МИНИСТЕРСТВО ВНУТРЕННИХ ДЕЛ
РОССИЙСКОЙ ФЕДЕРАЦИИ
АКАДЕМИЯ УПРАВЛЕНИЯ

Ш.Х. Гонов, И.В. Горошко

АКТУАЛЬНЫЕ ВОПРОСЫ АНАЛИЗА ДАННЫХ, ХАРАКТЕРИЗУЮЩИХ РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ВНУТРЕННИХ ДЕЛ

Учебное пособие

Москва · 2022

Академия управления МВД России

Ш. Х. Гонов, И. В. Горошко

АКТУАЛЬНЫЕ ВОПРОСЫ АНАЛИЗА ДАННЫХ, ХАРАКТЕРИЗУЮЩИХ РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ВНУТРЕННИХ ДЕЛ

Учебное пособие

Москва • 2022

УДК 004.03
ББК 32.97
Г65

Одобрено редакционно-издательским советом Академии управления МВД России

Рецензенты: *И. А. Кубасов,* главный научный сотрудник НИИСТ ФКУ НПО «СТиС» МВД России, доктор технических наук, доцент; *М. Ю. Паклянченко,* доцент кафедры специальных информационных технологий учебно-научного комплекса информационных технологий Московского университета МВД России имени В. Я. Кикотя, кандидат технических наук.

Гонов Ш. Х., Горошко И. В.

Актуальные вопросы анализа данных, характеризующих результаты деятельности органов внутренних дел : учебное пособие / Ш. Х. Гонов, И. В. Горошко. – Москва : Академия управления МВД России, 2022. – 128 с.

Г65

ISBN 978-5-907530-05-8

В учебном пособии рассматривается система информационно-аналитического обеспечения научных исследований и образовательного процесса в органах внутренних дел Российской Федерации. В системном и обобщенном виде описываются методы и модели исследования социальных и экономических систем, а также механизмы управления сложными организационными системами.

Учебное пособие может использоваться в образовательном процессе по дисциплинам, преподаваемым по программе подготовки научно-педагогических кадров по направлениям подготовки 09.07.01 – Информатика и вычислительная техника, 37.07.01 – Психологические науки, 38.07.01 – Экономика, 40.07.01 – Юриспруденция, 44.07.01 – Образование и педагогические науки, а также по программам магистратуры и дополнительным профессиональным программам повышения квалификации. Материалы учебного пособия могут быть полезны практическим работникам штабных и информационных подразделений, а также соискателям ученых степеней и специалистам, интересующимся актуальными вопросами использования компьютерных технологий в научных исследованиях и образовательном процессе.

УДК 004.03
ББК 32.97

ISBN 978-5-907530-05-8

Оглавление

Введение

Для изучения объекта исследования необходимо иметь ясное представление не только о внутренних (эндогенных), но и о внешних (экзогенных) переменных, характеризующих закономерности функционирования организационной системы. Моделируя состояние объекта, развитие явления или структуру процесса, исследователь должен анализировать и прогнозировать возможные воздействия на них со стороны внешней среды. Модели такого вида можно получить с помощью современных методов статистического анализа данных, в основе которых лежат основные положения теории управления, законы теории вероятностей и математической статистики, теории систем и т. д.

Существует множество различных подходов, методов и моделей исследования организационных систем, которые были созданы для их описания и решения практических задач. И сами модели, и информационные технологии, созданные для их реализации, представляют собой очень разнообразное сочетание различных математических подходов и методов. Это обстоятельство затрудняет систематизацию задач исследования организационных систем и отрицательно сказывается на ресурсном обеспечении.

Одна из задач пособия – систематизировать существующие математические подходы и методы описания организационных систем, а также дать эффективный инструментарий молодым исследователям для решения задач моделирования в первую очередь социально-правовых систем.

Актуальность данного учебного пособия обусловлена необходимостью обобщения и систематизации знаний и подходов к современным аспектам построения математических моделей описания организационных систем на основе компьютерных технологий.

Пособие призвано помочь реализации одной из главных задач комплекса мероприятий по совершенствованию системы подготовки кадров для органов внутренних дел Российской Федерации – повышению качества и усилению практической направленности системы подготовки высококвалифицированных научных и научно-педагогических кадров.

Глава 1. Методы визуализации данных

В настоящей главе рассматриваются основные методы отображения статистических данных при решении научно-исследовательских задач, а также способы визуализации данных с помощью гистограмм, графиков и других типов диаграмм и технологии их создания[1].

1.1. Виды данных и их визуализация

Визуализация данных – это набор методов, которые позволяют использовать визуальное представление для изучения, анализа и коммуникации количественных данных[2]. Это помогает изучать тенденции и закономерности в имеющемся у исследователя наборе данных. Конечная цель визуализации данных – способствовать принятию более эффективных решений и мер.

Чем больше становятся объемы доступных нам данных, тем важнее иметь возможность интерпретировать постоянно увеличивающиеся массивы информации, и визуализация данных позволяет эффективно решить эту задачу. Сказанное актуально не только для специалистов по обработке данных и аналитиков, владение методами визуализации данных необходимо в научной и образовательной сферах.

Данные можно визуализировать на различных этапах анализа и коммуникации. Во-первых, на ***этапе исследования данных*** производить анализ, визуализируя их с помощью инструментов статистического анализа и электронных таблиц, гораздо проще. С помощью инструментов визуализации можно выявлять различные связи, изучать распределения и сравнивать данные. На этапе исследования данных не столь важно, какой тип диаграммы выбрать, какие пояснительные надписи разместить и как оформить иллюстративный материал, главное, чтобы визуализация позволяла аналитику получить новую информацию.

После того, как удалось достаточно глубоко разобраться в наблюдаемых тенденциях и закономерностях, начинается ***этап презентации данных***. Цель презентации данных (иногда называемой представлением данных) – ознакомить целевую аудиторию с конечными результатами исследования. Они могут быть представ-

---

[1] Информационные технологии в науке и образовании: учебное пособие / И. В. Горошко, Б. А. Торопов. Москва: Академия управления МВД России, 2021. 76 с.

[2] Инструменты для качественной визуализации данных: искусство использования диаграмм. Копенгаген: Европейское региональное бюро ВОЗ, 2021. Лицензия: CC BY-NC-SA 3.0 IGO.

лены в отчете, справке, презентации или в диссертационном исследовании. На этом этапе большое значение приобретают визуальное оформление материала, тип выбранной диаграммы, пояснительные подписи и т. д. Переработав данные в информацию, мы должны помочь широкой аудитории усвоить сделанные нами выводы.

Силу визуализации можно проиллюстрировать на примере, который известен как квартет Энскомба. Это четыре набора данных, которые почти идентичны по описательным характеристикам, но имеют разное распределение и при графическом представлении дают совершенно разную картину. Каждый набор данных состоит из 11 точек (*x*, *y*). Эти наборы данных были разработаны в 1973 г. специалистом по статистике Фрэнсисом Энскомбом (*англ. Francis John Anscombe*), чтобы продемонстрировать, как важно перед анализом данных представить их в виде диаграммы (графика).

Таблица 1.1. Квартет Энскомба

	Набор I		**Набор II**		**Набор III**		**Набор IV**	
	x1	*y1*	*x2*	*y2*	*x3*	*y3*	*x4*	*y4*
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
Среднее	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
Дисперсия	10,00	3,75	10,00	3,75	10,00	3,75	10,00	3,75
Корреляция	0,82		0,82		0,82		0,82	
Уравнение регрессии	*y=0,5*x+3*		*y=0,5*x+4*		*y=0,5*x+5*		*y=0,5*x+6*	
Коэффициент детерминации	0,67		0,67		0,67		0,67	

В таблице 1.1 представлены наборы данных, у первых трех значения x одинаковы. Заметим, что приведенные в конце таблицы некоторые описательные статистические характеристики одинаковы, поэтому можно предположить, что эти наборы данных идентичны, но если представить их в виде диаграмм, то различия становятся очевидны (рис. 1.1–1.4).

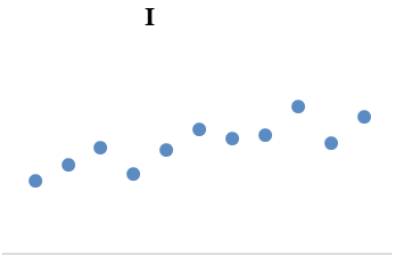


Рис. 1.1. Квартет Энскомба (I)

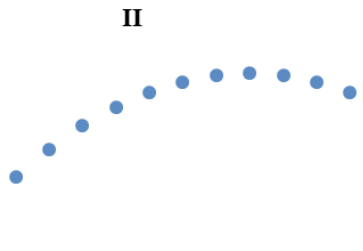


Рис. 1.2. Квартет Энскомба (II)

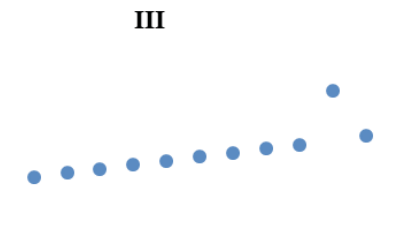


Рис. 1.3. Квартет Энскомба (III)

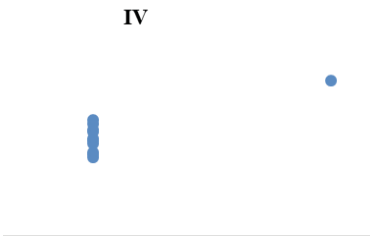


Рис. 1.4. Квартет Энскомба (IV)

Таким образом, визуализация данных может дать информацию, которую не всегда дают табличные данные и описательные статистики. Без визуального представления информации зачастую трудно понять истинное значение полученных результатов.

Способы визуализации данных

Рассмотрим основные способы представления данных. Визуализация данных имеет множество применений, и каждый ее вид можно использовать по-разному[1]. Самым распространенным способом применения визуализации данных являются временные изменения, или изменения в динамике. Этот способ наиболее распро-

[1] Более подробно различные типы диаграмм рассматриваются в следующем параграфе.

странен, поскольку в большинстве наборов данных присутствует временной фактор *t*. Для этого типа данных лучше всего подходят линейные графики. Однако нужно иметь в виду, что, если необходимо отобразить много линий тренда сразу, линейные графики часто оказываются перегружены и образуют диаграммы типа «спагетти» (рис. 1.5).

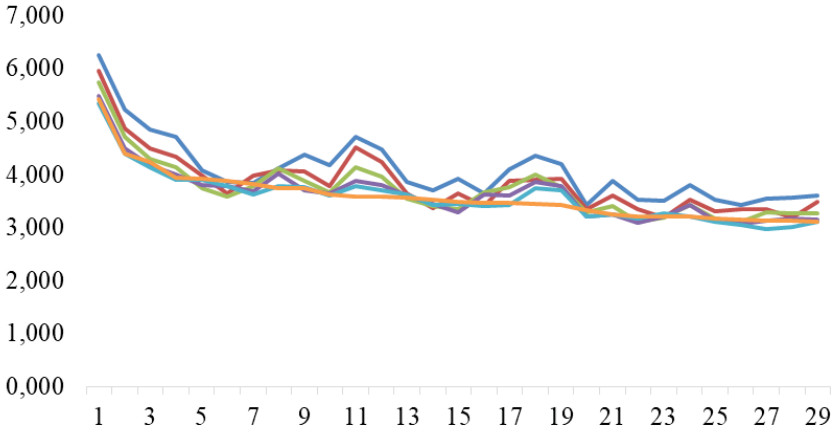


Рис. 1.5. «Спагетти» диаграмма

В качестве дополнительной графической переменной можно использовать цвет. Визуализация данных в данном случае столь же проста, как цветовое кодирование; такой прием может помочь при отображении больших наборов данных с большим количеством линий. Цветовая маркировка просто добавляет дополнительную переменную к координатам по осям *X* и *Y*.

Отображение частот – еще одно частое применение визуализации данных. Оно возможно, когда данные разделены на классы, которые могут быть порядковыми либо номинальными. Для этого типа визуализации лучше всего подходят столбчатые диаграммы, линейчатые диаграммы и гистограммы.

Выявление корреляций (определение отношений) – чрезвычайно ценное применение визуализации данных. Без визуализации трудно определить отношения между двумя переменными, однако знать о взаимосвязях в данных крайне важно. Для этого очень полезны точечные диаграммы и пузырьковые диаграммы. Визуализация наборов данных Энскомба (рис. 1.1–1.4) является примером точечной диаграммы. Это замечательный пример, пока-

зывающий ценность визуализации при анализе данных. Становится очевидно, что для понимания корреляций и выбросов одной описательной статистики недостаточно.

При изучении **пространственных закономерностей** (наличии пространственного компонента) хорошим способом визуализации распределения по территориальному признаку (например, по району, городу, региону или стране) являются картографические решения. Этот способ может быть удобен, например, для визуализации различий в уровне преступности между регионами страны.

Для определения сложных показателей, таких как значения с доверительными интервалами, необходимо учитывать множество различных переменных, что делает адекватный просмотр данных с помощью простой электронной таблицы почти невозможным. Для визуализации значений со степенью доверительной вероятности удобно использовать диаграмму **диапазонов с областями**.

Иногда может потребоваться сочетать разные способы визуализации данных. Если, например, необходимо визуализировать изменение частотного значения с течением времени, можно поместить рядом две линейчатых диаграммы, однако наилучшим вариантом будет **диаграмма наклона**. Сравнение частот с другой популяцией или целевым ориентиром может быть выполнено с помощью диаграммы-шкалы.

Любой набор данных имеет ряд характеристик. Рассмотрим две характеристики, которые важно учитывать при выборе типа диаграммы:

1) содержит ли набор данных индивидуальные (отдельные) единицы данных или агрегированные (сгруппированные);

2) какая используется шкала измерения.

Первой характеристикой набора данных является то, содержит ли он индивидуальные или сгруппированные данные. Так, индивидуальные данные могут быть визуализированы в виде точечной диаграммы или представлены в обобщенном виде как частоты на гистограмме.

