответа только да / нет. Есть методы анализа, при которых удобно использовать именно дихотомии.

Второй тип шкал – **порядковая или ранговая** (***рис. 13***).

Рис.13. Порядковая (ранговая) шкала

Еще ее называют **ординальная** (от order – *с англ. порядок*). Например, воинское звание, место в организационной иерархии или уровень образования. Тут закладывается степень проявления какого-то свойства между объектами, но непонятна ни его точность, ни расстояния между ними.

Генерал выше полковника. Работа может быть интересна, безразлична или неинтересна. Занявший I место по бегу выше II и III (хотя разница в их абсолютном результате могла составить между ними всего 5 секунд).

Эту шкалу, как и номинальную, используют для классификации объектов и подсчета количества или %. Но по ней можно применять уже и не только частотный анализ – к примеру, можно попробовать найти связь между частотой использования мата и воинским званием.

Третий тип – **количественные\интервальные шкалы** (***рис. 14***).

Рис. 14. Интервальная (количественная, относительная, метрическая) шкала

Если предыдущая порядковая шкала несла инфо о порядке данных, то количественная – это числа, реально отражающие размерности, разности, масштабы и расстояния между

объектами.

Например, точное время, за которое бегуны пробежали дистанцию. Возраст лет. IQ. Уровень лояльности или мотивации сотрудника. Доход.

С этими шкалами можно осуществлять любые виды анализа. Более того, их можно легко превращать в порядковые, объединяя диапазоны значений. Например, доход можно разбить на 4 диапазона – низкий, средний, выше среднего и высокий.

Оговорюсь, что количественные (метрические) шкалы могут выглядеть по-разному: есть с отрицательными значениями, есть с абсолютным нулем (например, возраст) есть те, которые в принципе не начинаются с нуля (например, IQ). Аналитики в разговорах, статьях, литературе их могут именовать по-разному (например, интервальная, шкала масштаба или шкала отношений с абсолютным нулем…) – но, по сути, все они с точки зрения использования методов аналитического инструментария одинаковы.

Гипотезы

Когда говорят слово гипотеза, у многих возникает ассоциация с учеными или теориями. На самом деле гипотезами оперируют и менеджеры, бизнесмены, сотрудники компаний, криминалисты и т. д.

Например, создавая рекламную кампанию, менеджер по рекламе выдвигает гипотезу, почему и как реклама должна сработать – и на их базе строит свою кампанию. Бизнесмен, принимая решение вкладываться в дело или нет, выдвигает и размышляет над целым набором гипотез-предположений. Криминалист, расследуя перестрелку, выдвигает гипотезы, которые проверяются в ходе расследования и изучения фактов.

Например, я при проведении исследований персонала проверяю гипотезу, что определенный набор организационных факторов (зарплата, карьера, обучение и развитие, морально-психологический климат и т.д.) влияет на лояльность и мотивацию персонала.

Или прогнозируя будет кандидат успешным продавцом или нет в конкретной компании, в качестве гипотезы могу заложить предположение, что успешность определяют результаты по нескольким тестам, пол и уровень образования.

Гипотезы очень важны. Хорошо о них было сказано на 32 минуте последней сериии фильма «Михайло Ломоно-

сов» (Мосфильм, 1986): «Запомните, в основе науки лежит ежечасная работа по спирали опыта. Но не бойтесь и гипотез! Они в естественных и философских трудах подчас единственный путь, которым величайшие умы постигли самых важных истин. Гипотезы! Полет! Порыв души!…»

Гипотезы могут или быть верными, или отклоняться.

> И в современных подходах отклонить или принять гипотезу помогает расчет вероятности, являются наблюдаемые закономерности случайными, или можно считать их реальными. Особенно это важно для социально-экономической реальности, где не работают жестко предопределенные законы.

Так, например, для успешности продавца могут оказаться верными предположения по тестам и уровню образования, но будет отвергнуто влияние пола.

Любая гипотеза (наше предположение) в статистике раскладывается на две статистических гипотезы:

– нулевая (H_0), которая гласит, что обнаруженных в наборе данных (выборке) закономерностей **в генеральной совокупности нет** – это исключительно случайность, которая имеет место только в исследуемой Вами выборке.