Основная характеристика сгруппированных данных заключается в том, что отдельные наблюдения объединяются в группы. Примеры:

- количество совершенных преступлений за месяц;
- количество новых случаев COVID-19 за сутки;
- количество погибших в ДТП.

Сгруппированные данные лучше всего визуализировать с помощью линейных графиков, столбчатых диаграмм или линейчатых диаграмм.

Вторая характеристика набора данных – это шкала измерения. Выбранная шкала определяет тип данных и операции, которые можно осуществлять с этими данными. Состояние объекта исследования оценивается по критериям, а оценки измеряются по определенной шкале. В 1940-х гг. американский психолог Стэнли Смит Стивенс (*англ. Stanley Smith Stevens*) выделил четыре шкалы измерения: номинальную (наименований), порядковую (ранговую), интервальную и шкалу отношений. Эти шкалы активно используются в научных и прикладных исследованиях для описания характеристик переменных. Определение шкалы измерения переменных очень важно для выбора корректного метода исследования и правильного способа визуализации данных.

Номинальная шкала (шкала наименований) характеризует переменную по принадлежности к определенной категории. Иными словами, это качественная шкала, компоненты которой представляют собой связанные между собой отдельные элементы, которые не предполагают какого-либо строгого порядка. Номинальные переменные можно кодировать числами, но порядок присвоения этих чисел будет произвольным, и любые вычисления с ними – некорректными.

Примерами номинальных переменных могут быть пол, цвет автомобиля, список субъектов Российской Федерации и др. Для этой шкалы измерения наилучшим образом подойдет линейчатая диаграмма. Поскольку качественные компоненты не имеют собственного порядка, их можно произвольно переставлять, чтобы выявлять закономерности в данных.

Порядковая (ранговая) шкала – это шкала, в которой важен порядок следования уровней, но не разница между значениями. Ее компоненты представляют собой элементы, которым свойственна некая естественная последовательность, например, «холодно – тепло – горячо» или «белый – серый – черный». Примерами порядковых переменных могут быть уровень образования (начальное, среднее, высшее), уровень дохода (высокий, средний, низкий), удовлетворенность качеством оказанной государственной (муниципальной) услуги (удовлетворительно, неудовлетворительно).

Для этой шкалы измерения лучше всего подходит столбчатая диаграмма.

Интервальная шкала – это шкала разностей, в которой уровни упорядочены, а интервалы между ними равны. Ее компоненты представляют собой элементы с постоянным числовым соотношени-

ем между собой. Примером интервальной шкалы может быть измерение времени (последовательность секунд, минут), температуры (по Фаренгейту, по Цельсию).

Переменная на ***шкале отношений*** (абсолютной шкале) имеет все свойства интервальной, но для нее также четко определен абсолютный ноль. Когда переменная равна нулю, означаемая ею сущность отсутствует. Примеры переменных, для которых используется шкала отношений: продолжительность, вес, длина, стоимость (цена).

При работе с переменными на шкале отношений, в отличие от интервальных переменных, можно получить содержательную интерпретацию, оценив соотношение двух значений. Как для интервальных переменных, так и для переменных на шкале отношений лучше всего использовать линейные графики.

Иногда интервальные данные могут отображаться при помощи довольно специфических инструментов технического анализа. Ниже приведена схема биржевой диаграммы (рис. 1.6), которая называется «Японские свечи» (*англ. candlestick chart / Japanese Candlesticks*), применяемой в первую очередь для отображения изменений биржевых котировок акций, цен на валюту и т. д. Кроме этого, на рисунке отображена схема диаграммы ящик с усами (*англ. box and whiskers plot*) с отображением квартилей (Q0-Q4).

Рис. 1.6. Японские свечи и ящик с усами

Для построения ***биржевой диаграммы*** (блочной, ящичной) (*от англ. box chart*) необходимо создать таблицу, в которой последовательно расположить следующие данные: цена открытия, максимальная цена, минимальная цена и цена закрытия. В качестве подписей обычно используются даты или наименования индикаторов (рис. 1.7).

Рис. 1.7. Биржевая диаграмма[1]

«Японские свечи» сочетают в себе свойства линейного графика и интервальной диаграммы и активно используются как эффективный инструмент теханализа, например, на основе чисел Фибоначчи (веер, дуги, зоны и др.).

Кроме этого, в визуализации данных активно применяются ***комбинированные диаграммы*** (*англ. combo chart*), сочетающие в себе разные типы. Ниже представлено сочетание линейчатой и столбчатой диаграмм (рис. 1.8).

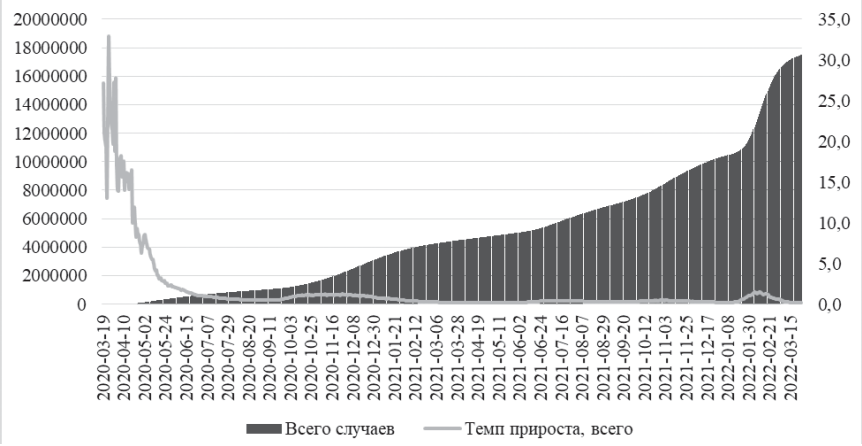


Рис. 1.8. Комбинированные типы диаграмм

[1] Данные для построения диаграммы взяты с портала «ProFinance» (https://www.profinance.ru).

На данном графике представлен временной ряд заболеваемости коронавирусом COVID-19 в Российской Федерации по состоянию на 28 марта 2022 г. Для сравнения на графике также отражен темп прироста[1].

Еще один инструмент визуализации данных – ***поверхностные диаграммы*** (*англ. surface chart*), которые могут использоваться для отображения трендов в значениях по двум измерениям (рис. 1.9).

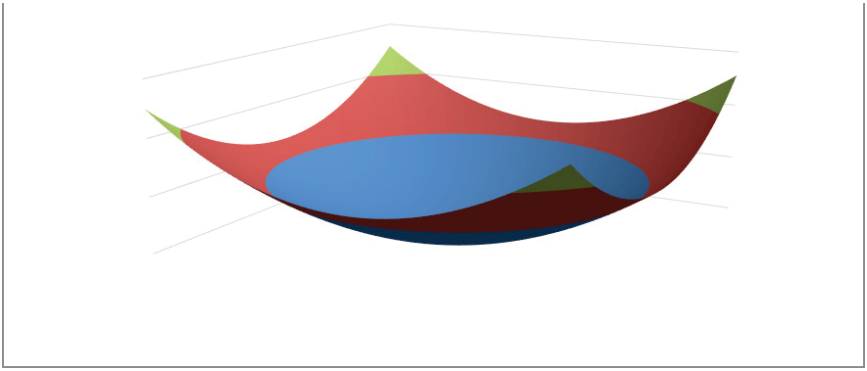


Рис. 1.9. График функции $z=x^2+y^2$

В исследовании экономических и социальных систем такие диаграммы могут использоваться для отображения построенных моделей, например, производственной функции Кобба – Дугласа. Методику построения регрессионной модели данного типа и ее визуализации рассмотрим в главе 3 настоящего учебного пособия.

Средством визуализации данных о численности населения может выступать возрастно-половая пирамида (*англ. population pyramid*). На следующем рисунке представлена ***возрастно-половая пирамида***, построенная на основе официальных данных Федеральной службы государственной статистики (Росстата) о численности населения Российской Федерации по полу и возрасту на 1 января 2021 г.[2] (рис. 1.10).

[1] Данные с портала «Our World in Data» (https://ourworldindata.org/covid-vaccinations?country=OWID_WRL).

[2] Данные с официального сайта Росстата (https://rosstat.gov.ru/compendium/document/13284).

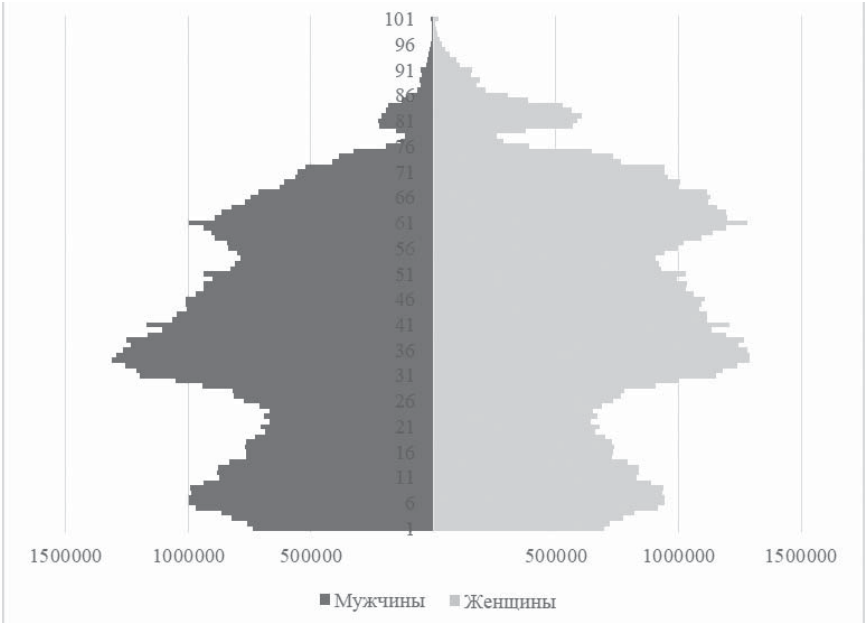


Рис. 1.10. Возрастно-половая пирамида по РФ за 2021 г.

Данный тип диаграммы хорошо иллюстрирует социально-экономические, социально-демографические и политические сдвиги в общественной жизни. На пирамиде хорошо заметны демографические волны снижения числа родившихся в середине 40-х, конце 70-х и 90-х гг., а также рост числа родившихся в конце 80-х и начале 60-х гг. Кроме этого, в верхней части пирамиды видно значительное превышение численности женщин над мужчинами, особенно старших возрастов.

1.2. Основные типы диаграмм

Линейный график *(англ. line chart)* используется для отображения интервальной шкалы или шкалы отношений по оси абсцисс *X* и количественного показателя по оси ординат *Y*. Часто по оси *X* расположена шкала времени, но могут использоваться и другие непрерывные шкалы. Линейный график хорошо подходит для иллюстрации развития процесса (явления) в динамике, например, количества зарегистрированных в отчетном периоде преступлений, квалифицируемых по статье 158 Уголовного кодекса Российской Федерации (далее – УК РФ) (рис. 1.11).

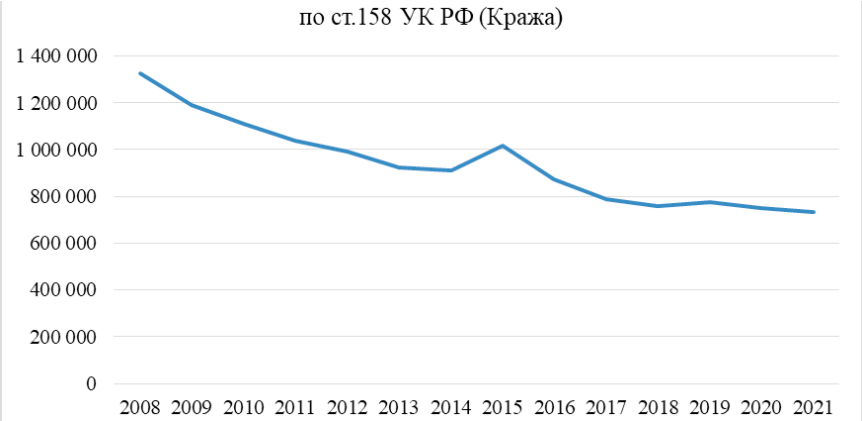


Рис. 1.11. Динамика краж, зарегистрированных в отчетном периоде

Диаграмма с областями (*англ. stacked area chart*) – это вариант линейного графика, где область под линией закрашена, чтобы подчеркнуть ее значимость. Если на графике отображено более одной переменной, диаграмму с областями можно использовать как линейный график с накоплением (рис. 1.12). Значения отдельных категорий, показанных на диаграмме, складываются в общую сумму. На диаграммах этого типа нет необходимости отображать строку «Всего количество преступлений», так как отдельные показатели сами складываются в итоговое значение.

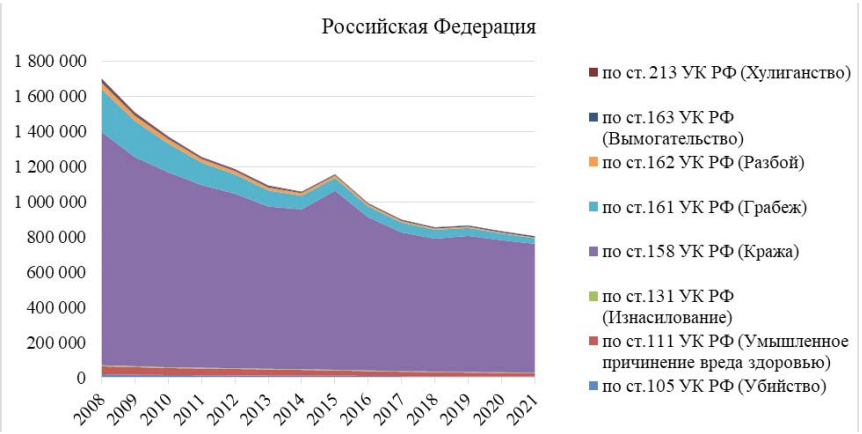


Рис. 1.12. Всего преступлений, зарегистрированных в отчетном периоде

Диаграмма диапазонов с областями – это диаграмма с областями, где область определяется двумя значениями: верхним и нижним. Диа-

грамма диапазонов с областями в основном используется для отображения предполагаемого диапазона конкретного показателя. В следующем примере в качестве диапазона используется стандартное отклонение (рис. 1.13).

Рис. 1.13. Всего зарегистрировано преступлений по России

Использование ***столбчатой диаграммы*** (*англ. column chart)* – лучший способ отобразить распределение значений порядковых переменных. На рис. 1.14 показан пример распределения безработных по возрастным группам по Российской Федерации.

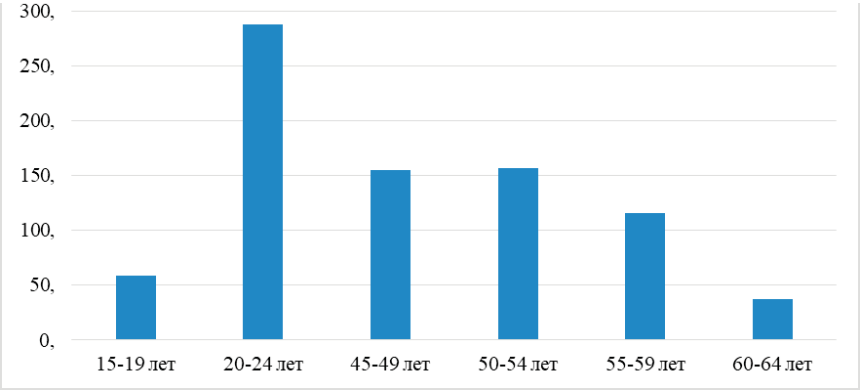


Рис. 1.14. Численность безработных по возрастным группам

Этот же вид диаграммы подходит для сравнения двух и более показателей. Например, для сравнения соотношений численности безработных по возрастным группам и полу (рис. 1.15).

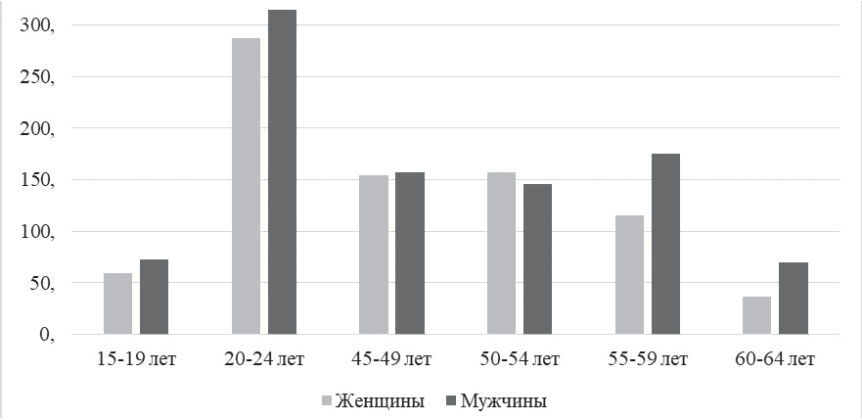


Рис. 1.15. Численность безработных по возрастным группам и полу

Линейчатая диаграмма *(англ. bar chart)* лучше всего подходит для отображения распределения значений номинальных переменных. Особенностью отображения номинальных переменных является обязательность подписи горизонтальной оси. Поскольку подписи, обозначающие номинальные переменные, могут быть длинными, линейчатая диаграмма в данном случае удобнее, чем столбчатая.

Если для отображения номинальных переменных используется столбчатая диаграмма, подписи, идущие вдоль оси X, могут стать слишком длинными для размещения по горизонтали. Поэтому для номинальных данных более оптимальным решением является линейчатая диаграмма (рис. 1.16). С технической точки зрения на линейчатой диаграмме оси X и Y расположены не так, как на столбчатой: ось Y ориентирована горизонтально, а ось X – вертикально.

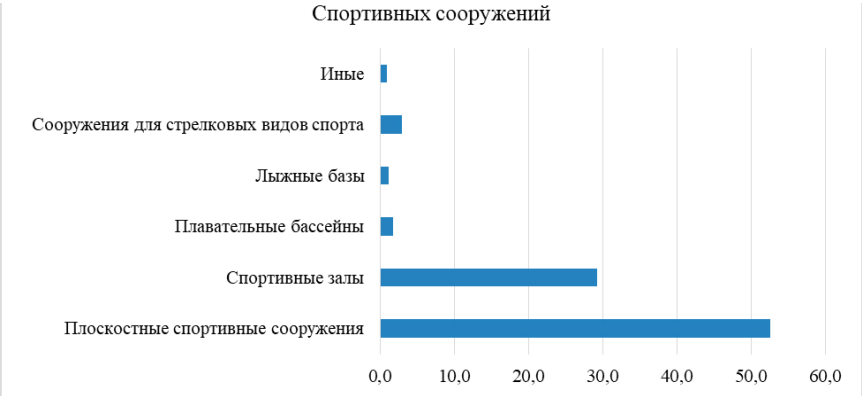


Рис. 1.16. Количество спортивных сооружений

Диаграмма-шкала с маркером – это линейчатая диаграмма с дополнительным маркером (засечкой, отметкой) на каждой линии (полосе). Такой маркер, например, может отмечать значение аналогичного показателя для другой группы или целевое значение для сравнения (рис. 1.17).

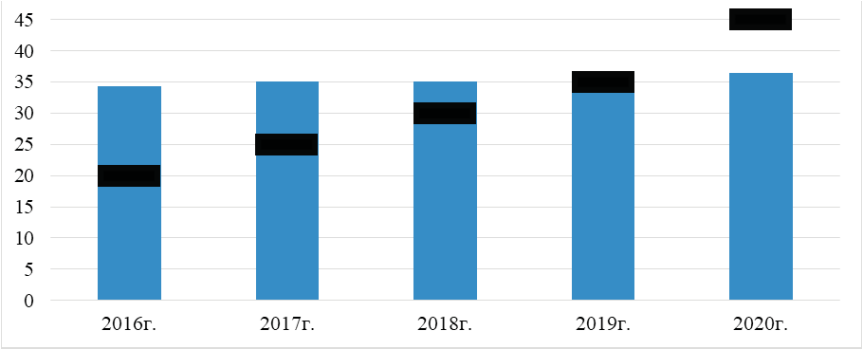


Рис. 1.17. Целевые индикаторы

Гистограмма *(англ. histogram)* выглядит как линейчатая диаграмма, но отражает распределение частот, а не тренд на порядковой шкале. По оси X гистограммы перечислены разряды (интервалы) переменной; по оси Y отсчитывается частота, поэтому каждая полоса пропорциональна частоте соответствующего разряда (рис. 1.18).

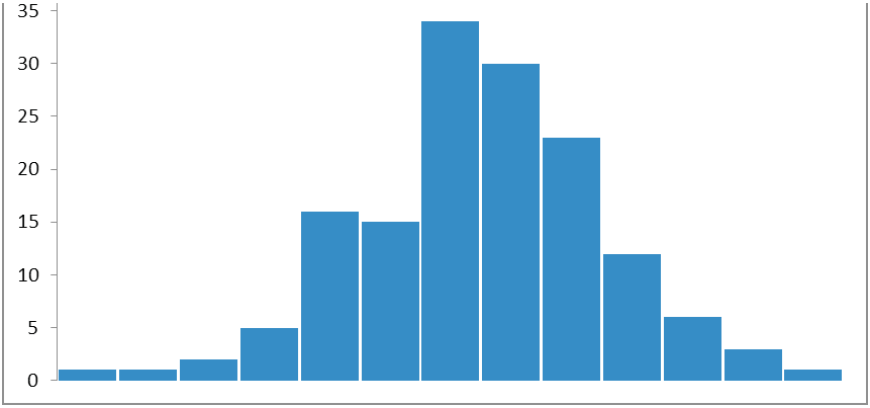


Рис. 1.18. Гистограмма

Гистограмма – это представление нормального распределения (распределения Гаусса) числовых данных. Впервые была предложена Карлом Пирсоном (*англ. Carl Pearson*) в 1895 г. Для того чтобы

построить гистограмму, необходимо сначала выделить серию разрядов (интервалов)[1], а следом подсчитать, сколько значений попадает в каждый разряд. Обычно разряды имеют одинаковый размер и располагаются по соседству, так как гистограмма показывает частоту.

Разряды могут выделяться разной ширины и в разном количестве. Самый распространенный подход к выбору количества разрядов – это метод квадратного корня. Для этого извлекается квадратный корень из числа элементов данных в выборке и округляется до следующего целого числа[2].

Точечные диаграммы *(англ. scatter chart)* (диаграммы рассеяния, разброса) чаще всего используются для поиска статистической связи (корреляций). Каждая точка на точечной диаграмме имеет координаты по оси абсцисс (x) и оси ординат (y) $A_1(x_1,y_1)$.. $A_n(x_n,y_n)$. Таким образом, если в распределении точек наблюдается определенный тренд (повышающий, понижающий и т. д.), между ними существует связь. Если точки полностью рассеяны и трендов не наблюдается, можно заключить, что переменные вообще не влияют друг на друга. Однако для того, чтобы точно сказать, есть связь или нет, используется корреляционный анализ, который мы рассмотрим в следующей главе.

Точечная диаграмма на рис. 1.19 показывает взаимосвязь между весом и ростом 5 000 человек с разбивкой по полу[3].

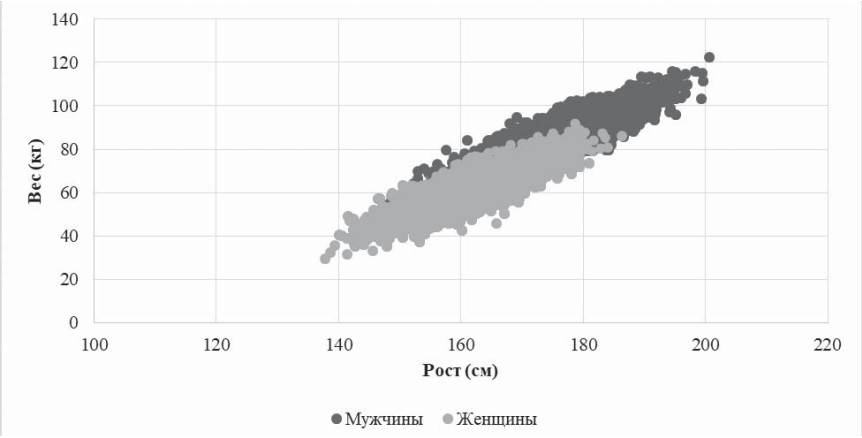


Рис. 1.19. Диаграмма разброса

[1] Иногда их называют карманами.

[2] Этот метод используется для построения гистограмм в MS Excel.

[3] Диаграмма построена на основе информации из набора данных, опубликованных на веб-платформе по адресу: https://www.kaggle.com/mustafaali96/weight-height.

Пузырьковая диаграмма *(англ. bubble chart)* – это вариант точечной диаграммы, где каждая точка изображена в виде «пузырька», площадь которого также несет определенную информацию в дополнение к положению точки по координатным осям (рис. 1.20). Трудность, связанная с пузырьковыми диаграммами, заключается в том, что пузырьки могут не помещаться на осях; поэтому не все данные подходят для этого типа визуализации. На следующей диаграмме размер пузырьков отражает данные о численности населения указанных стран[1].

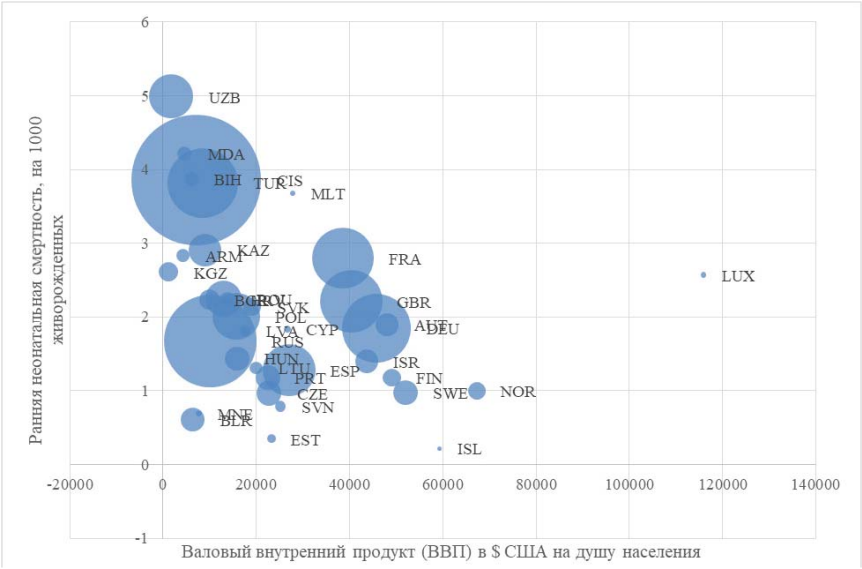


Рис. 1.20. Пузырьковая диаграмма

Круговая диаграмма *(англ. pie chart)* – это способ иллюстрации процентных долей, поскольку она показывает каждый элемент как часть целого. Ее основное преимущество заключается в том, что идея об «отношении части и целого» отражена в ней самым непосредственным образом (рис. 1.21).

[1] Диаграмма построена на основе данных с Европейского портала информации здравоохранения (https://gateway.euro.who.int/ru/hfa-explorer/).

Рис. 1.21. Круговая диаграмма

Существенным недостатком таких диаграмм является то, что размер любых других долей на круговой диаграмме оценить трудно. Причем такие диаграммы могут быть как объемными (*англ. 3D pie chart*) (рис. 1.22), так и кольцевыми (*англ. donut chart*) (рис. 1.23).