– альтернативная (H_1), которая гласит противоположное: что обнаруженная в выборке **закономерность имеет место и в генеральной совокупности.**

Пока о гипотезах все. Больше о нулевых и альтернативных гипотез будут рассмотрены в следующей главе в привяз-

ке к понятию вероятности.

Вероятность

Вероятность в статистике выражается в % и лежит в диапазоне от 0 до 1 (0—100%). Обозначается буквой P – от *англ. probability*.

В повседневной жизни мы привыкли оценивать вероятность события или вероятность истинности каких-то утверждений. Например, 80% что пойдет дождь, 99% что я сдам этот тест, вероятность выбить с клиента долг менее 10%…

Но практическая статистика оперирует не вероятностью наступления события (или истинности утверждения), а **вероятностью ошибиться в случае применения обнаруженной закономерности ко всей генеральной совокупности**.

> Самым страшным и критичным в анализе считается именно обнаружить закономерности, взаимосвязи или различия, которых на самом деле в генеральной совокупности не существует.

А не обнаружить какие-то реально существующие взаимосвязи – это не так страшно. Это как в правосудии: выпустить виновного считается менее критичным, нежели обвинить невиновного…

Статисты придали этим вещам названия в виде **нулевой (H_0)** и **альтернативной (H_1)** гипотез. H_0 говорит, что обнаруженных закономерностей, взаимосвязей или отличий в ге-

неральной совокупности нет – это исключительно случайность, которая имеет место только в исследуемой Вами выборке.

> Я в свое время для себя просто запомнил, что нулевая гипотеза (H_0) – это **ноль различий / взаимосвязей / закономерностей**.

Только если вероятность H_0 крайне низка – принимается альтернативная гипотеза (H_1), что обнаруженная в выборке закономерность имеет место и в генеральной совокупности.

Т.е., в практике мы пытаемся в первую очередь ответить на вопрос – какова вероятность, что выведенная нами взаимосвязь между параметрами или закономерность является случайной и ее на самом деле нет в генеральной совокупности?

Например, криминалист, собрав все известные случаи, видит вроде как закономерность, что серийные маньяки орудуют в пределах трех кварталов от места жительства. Можно ли это распространить на всю генеральную совокупность? Или это просто случайное «стечение обстоятельств» в его выборке данных?

Конечно, проще всего было бы взять еще пару выборок из генеральной совокупности и убедиться, что в них также наблюдается такая связь. Но это не всегда возможно. И все равно ответ не может быть точным, пока не будет изучена вся генеральная совокупность.

Для того, чтобы чувствовать себя поувереннее, распространяя полученные на выборке закономерности на всю генеральную совокупность, используется очень узкий интервал – **не более 5% вероятности ошибки.**

Все закономерности (взаимосвязи, различия), вероятность ошибки по которым ниже этого интервала (т.е. менее 5%), считаются **статистически значимыми**. В англоязычной литературе обозначаются *Sig., Significant.*

> Именно наличие **значимых** закономерностей позволяет распространять полученные на выборке результаты на всю генеральную совокупность.

Как это работает? Например, мы хотим выяснить, проводят ли женщины больше времени в соцсетях, чем мужчины. Мы взяли определенную выборку из 1000 женщин и мужчин и обнаружили, что мужчины в среднем проводят в сетях 5 часов в неделю, а женщины 7 часов. Получается, что женщины на 2 часа (на 40%!) больше сидят в сетях.

Но можем ли мы на этих результатах утверждать, что в принципе все другие женщины больше сидят в соцсетях, чем мужчины? Возможно, мы получили различие случайно, и оно характеризует только эту выборку, а не всю генеральную совокупность…

И вот тут мы сначала определяем вероятность для H_0: что разницы по «просиживанию» в соцсети между мужчинами и женщинами нет. Или, другими словами, рассчитываем вероятность ошибки насчет того, что женщины сидят в соцсе-

ти больше мужчин.

И если вероятность ошибиться будет менее 5%, то мы можем говорить о том, что обнаружили **статистически значимое различие** – и таки можем говорить, что все женщины проводят в сети больше времени.

Почему берется такое низкое значение вероятности ошибки? Скажу, что на самом деле часто используют даже ниже 1% или менее. От чего зависит? На самом деле от отрасли и сложившейся в ней практики. Например, в медицине цена ошибки может быть высокой и там значения вероятности ошибок принимают обычно очень низкими.