Рис. 1.22. Объемная круговая диаграмма

Однако, несмотря на очевидность содержимого круговой диаграммы, линейчатые диаграммы гораздо лучше приспособлены для сравнения величин каждой из частей. Круговые диаграммы позволяют легко

оценить величину сегмента, только когда она близка к 0 %, 25 %, 50 %, 75 % или 100 %.

Рис. 1.23. Кольцевая диаграмма

На круговой диаграмме необходимо использовать разные цвета для разных сегментов, а это означает, что для визуализации данных может потребоваться множество цветов. На линейчатой диаграмме разные элементы могут иметь одинаковые цвета, хотя при необходимости можно выделить один элемент, назначив его полосе другой цвет. Причем выбор конкретного типа диаграммы зависит от фантазии исследователя или аналитика, а также от конкретной задачи и целевой аудитории.

Для решения различных задач могут применяться разные виды диаграмм. Например, при решении задач планирования и управления проектами могут применяться ***диаграммы Ганта***[1], представляющие собой метод визуального планирования задач. Для ее построения используются две оси: по вертикали находится перечень задач, а по горизонтали время для их выполнения. Сначала создается таблица с исходными данными, с указанием перечня задач и продолжительности их выполнения, времени начала и окончания работ и др. (таблица 1.2).

Таблица 1.2. Исходные данные для построения диаграммы Ганта

№ п/п	Задача	Начало (план)	Конец (план)	Длительность (план)	Начало (факт)	Длительность (факт)	% выполнения

[1] Разработаны американским инженером Генри Гантом во время Первой мировой войны.

1	Задача 1	01.01.2021	02.01.2021	1			
2	Задача 2	01.01.2021	03.01.2021	2			
3	Задача 3	02.01.2021	04.01.2021	2			
4	Подзадача 3.1	02.01.2021	03.01.2021	1			
5	Задача 4	02.01.2021	04.01.2021	2			
6	Задача 5	03.01.2021	05.01.2021	2			
7	Задача 6	04.01.2021	08.01.2021	4			
8	Задача 7	05.01.2021	09.01.2021	4			
9	Задача 8	06.01.2021	08.01.2021	2			
10	Подзадача 8.1	06.01.2021	07.01.2021	1			
11	Задача 9	07.01.2021	10.01.2021	3			

Чаще всего для построения диаграмм Ганта используется специализированное программное обеспечение, такое как dotProject, ProjectLibre, GanttProject и др[1]. Но для решения простейших задач могут использоваться и табличные процессоры. На следующем рисунке представлена диаграмма Ганта в виде линейчатой диаграммы, построенной стандартными средствами MS Excel (рис. 1.24).

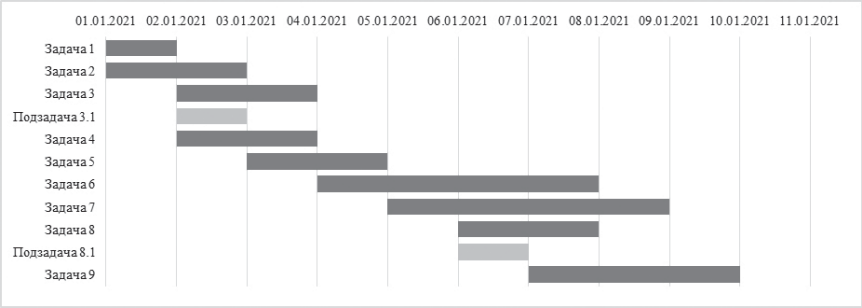


Рис. 1.24. Диаграмма Ганта

Диаграмма наклона является одной из модификаций линейного графика. На линейном графике отображаются три или более точки времени, а на диаграмме наклона – ровно две (рис. 1.25). Диаграммы наклона могут быть эффективным инструментом для визуализации

[1] В данном списке представлено программное обеспечение, распространяемое по лицензиям General Public License (GPL) или Common Public Attribution License Version (CPAL).

изменений значений для одного объекта и сравнения его с другими. Этот тип диаграммы может быть полезен для демонстрации:

– иерархических отношений набора объектов в два момента времени;

– процентной доли каждого из объектов в определенный момент времени;

– темпов изменения доли объекта в сравнении с другими объектами;

– любых отклонений в общем тренде.

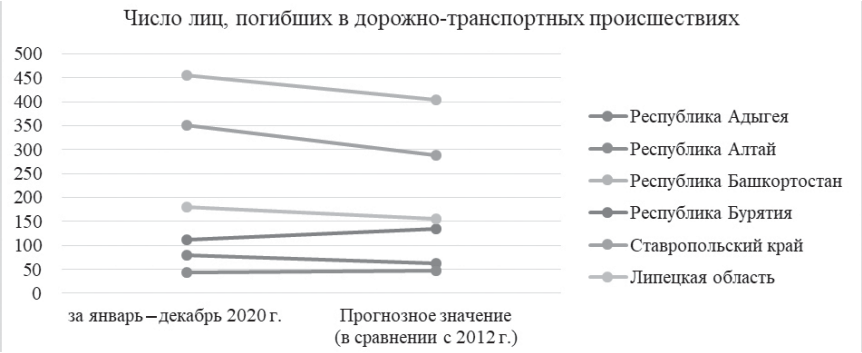


Рис. 1.25. Диаграмма наклона

Существует достаточно много специфических видов диаграмм, такие как биржевая, поверхностная, лепестковая (паутинообразная) и др. Один из вариантов – это график ***лепестковой диаграммы*** *(англ. radar chart)*, который по форме напоминает колесо, где каждый набор переменных отображается вдоль отдельной оси (рис. 1.26).

Рис. 1.26. Лепестковая диаграмма

Если диаграмма наклона может демонстрировать соотношения набора сущностей в два момента времени, то паутинообразная диаграмма намного больше. Диаграммы данного типа демонстрируют максимальную наглядность, а также хорошо подходят для иллюстрации изменчивости показателей по нескольким блокам. Кроме этого, данный тип подойдет при необходимости отобразить на одном графике отклонение переменных от некого набора стабильных значений.

Существуют некоторые рекомендации при выборе того или иного типа диаграммы. Например, процедуру выбора определенного способа визуализации иллюстрирует следующая схема (рис. 1.27)[1].

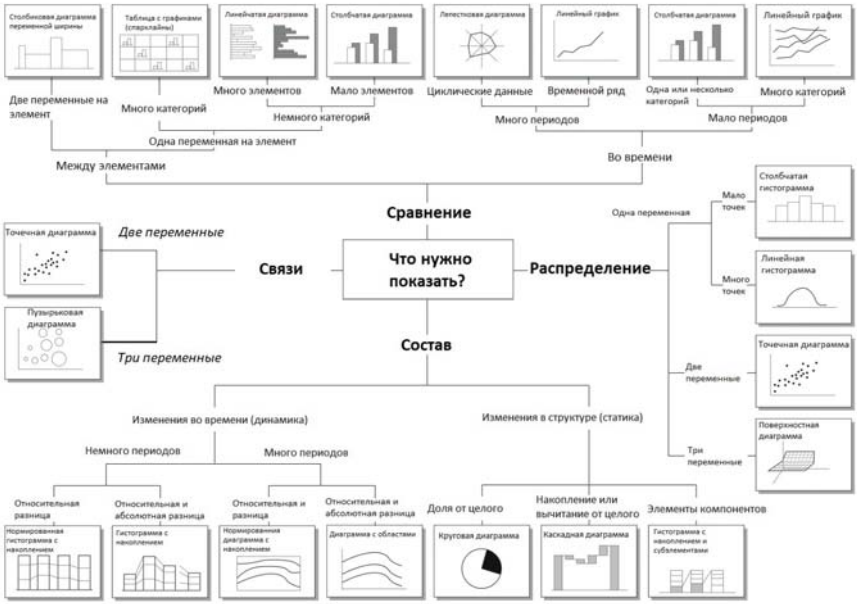


Рис. 1.27. Выбор типа диаграммы

Рассмотренные в данной главе методы визуализации данных являются общими для всех сфер деятельности[2]. Вместе с тем при решении задач научного исследования необходимо учитывать то,

[1] Опубликована на портале Extremepresentation.com (автор схемы A. Abela, перевод выполнен авторами учебного пособия).

[2] *Яу Н.* Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами / пер. с англ. Москва: Манн, Иванов и Фербер, 2013. 352 с.

что не существует типовых, шаблонных и стандартных задач в творческом процессе, к которым, безусловно, относится и научно-исследовательская деятельность.

1.3. Программное обеспечение

На современном этапе развития общества информационные технологии основаны на использовании разнообразных компьютерных программ анализа и визуализации данных. Как в научных исследованиях, так и в образовательном процессе предметом этой деятельности являются компьютерные файлы различных форматов – тексты, презентации, числовые таблицы, изображения и, на нынешнем этапе ускоренного развития образовательных технологий, файлы мультимедиа (аудио- и видеофайлы).

Наибольшее распространение получило прикладное (пользовательское) программное обеспечение (далее – ППО), отличительной особенностью которого является ориентация на выполнение определенных пользовательских специализированных задач. Программное обеспечение (далее – ПО) может быть классифицировано по типам, основным из которых является ПО офисного назначения: текстовые редакторы (MS Office Word, LibreOffice Writer и др.), электронные таблицы (MS Office Excel, LibreOffice Calc и др.), средства подготовки презентаций (MS Office PowerPoint, LibreOffice Impress и др.) и т. п. Кроме этого, к офисному ППО можно отнести некоторые графические редакторы (MS Paint, LibreOffice Draw и др.), а также пользовательские системы управления базами данных (ПСУБД), такие как MS Office Access, LibreOffice Base и др.

Отдельной категорией ППО являются браузеры, предназначенные для просмотра web-страниц, такие как Chrome, MS Explorer (Edge), Opera и др., которые относятся к проприетарному программному обеспечению. В классе браузеров к открытому (свободному) программному обеспечению относится Mozilla FireFox.

В отдельную группу можно вынести статистические пакеты, предназначенные для выявления явных и скрытых закономерностей в социально-экономических явлениях и процессах. Данный класс программ можно разделить на три группы: пользовательские (MS Excel, LibreOffice Calc. PSPP, Minitab), профессиональные (SPSS, Statistika, Gretl, IBM i2 Analyst's) и специализированные (Statistika Adv, SciDAVis, MATLAB, GNU Octave).

В последнее время наибольшее распространение получает ПО, основанное на облачных технологиях. Облачные вычисления (*англ.*

cloud computing) – модель обеспечения удобного сетевого доступа по требованию к некоторому общему фонду конфигурируемых вычислительных ресурсов (сетям передачи данных, серверам, средствам хранения данных, приложениям и сервисам – как вместе, так и по отдельности), которые могут быть оперативно предоставлены и освобождены с минимальными эксплуатационными затратами или обращениями к провайдеру. Наиболее крупными игроками в сфере публично-облачных вычислений являются такие гиганты как Amazon, Google, VMware, Microsoft. Наиболее часто встречающейся моделью облачных вычислений является технология «Программное обеспечение как услуга» (*англ. Software-as-a-Service, SaaS*), например, 1С: Предприятие, Microsoft Office 365 и др.

В зависимости от типа используемой лицензии программное обеспечение можно разделить на две категории (рис. 1.28): проприетарное (несвободное) и открытое (свободное)[1].

Рис 1.28. Лицензия на программное обеспечение

В первом случае в целях предотвращения использования, копирования или модификации программное обеспечение выпускается только в виде машинно-читаемых двоичных файлов, таким образом ограничивается доступ к исходному коду (закрытый исходный код). Вторая группа предполагает открытый доступ к исходным текстам программ. Данные типы программ предполагают использование GNU General Public License, то есть лицензии на свободное программное обеспечение, по которой автор передает разработанное им программное обеспечение в общественную собственность. Заметим,

[1] Данная классификация достаточно условна.

что использование открытого программного обеспечения позволяет повысить уровень информационной безопасности, так как в программном коде отсутствует бэкдор (*англ. back door* – тайный ход), то есть недекларированные (недокументированные) возможности (НДВ). Вместе с тем данный тип программного обеспечения также не застрахован от уязвимости «нулевого дня» (*англ. Zero-day exploit,* 0-*day attack*). Некоторые примеры разных видов программного обеспечения приведены в таблице 1.3.

Таблица 1.3. Программное обеспечение

Назначение	**Проприетарное**	**Открытое**
Графические и видеоредакторы	Microsoft Visio	Dia
	Adobe Illustrator, CorelDraw	Inkscape
	Movie Maker, Movavi	AviDeMux
Коммуникации	MS IE, Opera, Chrome	Mozilla Firefox
	MS Outlook, Becky, The Bat!	Mozilla Thunderbird
	ICQ, Trillian, MSN Messenger	Pidgin
Медиа	Windows Media Player	VLC media player
Антивирусы	Kaspersky Antivirus	ClamWin
Архиваторы	WinZip, WinRar	7-Zip
Файловый менеджер	Total Commander	MuCommander
Офисные приложения	Microsoft Office	OpenOffice, LibreOffice
Просмотр и печать документов	Adobe Acrobat	Sumatra

Безусловно, решать задачи визуализации статистических данных можно при помощи общедоступного программного обеспечения. Однако зачастую удобнее производить анализ данных и визуализацию при помощи специализированных инструментов, таких как статистические пакеты анализа. Условно данный вид ПО можно разделить на три большие группы: пользовательские, профессиональные и специализированные (таблица 1.4).

Таблица 1.4. Статистические пакеты анализа

Профессиональные	Пользовательские	Специализированные
SPSS	MS Excel	Statistika Adv
Statistika	LibreOffice Calc	SciDAVis
Gretl	PSPP	MATLAB
IBM i2 Analyst's	Minitab	GNU Octave

В свою очередь указанное ПО может быть также разделено на проприетарное и открытое. Данный перечень далеко не исчерпывающий и не включает многие хорошо известные пакеты анализа, такие как SOFA Statistics, STATA, EViews, Prognoz Platform, EViews, SalStat Statistics Package и др. Кроме этого, известно достаточно много надстроек и расширений MS Excel (StatTools, XLSTAT, StatPlus, XLR и др.), позволяющих реализовать множество статистических методов. Далее рассмотрим несколько статистических пакетов анализа, позволяющих в том числе и визуализировать данные.

GNU PSPP – программа для статистического анализа данных. Это бесплатный аналог проприетарной программы SPSS IBM и очень похожа на нее. PSPP[1] позволяет рассчитать описательную статистику, произвести различные статистические тесты, дисперсионный анализ, линейную и логистическую регрессию, кластерный анализ, анализ надежности и факторный анализ, непараметрические тесты и многое другое. PSPP предназначена для статистиков, социологов и обучающихся, которым требуется быстрый и удобный анализ статистических данных. К плюсам следует отнести ее бесплатность и свободное распространение, простоту в изучении и освоении. К недостаткам – ограниченный функционал методов и статистических тестов, отсутствие поддержки методов Data Mining[2] и AI (*англ. Artificial Intelligence* – искусственный интеллект).

GRETL (*англ. Gnu Regression, Econometrics and Time-series Library*) – это открытый, свободный и бесплатный кросс-платформенный программный пакет для эконометрического анализа. Отличительной особенностью GRETL является простой и интуитивно понятный мультиязычный интерфейс, множество методов оценивания (OLS, MLE, GMM)[3], обширный инструментарий для анализа временных рядов

[1] Данное название в виде аббревиатуры не имеет официальной расшифровки.

[2] Обобщенная группа методов «добычи» данных; глубинный анализ данных.

[3] Методы оценки неизвестных параметров моделей: OLS (Ordinary Least Squares) – метод наименьших квадратов; GMM (Generalized Method of Moments) – обобщенный метод моментов; MLE (Maximum Likelihood Estimation) – метод максимального правдоподобия.

(ARIMA, GARCH, ADL, VAR и др.)[1], поддержка моделей с ограниченной зависимой переменной (logit, probit, tobit и т. д.) и др. К недостаткам GRETL нужно отнести наличие только эконометрических методов, хотя и очень обширных, ориентацию на специалистов в области анализа данных (аналитиков) и некоторую сложность в освоении.

IBM i2 Analyst's Notebook предоставляет широкие возможности анализа данных, которые помогают преобразовать наборы разрозненных данных в высококачественную информацию. i2 Analyst's Notebook позволяет обрабатывать структурированные и неструктурированные данные, помогая аналитикам создавать единую аналитическую картину. Система обладает интуитивно понятным пользовательским интерфейсом, что сокращает время обучения пользователей, имеет широкий спектр инструментов визуального анализа, призванных помочь пользователям обнаружить ключевые связи, отношения, события, закономерности и тенденции, также включает в себя возможности анализа социальных сетей. Данные, полученные в результате анализа, можно использовать в виде интуитивно понятных визуальных диаграмм.

MS Excel – это программа, входящая в состав пакета Microsoft Office и предназначенная для работы с электронными таблицами. MS Excel – это универсальный инструмент, позволяющий обрабатывать разные форматы данных, а также хранить, организовывать и анализировать информацию. MS Excel позволяет работать с числовыми данными, текстом, создавать графики и различные варианты диаграмм. Помимо графиков и диаграмм MS Excel позволяет создавать различные типовые схемы за счет технологии SmartArt. Для решения нетривиальных (нетиповых) задач можно использовать инструменты для работы с фигурами, что значительно увеличивает возможности визуализации данных.

Начиная с версии MS Excel 2016, появилось несколько новых типов диаграмм: дерево (*англ. Treemap*), солнечные лучи (*англ. Sunburst*), ящик с усами (*англ. Box&Whisker*), каскадная (*англ. Waterfall*), обычная гистограмма (*англ. Histogramm*) и гистограмма типа Парето (*англ. Pareto*).

Практически аналогичным функционалом обладает программа ***Calc***, входящая в пакет LibreOffice. LibreOffice – это кроссплатформенный, свободно распространяемый офисный пакет с открытым исходным кодом, основанный на OpenOffice.org. Основные возможности Calc включают в себя выполнение достаточно

[1] ARIMA – autoregressive integrated moving average; GARCH – Generalized AutoRegressive Conditional Heteroscedasticity; ADL – autoregressive distributed lags; VAR – Vector AutoRegression.

сложных вычислений, позволяют организовывать, хранить и обрабатывать данные, создавать разнообразные диаграммы, а также открывать, редактировать и сохранять файлы в формате Microsoft Excel. Заметим, что при работе с электронными таблицами формата MS Excel наблюдаются некоторые проблемы с совместимостью.

Отдельно стоит обсудить ***Microsoft PowerPoint***, которая широко используется для подготовки и редактирования презентаций в различных областях человеческой деятельности, особенно в сфере науки и образования. Например, научные доклады во время защиты диссертации обычно сопровождаются презентацией. В образовательных организациях профессорско-преподавательский состав постоянно использует презентации во всех формах обучения.

Наиболее универсальным инструментом для подготовки числовых данных и их визуализации является таблица Microsoft Excel или ее аналоги с открытой лицензией LibreOffice Calc, OpenOffice Calc. Таблицы и диаграммы, созданные в этих программах, можно легко скопировать в текстовые документы в форматах doc, docx, rtf. Сами файлы, управляемые электронными таблицами, имеют множество различных расширений, но наиболее распространенными являются xls, xlsx и формат данных открытой электронной таблицы, не связанный с продуктами Microsoft – ods (*англ. Open Document Spreadsheet*).

Помимо графиков, диаграмм и типовых схем SmartArt, офисные пакеты позволяют создавать различные нетиповые схемы за счет технологии работы с иллюстрациями. Рассмотрим несколько нестандартных схем. На следующем рисунке может быть представлено схематическое описание концептуального правового или нормативного документа (рис. 1.29).

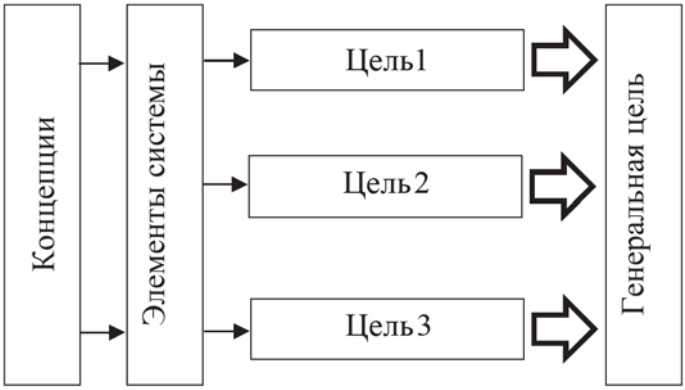


Рис. 1.29. Схематическое описание концептуального документа

Далее представлено схематическое отображение системы взаимосвязей элементов с целевыми индикаторами (рис. 1.30).

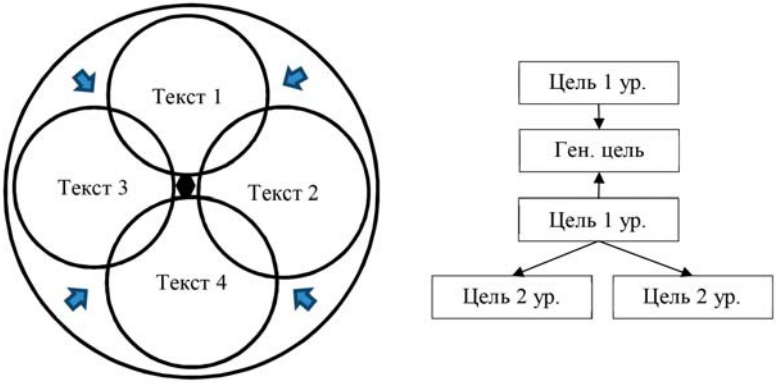


Рис. 1.30. Схематическое отображение системы взаимосвязей элементов с целевыми индикаторами

На следующем рисунке представлено схематическое описание объекта с системой целевых показателей (рис. 1.31).

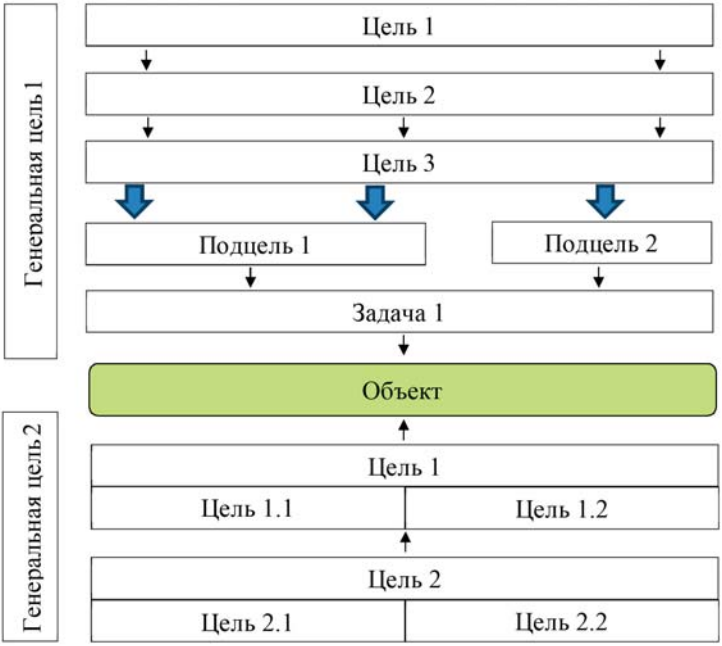


Рис. 1.31. Схематическое описание объекта с системой целевых показателей

Одной из наиболее часто используемых схем является описание иерархической системы. В большинстве случаев исследователь имеет дело с различными иерархическими системами, как типовыми, так и не типовыми. В качестве примера на рисунке представлена система управления Министерства внутренних дел Российской Федерации (рис. 1.32).

Рис. 1.32. Система управления Министерства внутренних дел Российской Федерации

В теории управления часто применяется следующая схема, иллюстрирующая систему управления (рис. 1.33).

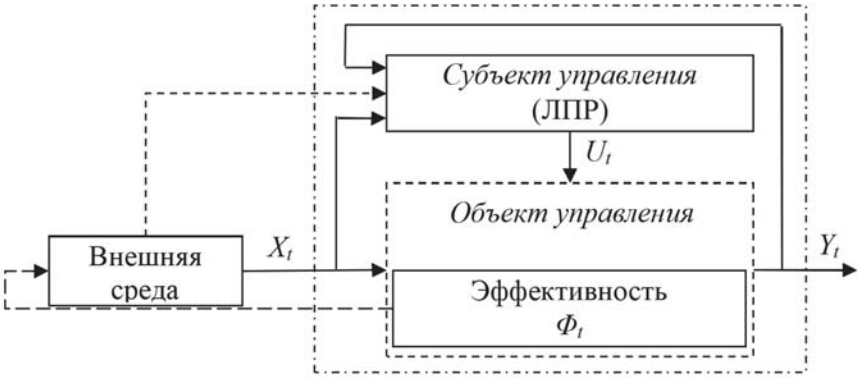


Рис. 1.33. Общая система управления

Эти схемы относятся к категории ***информационной визуализации*** и хорошо дополняют статистические данные, визуализируемые при помощи различных графиков и диаграмм (визуализация данных).

Глава 2. Основы математической статистики и анализ временных рядов

В главе рассматриваются основные понятия математической статистики, корреляционный анализ и дисперсия, а также компьютерные технологии анализа временных рядов[1].

2.1. Основные показатели математической статистики

Математическая статистика – это раздел прикладной математики, изучающий методы нахождения свойств случайных величин на основе результатов наблюдений и экспериментов. Математическая статистика основана на теории вероятностей и, в свою очередь, служит базой для разработки методов обработки и анализа статистических результатов в определенных областях человеческой деятельности.

Первая задача математической статистики – указать методы сбора и группировки статистической информации, полученной в результате наблюдений или в результате специальных экспериментов.

Вторая задача математической статистики – разработать методы анализа статистических данных в зависимости от цели исследования.

Современная математическая статистика разрабатывает способы определения количества тестов, необходимых до начала исследования (планирование тестирования) и во время исследования (последовательный анализ). Ее можно определить как науку о принятии решений в условиях неопределенности.

Таким образом, задача математической статистики заключается в создании методов сбора и обработки статистических данных для продвижения, подтверждения или опровержения обоснованных научных гипотез.

Генеральная совокупность – это совокупность всех мыслимых объектов определенного типа, над которыми производятся наблюдения с целью получения определенных значений определенной случайной величины.

Выборка (выборка совокупности) – это набор объектов, выбранных случайным образом из генеральной совокупности.

Последовательность вариаций в порядке возрастания называется ***серией вариаций***.

Дискретный статистический ряд – это ряд вариантов с соответствующими им частотами.

---

[1] Информационные технологии управления и организация защиты информации: учебник / В. В. Баранов и др. Москва: Академия управления МВД России, 2018. 456 с.

Характеристики дискретного статистического ряда: диапазон вариации, режим – вариант с наибольшей частотой, медиана – значение случайной величины в середине ряда.

Среднее арифметическое – это условное значение, которого на самом деле не существует. Есть действительно общая сумма. Следовательно, среднее арифметическое не является характеристикой наблюдения, а характеризует серию в целом.

Нормальное распределение, также называемое распределением ***Гаусса***, представляет собой распределение вероятностей, которое играет решающую роль во многих областях знаний. Социальная сфера не является исключением, где большинство ценностей обычно распределяется нормально.

Выявление закономерностей, которым подвержены случайные массовые явления, основано на исследовании статистических данных методами теории вероятностей и математической статистики[1].