В целом, общепринятая интерпретация вероятности ошибки (или значимости результатов) в среде аналитиков следующая (***рис. 15***):

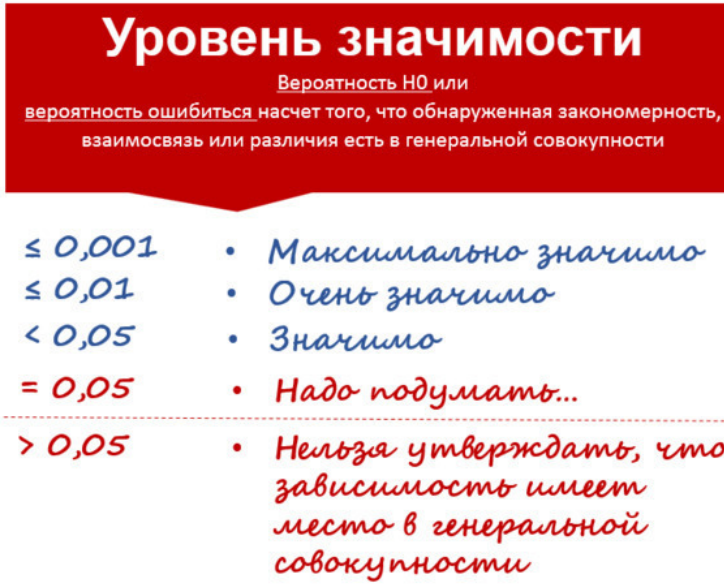


Рис. 15. Уровни значимости и их интерпретация

Прочитав этот раздел, я думаю, Вы уже поняли, насколько нами могут манипулировать с помощью различных опросов и исследований, в которых утверждается, что «женщины / мужчины лучше руководят», «опрошенные считают честным кандидата в президенты», «у ряда пациентов наблюдалось улучшается самочувствие после применения препарата» и т. д.

Широкой публике просто часто выдают информацию без

обозначения репрезентативности выборки, заложенной модели, еще и в придачу не указывая, являются ли эти взаимосвязи статистически значимыми.

Нормальное распределение

Колоколообразную кривую знают и наслышаны все (она же колокол Гаусса, гауссовское распределение – ***рис. 16***).

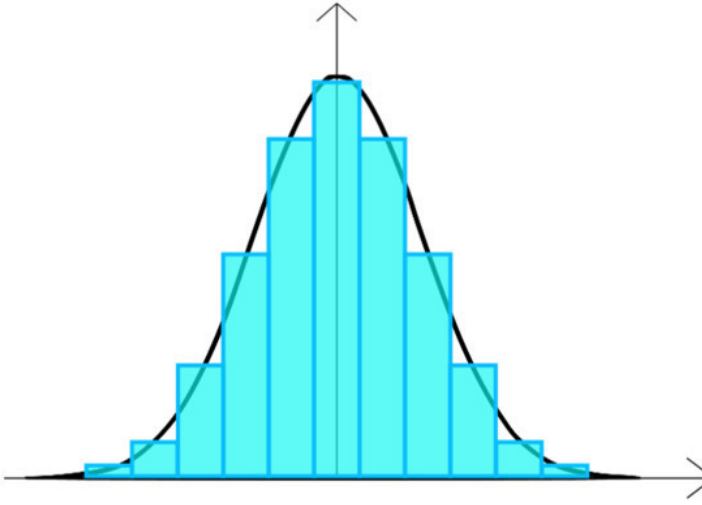
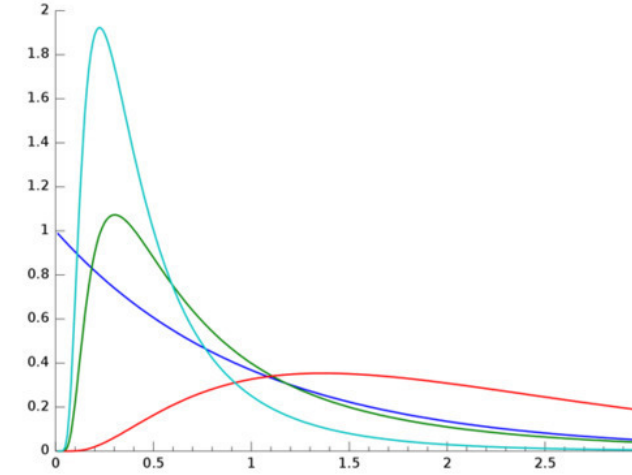


Рис. 16. То самое ОНО – нормальное распределение

Я о ней уже упоминал вначале, когда говорил об особенностях социально-экономической реальности в сравнении с естественно-технической.