Прикладное программное обеспечение MS Excel, входящее в пакет MS Office, позволяет осуществлять анализ данных при помощи двух основных подходов. Первый позволяет использовать специальный пакет анализа данных, а второй использует статистические функции, вводимые «вручную». Здесь и далее будем приводить как первый подход, так и второй.

Включение функции «Анализ данных» в MS Excel осуществляется в меню: Файл\Параметры\Надстройки, где следует нажать кнопку «Перейти» (рис. 2.1).

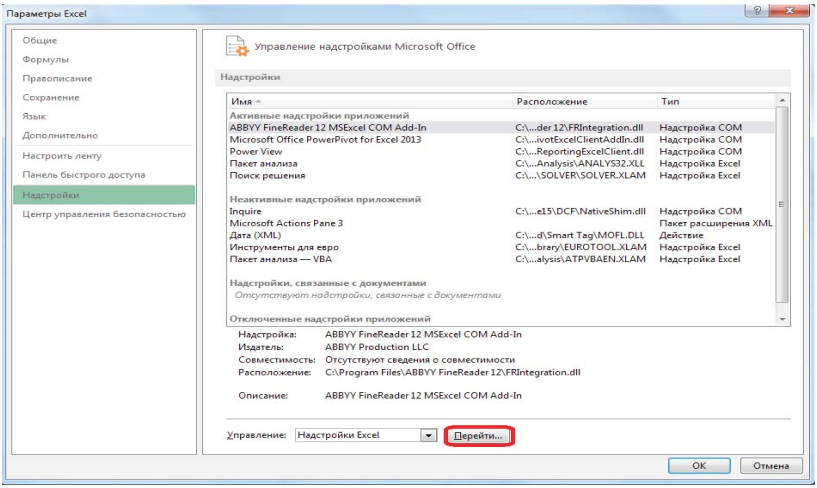


Рис. 2.1. Управление надстройками MS Excel

[1] *Носко В. П.* Эконометрика для начинающих. Москва, 2005. 379 с.

В появившемся окне установить галочку напротив пункта «Пакет анализа» и нажать кнопку «ОК» (рис. 2.2).

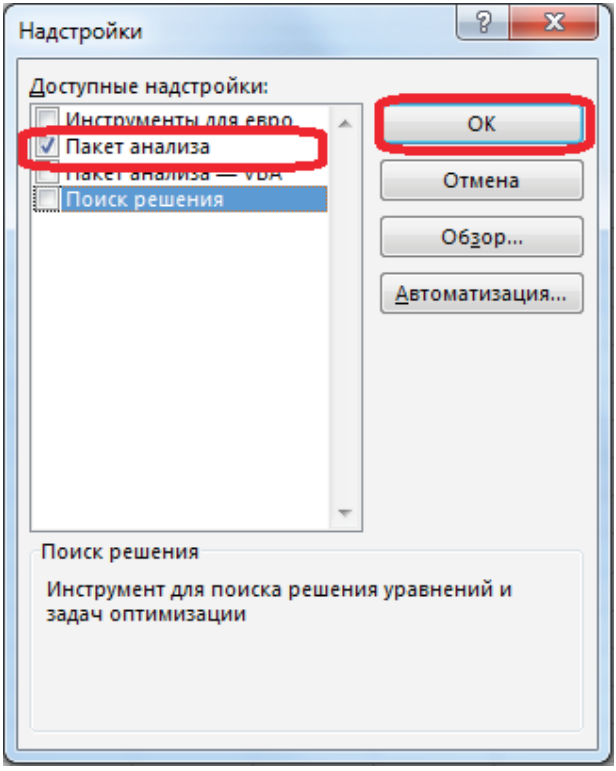


Рис. 2.2. Надстройки MS Excel

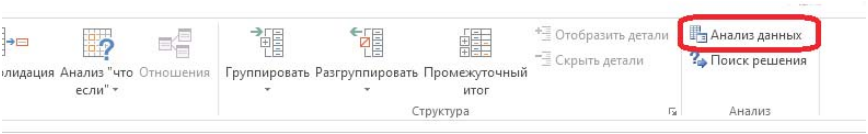


Рис. 2.3. Анализ данных MS Excel

Для расчета статистических показателей в LibreOffice Calc будут использоваться встроенные функции. Выбираем в главном меню пункт «Данные», далее «Статистика», после чего появится дополнительное меню с различными статистическими инструментами (рис. 2.4).

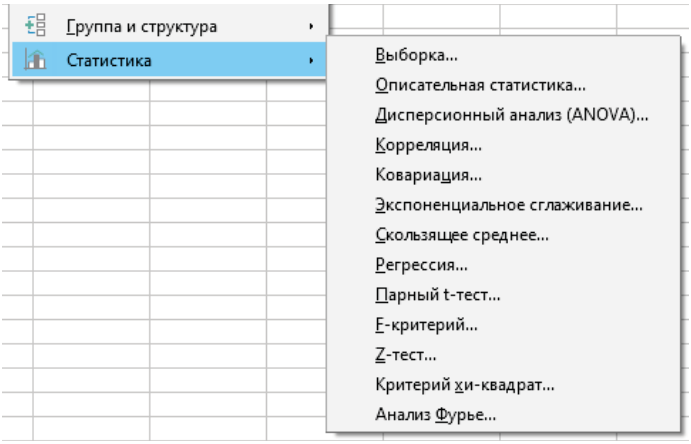


Рисунок 2.4. Анализ данных LibreOffice Calc

Заметим, что большинство функций LibreOffice и MS Excel совместимы, то есть наименование и параметры одинаковы. Это сделано для обеспечения совместимости двух популярных пакетов. Поэтому в большинстве случаев файлы формата xls и ods с автоматизированными расчетами будут без особых проблем работать в LibreOffice Calc и MS Excel.

Для расчета *описательной статистики* используем пакет анализа, встроенный в MS Excel. Выбираем в главном меню пункт «Данные», нажимаем кнопку «Анализ данных», в появившемся окне выбираем пункт «Описательная статистика» и нажимаем кнопку «ОК» (рис. 2.5).

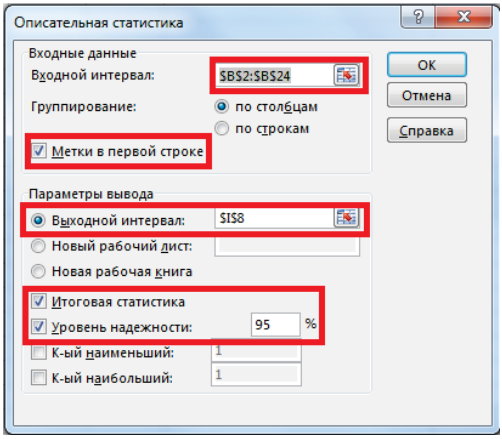


Рис. 2.5. Описательная статистика Excel

Задаем соответствующие параметры, и после нажатия кнопки «ОК» система выдаст результат следующего содержания (рис. 2.6):

Преступления	
Среднее	40188,04545
Стандартная ошибка	1223,756207
Медиана	40326,5
Мода	#Н/Д
Стандартное отклонение	5739,925398
Дисперсия выборки	32946743,57
Эксцесс	-1,496075474
Асимметричность	0,066451648
Интервал	16488
Минимум	31726
Максимум	48214
Сумма	884137
Счет	22
Уровень надежности(95,0%)	2544,94035

Рис. 2.6. Результаты описательной статистики Excel

Для расчета описательной статистики в LibreOffice Calc активируем главное меню и выберем пункт «Данные» – «Статистика» – «Описательная статистика» (рис. 2.7).

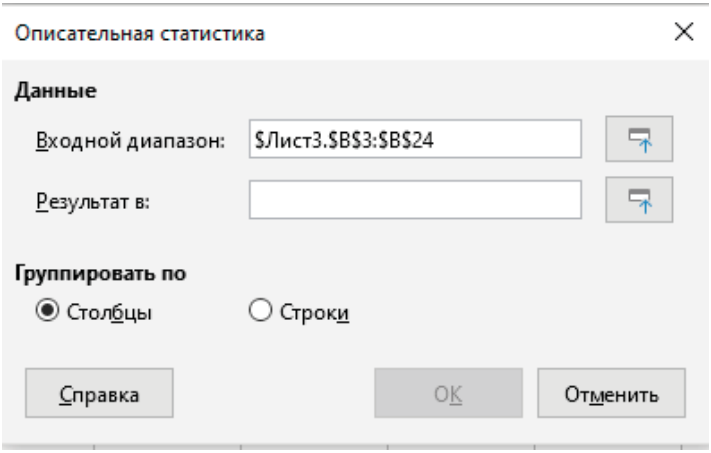


Рис. 2.7. Описательная статистика Calc

В появившемся окне задаем входной диапазон и результирующий диапазон, в нашем случае это B3:B24 и ячейка E2[1]. Задание диапазонов в Calc осуществляется нажатием кнопки  . После нажатия кнопки «ОК» получим следующий результат (рис. 2.8):

Среднее	40188,0454545455
Среднеквадратическое отклонение	1223,75620658216
Мода	#ЗНАЧ!
Медиана	40326,5
Первый квартиль	34925,5
Третий квартиль	45149,75
Дисперсия	32946743,5692641
Среднеквадратическое отклонение	5739,92539753472
Эксцесс	-1,49607547378158
Асимметрия	0,066451648385728
Диапазон	16488
Минимум	31726
Максимум	48214
Сумма	884137
Количество	22

Рис. 2.8. Результаты описательной статистики Calc

В отличие от MS Excel большинство инструментов анализа данных в LibreOffice Calc не позволяют задать подписи данных (метки).

Рассмотрим один из наиболее распространенных инструментов статистического анализа – корреляционный. Для этого обратимся к процедуре расчета коэффициента корреляции, который показывает меру тесноты взаимосвязи между двумя и более случайными величинами и рассчитывается по формуле:

$$r_{xy} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(y_i - \bar{y})^2}}; \; -1 < r_{xy} < 1.$$

Рассчитаем коэффициент корреляции «вручную» (рис. 2.9).

	A	B	C	D	E	F	G	H	I	J
1		y	x	$x_i - x_{ср}$	$y_i - y_{ср}$	$(x_i - x_{ср})*(y_i - y_{ср})$	$(x_i - x_{ср})^2$	$(y_i - y_{ср})^2$		
2	1	16 842	40,6	7,3	1648,0	11983,3	52,9	2715904,0		
3	2	15 895	33,2	-0,1	701,0	-90,1	0,0	491401,0		
4	3	14 888	29,5	-3,8	-306,0	1171,5	14,7	93636,0		
5	4	14 930	32,2	-1,1	-264,0	297,9	1,3	69696,0		
6	5	14 296	32,5	-0,8	-898,0	744,1	0,7	806404,0		
7	6	15 792	32,8	-0,5	598,0	-316,1	0,3	357604,0		
8	7	13 715	32,5	-0,8	-1479,0	1225,5	0,7	2187441,0		
9	СР	15194,0	33,3							
10	Σ	106358	233,3			15016,1	70,47428571	6722086	21765,44	0,689906

Рис. 2.9. Расчет корреляции

[1] Система автоматически расширит указанный диапазон до нужного размера.

Для автоматизированного расчета коэффициента корреляции используем пакет анализа, встроенный в MS Excel. Выбираем в главном меню пункт «Данные» и нажимаем кнопку «Анализ данных», появится окно с инструментами анализа, в котором выбираем пункт «Корреляция», нажимаем кнопку «ОК» и в появившемся окне (рис. 2.10) задаем следующие параметры:

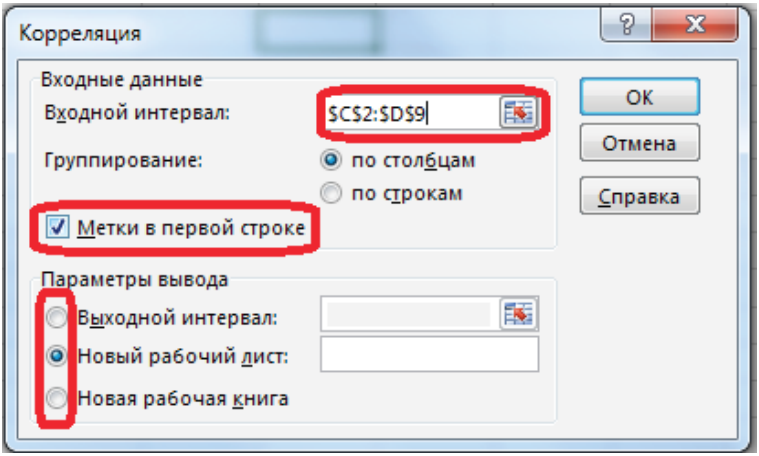


Рис. 2.10. Параметры расчета корреляции

Выбираем диапазон данных нажатием кнопки  MS Excel[1]. Указываем исходные данные, представленные ниже, и наличие меток[2].

Преступления	Безработица
16 842	40,6
15 895	33,2
14 888	29,5
14 930	32,2
14 296	32,5
15 792	32,8
13 715	32,5

Кроме этого, можно задать и параметры вывода (по умолчанию установлен вывод «Новый рабочий лист», и система направит

[1] Здесь и далее нажатие данной кнопки позволяет задать диапазон данных.
[2] Метки – это наименования полей данных.

результаты во вновь создаваемый следующий лист). После нажатия кнопки «ОК» система выдаст результат следующего содержания:

	Преступления	**Безработица**
Преступления	1	
Безработица	0,689905772	1

Полученный коэффициент корреляции равен ≈ 0,7 (0,689905772), который можно интерпретировать по шкале Чеддока.

Количественная мера тесноты связи	**Качественная мера силы связи**
0,1–0,3	Слабая
0,3–0,5	Умеренная
0,5–0,7	Заметная
0,7–0,9	Высокая
0,9–0,99	Весьма высокая

Таким образом, можно говорить о наличии положительной высокой связи переменных «Преступления» и «Безработица». При расчете коэффициента корреляции большое значение имеет знак. Так, при положительном знаке говорят о положительной (прямой) связи, а при отрицательном – об отрицательной (обратной) связи (рис. 2.11).

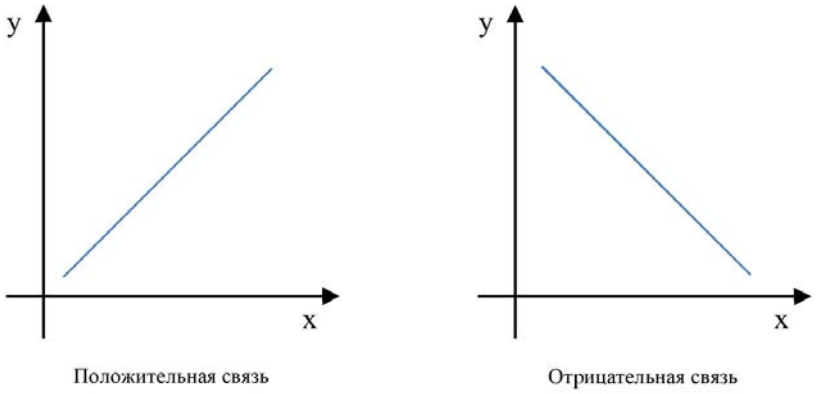


Рис. 2.11. Корреляционная связь

Аналогичные результаты получим, применяя «ручной» расчет коэффициента корреляции. Для этого активируем ячейку, в которую необходимо поместить результат, и нажимаем кнопку «Вста-

вить функцию». В появившемся окне выберем из выпадающего списка «Категория» группу функций «Статистические» (рис. 2.12)[1].

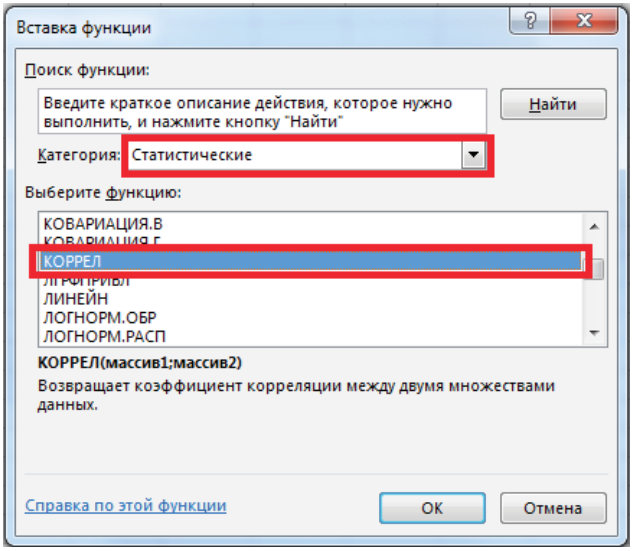


Рис. 2.12. Вставка функции «коррел»

В списке функций найдем функцию «КОРРЕЛ», и после нажатия кнопки «ОК» появится окно ввода аргументов функции следующего вида (рис. 2.13), в котором необходимо задать «Массив 1» и «Массив 2».

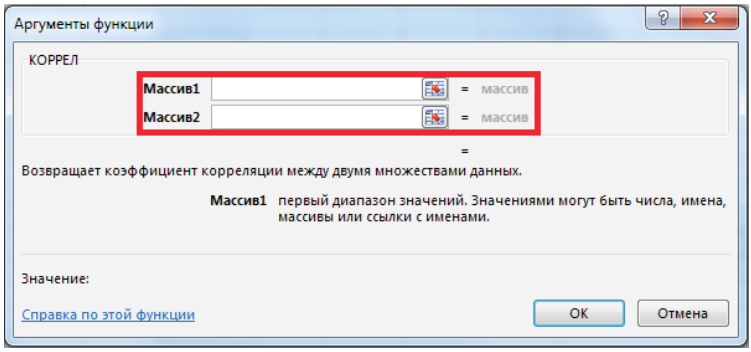


Рис. 2.13. Выбор аргументов функции

[1] По умолчанию установлена категория «10 недавно использовавшихся», и если в списке требуемой функции нет, то нужно выбрать соответствующую категорию или «полный алфавитный перечень».

После нажатия кнопки «ОК» в целевой ячейке появится результат следующего вида (рис. 2.14):

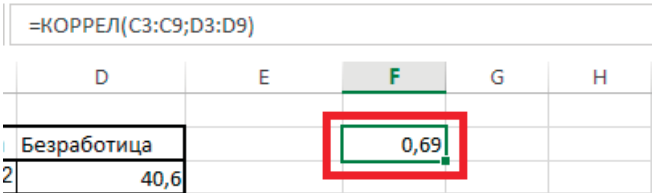


Рис. 2.14. Результат функции

Теперь рассмотрим процедуру расчета коэффициента корреляции в LibreOffice Calc. В ней, как и в MS Office, рассчитать коэффициент корреляции можно как в «ручном» режиме, так и через интерфейс, то есть диалоговое окно.

Инструменты анализа данных в LibreOffice Calc находятся также в главном меню, в котором выбираем пункт «Данные» – «Статистика» – «Корреляция» (рис. 2.15).

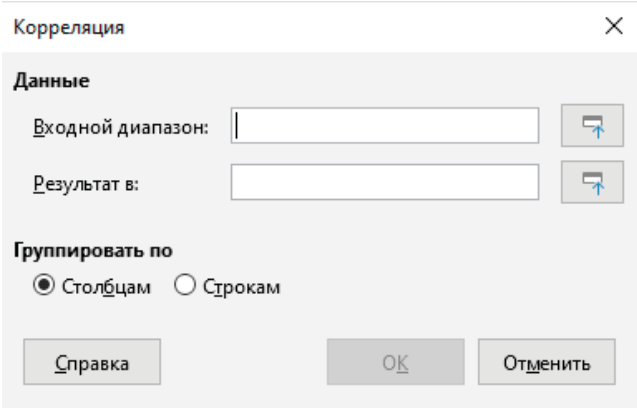


Рис. 2.15. Форма «Корреляция»

Запустится окно, в котором задаем входной диапазон и результирующий диапазон, в нашем случае это C3:D9 и ячейка F2. После нажатия кнопки «ОК» получим результат следующего вида (рис. 2.16):

Корреляции	Столбец 1	Столбец 2
Столбец 1	1	
Столбец 2	0,68990577	1

Рис 2.16. Коэффициент корреляции

В «ручном» режиме можно рассчитать коэффициент корреляции, воспользовавшись функцией «КОРРЕЛ», которая есть как в MS Excel, так и в LibreOffice Calc. Активируем нужную ячейку, куда необходимо поместить результат, и нажмем кнопку «Мастер функций». В появившемся окне находим функцию «КОРРЕЛ», выбираем ее и нажимаем кнопку «Далее». В следующем окне введем следующие параметры: «Данные 1» – это имеющиеся показатели *y*, то есть диапазон данных C3:C9; «Данные 2» – это имеющиеся показатели *x*, то есть диапазон номеров ряда D3:D9. После нажатия кнопки «ОК» в целевую ячейку будет помещен результат – коэффициент корреляции (≈ 0,69).

2.2. Дисперсионный анализ и временные ряды

Под дисперсионным анализом понимается статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния[1]. Дисперсионный анализ (*англ. Analysis of variance* – ANOVA) характеризует «разброс» значений переменных[2].

Одним из ключевых элементов дисперсионного анализа, да и всей математической статистки, является понятие статистической гипотезы. Статистическая гипотеза – это некоторое предположение относительно исследуемой генеральной совокупности[3].

Основная (нулевая) гипотеза H_0 – это гипотеза, которой мы придерживаемся, пока наблюдения не заставят признать обратное. Ей всегда сопутствует **альтернативная** (конкурирующая) гипотеза H_1.

Статистические методы не позволяют доказать гипотезу. На основе наблюдений мы можем опровергнуть гипотезу.

---

[1] *Шеффе Г.* Дисперсионный анализ. Москва: Наука: Главная редакция физико-математической литературы, 1980. 512 с.

[2] *Новиков Д. А.* Статистические методы в педагогических исследованиях (типовые случаи). Москва: МЗ-Пресс, 2004. 67 с.

[3] *Фадеева Л. Н.* Теория вероятностей и математическая статистика: учебное пособие / Л. Н. Фадеева, А. В. Лебедев. 2-е изд., перераб. и доп. Москва: Эксмо, 2010. 496 с.

Процедура проверки гипотезы состоит из нескольких этапов (рис. 2.17):

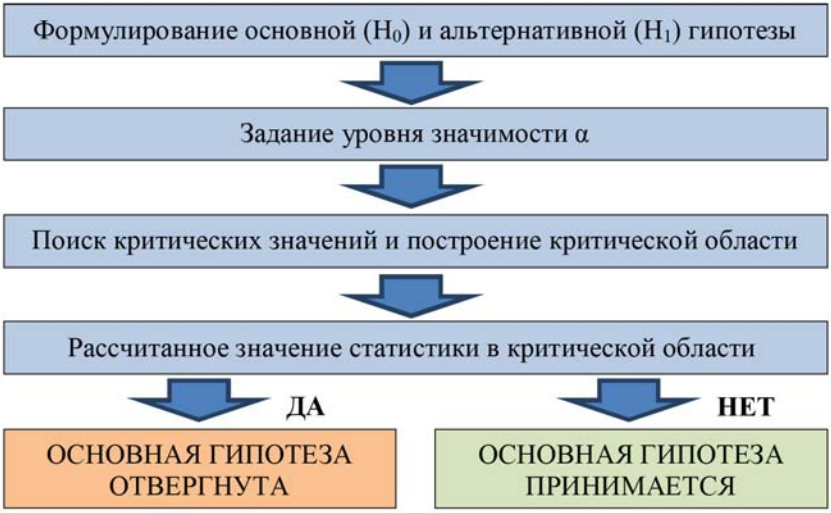


Рис. 2.17. Процедура проверки гипотез

Для определения ситуации, при которой необходимо опровергнуть гипотезу, используются понятия ошибки первого и второго родов. **Ошибка первого рода** – это ситуация, когда H_0 отвергается, но на самом деле она верна. **Ошибка второго рода** – это ситуация, когда H_0 принимается, хотя она неверна.

Чаще всего **уровень значимости или вероятность** ошибки первого рода обозначается греческой буквой α, а вероятность ошибки второго рода обычно обозначается буквой β. Уменьшение ошибки первого рода приводит к увеличению ошибки второго рода и наоборот. В научных исследованиях α чаще всего берут равным 0,1 (10 %), 0,05 (5 %) или 0,01 (1 %). В зависимости от того, какая гипотеза является основной, а какая альтернативной, ошибки первого и второго родов меняются местами.

Случайная величина, построенная по наблюдениям для проверки нулевой гипотезы, называется статистикой критерия и чаще всего обозначается Z. Для решения задачи выбирается уровень значимости и статистика критерия, на основе которой делается вывод о справедливости гипотезы. При справедливости основной гипотезы известно, с какой вероятностью какое значение принимает статистика критерия. Если эта вероятность очень маленькая, то гипотезу придется отвергнуть.

Мощностью критерия называется вероятность не совершить ошибку второго рода, то есть $1 - \beta$. А наиболее мощным критерием из всех

критериев с уровнем значимости α называется тот, который обладает наибольшей мощностью.

Критической областью называется область значений статистики критерия, при которых отвергается H_0. Критические значения – это граница критической области. Существует три вида критических областей: левосторонняя (рис. 2.18), правосторонняя (рис. 2.19) и двусторонняя (рис. 2.20).

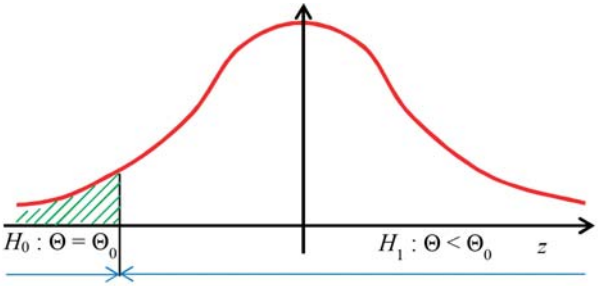


Рис. 2.18. Левосторонняя критическая область

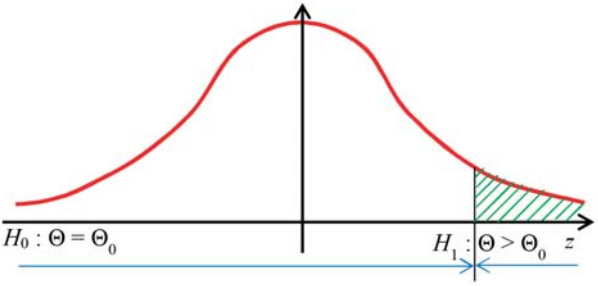


Рис. 2.19. Правосторонняя критическая область

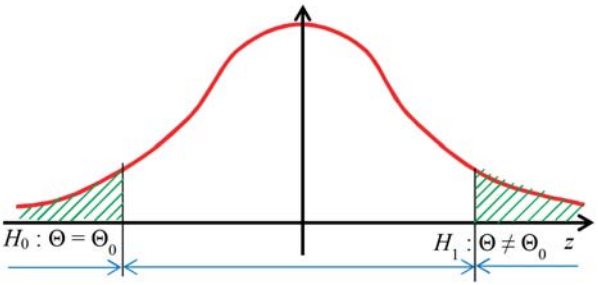


Рис. 2.20. Двусторонняя критическая область

Минимальный уровень значимости[1] (P–value) – это минимальное значение α, при котором основная гипотеза еще отвергается, обычно принимается на уровне значимости 1 %, 5 % или 10 %[2].