И почему-то многие уверены, что этой кривой подчиняется все. На самом деле в реальности кривая нормального распределения чаще всего проявляется в физических параметрах, ограниченных физическими законами – гравитация, размеры, вес организмов определенного вида и т. д.

В социально-экономической реальности скорее наоборот – Вы будете встречать отсутствие нормального распределения. Оно буде скорее скошено вправо или влево, или очень сжато по оси OX или OY (***рис. 17***).



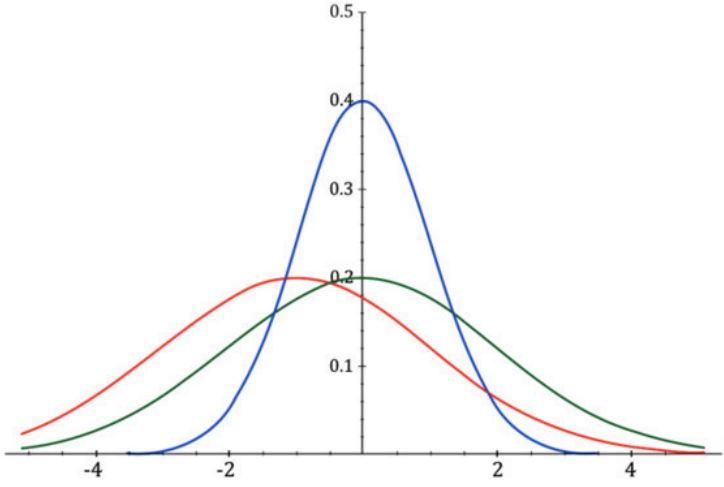


Рис. 17. Примеры реальных распределений в социаль-

но-экономической реальности

90% жителей страны владеют 2% капитала. 2 певца забирают 95% популярности. 99% тиража всех книг приходится на 1% авторов и т. д.

В любом случае на практике реальное распределение отклоняется от этой кривой. Да и выборки данных, строго соответствующие нормальному распределению, на практике, как правило, не встречаются.

> Но тем не менее, в статистике перед исследованием важно понимать соответствует ли распределение наших данных по каждой переменной нормальному распределению.

Для переменных, которые **нормально распределены** – используются одни параметры и критерии для сравнения (и среднее значение, дисперсия, стандартное отклонение – в этом случае информативные показатели).

Для тех переменных, которые **не соответствуют нормальному распределению** – другие критерии (тут скорее более информативными будут ранги, мода, медиана и т.д.).

Понять «на глаз» нормально ли распределены данные на самом деле может быть достаточно сложно. Бывает внешне похожее на нормальное распределение значимо от него отличается. А бывает наоборот – визуально не выглядящее нормальным распределение не имеет значимых отличий от нормального.

Поэтому для определения «нормальности» распределения разработаны специальные статистические тесты. Мы на этом остановимся позже в практических разделах книги.

Итоги раздела

В этом разделе основные мысли, которые хотелось бы «осадить» в памяти читателя, следующие:

1. Есть описательная и аналитическая статистика. Описательная статистика «ужимает» миллионы и миллиарды цифр к какому-то компактному числу, типичному для всего миллиона цифр. Аналитика позволяет находить **скрытые закономерности, которые дают нам больше понимания о реальности и как она работает, а также строить прогнозы**.

2. Выборка и генеральная совокупность. Генеральная совокупность – вся целиком популяция исследуемых объектов. Выборка – выбранные из этой популяции объекты (часть генеральной совокупности). Но **выборка должна быть репрезентативной** – т.е., отражать генеральную совокупность.

3. Переменные – это **признаки / характеристики изучаемых нами объектов** (люди, животные, товар, клиенты, организации и т.д.), которые могут принимать разные значения. Доход, пол, возраст, цвет и т. д.

4. В практике стоит различать три типа шкал для измерения переменных. **Номинальная:** шкала наименований – город, пол, профессия и т. д. **Ординальная / порядковая:** отражающая степень проявления какого-либо свойства, без

точных измерений – высокий-низкий; больше-меньше; I – II – III место и т. д. **Интервальная:** отражает размерность или масштаб каждой переменной – доход, возраст в годах, расстояние и т. д.

5. Мы выдвигаем наши предположения / суждения (как в виде мнений или домыслов, так и опыта) в виде гипотез, которые потом проверяем цифрами и аналитикой. В статистике фигурируют две гипотезы. Нулевая гипотеза (H_0), гласящая что закономерностей, взаимосвязей, различий в генеральной совокупности **не существует** – все что мы обнаружили всего лишь нелепая случайность в нашей выборке. И альтернативная (H_1), которая гласит, что обнаруженные в выборке различия нельзя объяснить случайностью: **они вероятнее всего имеют место и «материальны» в генеральной совокупности**.

6. Практическая статистика оперирует не вероятностью наступления события (или истинности утверждения), а вероятностью ошибиться в случае применения обнаруженной закономерности ко всей генеральной совокупности. **Самым страшным и критичным в анализе считается именно обнаружить закономерности, взаимосвязи или различия, которых на самом деле в генеральной совокупности не существует.**

7. Все закономерности (взаимосвязи, различия), по которым вероятность ошибки относительно их отсутствия в генеральной совокупности **менее 5% (менее 0,05)**, считают-

ся **статистически значимыми**.

8. В социально-экономической реальности Вы **редко будете встречать нормальное распределение**. Оно будет скорее скошено вправо или влево, или очень сжато к оси OX или OY. 90% жителей страны владеют 2% капитала, 2 певца забирают 95% популярности, 99% тиража всех книг приходится на 1% авторов и т. д.

КРАТКО О ПОДГОТОВКЕ МАССИВА ДАННЫХ ДЛЯ АНАЛИЗА

Что такое массив данных

Массивом данных для пользователей как мы с Вами по большому счету является таблица, в которую внесены данные. Главное: в массиве все данные по той или иной переменной должны соотноситься с конкретным случаем, объектом, процессом, явлением.

Строки таблицы – это случаи или объекты (ФИО, завод, филиал, клиент и т.д.).

Столбцы\Колонки – это наши переменные, то есть характеристики этих случаев или объектов (доход, % брака, возраст, пол, страна и т.д.).

Массивом для последующей аналитической обработки является «плоская» таблица (не сведенный отчет). См. ***рис. 18***.

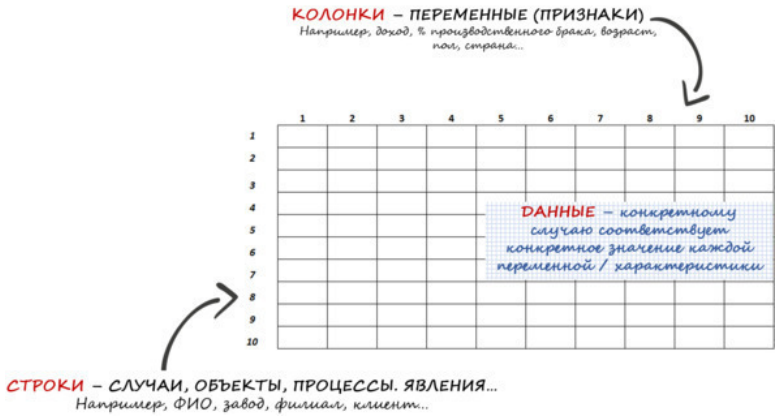


Рис. 18. Базовая структура массива данных

В массивах **по строкам** идут случаи / объекты / процессы (компания, дата замера, человек, клиент и т.д.), а **по столбцам\колонкам** – исследуемые переменные с их **значениями** для этих случаев / объектов / процессов по ячейкам.

В массиве не должно быть никаких объединений ячеек или по несколько разных переменных в одной ячейке. Каждая переменная – отдельная колонка и ее значение для каждого объекта / случая записывается в отдельную ячейку.

Конец ознакомительного фрагмента.