Далее рассмотрим элементы **анализа временных рядов**. В теории выделяют следующие элементы временного ряда[3]: регулярную компоненту, включающую в себя тренд (T), сезонность (S) и цикличность (C), а также случайную компоненту (ε). В основном применяется либо аддитивная форма[4] представления динамического ряда (T+S+C+ε), либо мультипликативная (T·S·C·ε). В сокращенном виде аддитивную модель можно записать как $y_{add} = \sum_{i=1}^{n} \beta_n x_n$, а мультипликативную как $y_{mlt} = \prod_{i=1}^{n} \beta_n x_n$.

В информационно-аналитической деятельности[5] принято выделять прогнозы со следующей периодичностью: за текущий период (ежедневно, ежемесячно), за отчетный год и прогнозы за длительный период (год и более). Ясно, что сезонность может быть только в пределах года, а циклическая компонента может присутствовать в достаточно длинных рядах. Выделить цикличность в исследовании преступности достаточно сложно, поэтому в практической деятельности данную компоненту временного ряда обычно не учитывают[6]. Таким образом, аддитивная модель временного ряда без учета цикличности будет представлена в виде (T+S+ε).

Метод экстраполяции основан на распространении тенденций развития объекта, процесса или явления в ретроспективе на будущее состояние объекта прогнозирования. Метод **статистической экстраполяции временного ряда** является одним из основных методов прогнозирования. Вместе с тем к недо-

---

[1] В прикладных эконометрических исследованиях уровень значимости обычно указывается в виде звездочки «*» рядом с коэффициентом: * - 10%-ый уровень значимости; ** - 5%-ый уровень значимости; *** - 1%-ый уровень значимости. По умолчанию применяется 5%-ый уровень значимости.

[2] *Новиков Д. А., Новочадов В. В.* Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: ВолГМУ, 2005. 84 с.

[3] Здесь и далее временной ряд и динамический рассматриваются как синонимы.

[4] Здесь и далее рассматривается только аддитивная форма представления временного ряда.

[5] Вопросы организации информационно-аналитической работы в управленческой деятельности органов внутренних дел: приказ МВД России от 26 сентября 2018 г. № 623.

[6] Заметим, что в отличие от экономических циклов (циклы Дж. Китчина, К. Жюгляра, С. Кузнеца и волны Н. Кондратьева) в исследовании социальных систем (преступности) нет единого мнения о наличии цикличности.

статкам данного метода можно отнести невозможность учесть сезонность и цикличность.

Рассмотрим методику аналитического выравнивания временного ряда (подбора аналитической функции). Основная идея метода аналитического выравнивания временного ряда заключается в применении рассмотренного ранее метода наименьших квадратов (далее – МНК) для подбора функции, наилучшим образом описывающей объект прогнозирования.

Наиболее простым и распространенным видом функции является линейная, в которой есть два коэффициента. Коэффициент сдвига, или свободный член, отражает точку пересечения с осью ординат, коэффициент при параметре x показывает наклон графика. В общем виде линейный тренд можно представить как: $y_t=\alpha+\beta x_t$ где y – значение, t – номер ряда, α и β –неизвестные параметры, оцениваемые МНК.

Для подбора аналитической функции с помощью MS Excel на первом этапе нужно построить график, для чего выделяем набор данных (рис. 2.21).

		Данные	Прогноз
1	2009	110565	
2	2010	130848	
3	2011	256007	
4	2012	271631	
5	2013	322181	
6	2014	353879	
7	2015	385152	
8	2016	468235	
9	2017	517694	
10	2018	557672	

Рис. 2.21. Выбор данных

В меню выбираем пункт «Вставка» – «Вставить график» – «График», система вставит график, и после нажатия правой клавишей мыши на графике появится контекстное меню следующего вида (рис. 2.22):

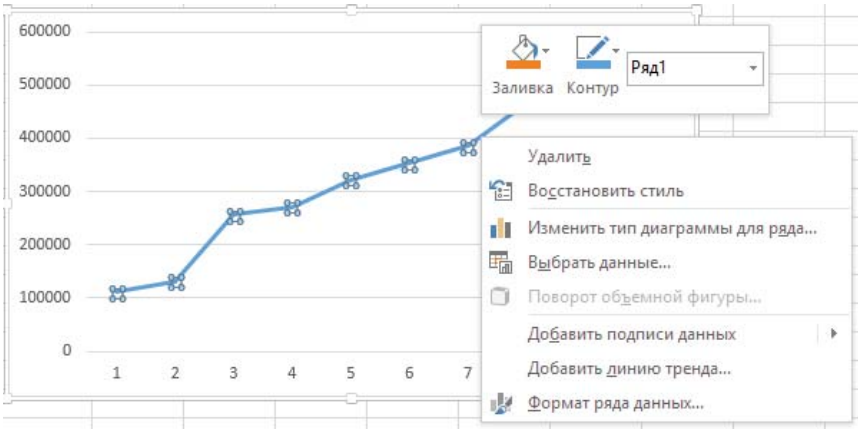


Рис. 2.22. Добавление линии тренда

Далее выбираем пункт «Добавить линию тренда», после чего появится окно справа следующего вида (рис. 2.23):

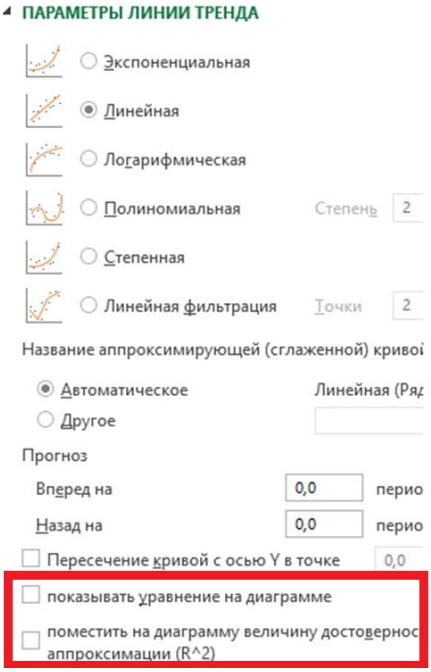


Рис. 2.23. Параметры линии тренда

В появившемся окне выберем пункт «Линейная» (она будет выбрана по умолчанию) и поставим галочки напротив пунктов «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2)»[1]. На графике (рис. 2.24) появятся следующие элементы (линия тренда, уравнение и R^2):

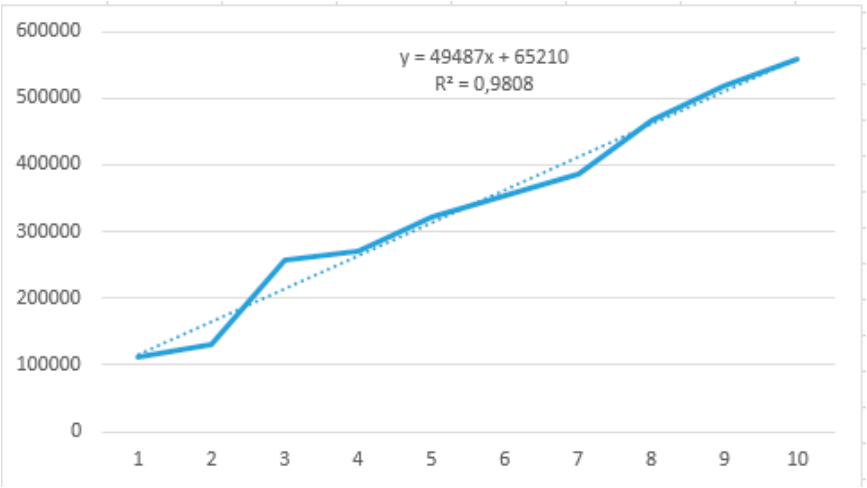


Рис. 2.24. Линия тренда

Рассмотрим вкратце полученные результаты и их значение:

1. Линия тренда – это теоретическая линия, подобранная к эмпирическим данным на основе МНК;

2. Уравнение – линейное уравнение вида $y=bx+a$, где a и b – неизвестные параметры[2], a – свободный член, который показывает точку пересечения с осью ординаты, а коэффициент b (коэффициент при x) показывает наклон графика. R^2 – коэффициент аппроксимации[3], принимает значения от 0 до 1 и показывает степень приближения теоретических (рассчитанных значений) значений $\hat{y}$ к эмпирическим значениям y. В нашим примере уравнение y=49487·x+65210, а $R^2 \approx 0{,}98$, что является довольно высоким значением[4]. X – это номер

[1] То есть коэффициент корреляции, возведенный в квадрат.

[2] Очевидно, что свободный член может располагаться как в конце, так и в начале уравнения. Заметим, что в литературе чаще всего встречается классическое расположение свободного члена в начале уравнения y=a+b*x.

[3] Иногда его также называют коэффициентом детерминации R^2.

[4] В практической деятельности такое высокое значение коэффициента встречается достаточно редко, обычно он намного ниже.

ряда, на получившемся графике он указан по оси абсцисс. Введем полученные данные в нашу таблицу (рис. 2.25).

	A	B	C	D	E
1					
2			Данные	Прогноз	
3	1	2009	110565		
4	2	2010	130848		
5	3	2011	256007		
6	4	2012	271631		
7	5	2013	322181		
8	6	2014	353879		
9	7	2015	385152		
10	8	2016	468235		
11	9	2017	517694		
12	10	2018	557672		
13	11	2019		=49487*A13+65210	
14	12	2020			

Рис. 2.25. Ввод уравнения

Зафиксируем введенное уравнение нажатием кнопки «Enter» и распространим на следующую ячейку. Таким образом, неизвестный параметр x – это номер временного ряда, для 2019 г. x=11, для 2020 г. x=12, для 2021 г. x=13 и т. д. Для прогноза на 2019 г. уравнение будет y=49487*11+65210. Построенный прогноз будет представлен на следующем графике (рис. 2.26):

		Данные	Прогноз
1	2009	110565	
2	2010	130848	
3	2011	256007	
4	2012	271631	
5	2013	322181	
6	2014	353879	
7	2015	385152	
8	2016	468235	
9	2017	517694	
10	2018	557672	560080
11	2019		609567
12	2020		659054

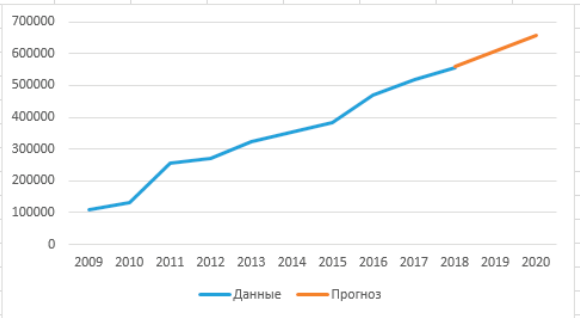


Рис. 2.26. Прогноз

Неизвестные параметры можно рассчитать и «вручную», используя инструмент «Регрессия» в пакете анализа. В появившемся окне в качестве входного интервала *Y* задаем известные значения временного ряда, а в качестве входного интервала X – номера временного ряда (рис. 2.27).

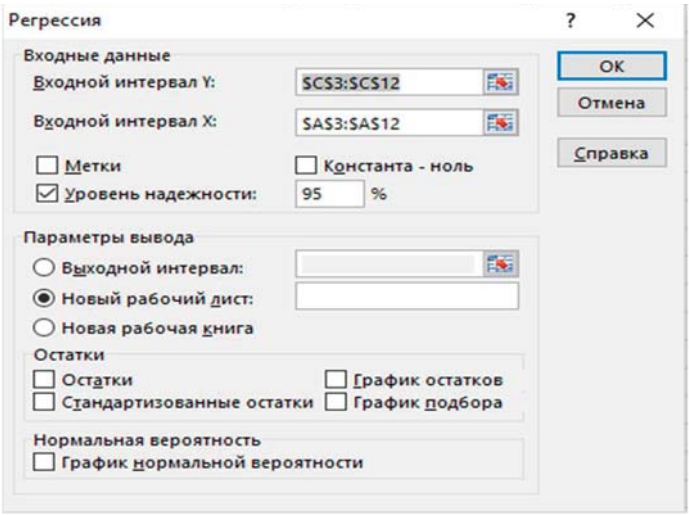


Рис. 2.27. Регрессия

После нажатия кнопки «OK» на отдельном листе появятся следующие результаты (рис. 2.28):

1	ВЫВОД ИТОГОВ								
2									
3	*ессионная статистика*								
4	Множеств	0,990344							
5	R-квадрат	0,980781							
6	Нормиров	0,978073							
7	Стандарти	22246,03							
8	Наблюдени	10							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*ачимость F*			
12	Регрессия	1	2,02E+11	2,02E+11	408,2479	3,76E-08			
13	Остаток	8	3,96E+09	4,95E+08					
14	Итого	9	2,06E+11						
15									
16	*Ко*	*ффициен*	*ртная о*	*татисти*	*²-Значени*	*ижние 95%*	*рхние 95%*	*жние 95,0*	*рхние 95,0%*
17	Y-пересе	65210,2	15196,93	4,291012	0,002648	30166,02	100254,4	30166,02	100254,4
18	Перемен	49486,58	2449,207	20,20515	3,76E-08	43838,7	55134,46	43838,7	55134,46
19									

Рис. 2.28. Регрессионный анализ

Обратите внимание, что рассчитанные значения коэффициентов аналогичны предыдущим результатам.

Технология аналитического выравнивания в LibreOffice Calc практически такая же, как и в MS Excel. На первом этапе осуществляется вставка графика, для чего выделяем набор данных и выбираем в главном меню пункт «Вставка» и элемент «Диаграмма». Выбираем тип диаграмм «Линии» и вид диаграммы «Только линии» и нажимаем кнопку «Готово». На получившейся диаграмме двойным щелчком мыши переходим в режим редактирования, выделяем ряд данных и нажимаем на нее правой кнопкой мыши. В появившемся контекстном меню выделяем пункт «Вставить линию тренда». В появившемся окне (рис. 2.29) переключаемся на закладку «Тип» и в следующем окне ставим галочки напротив пунктов «Показать уравнение» и «Показать коэффициент детерминации (R^2)».

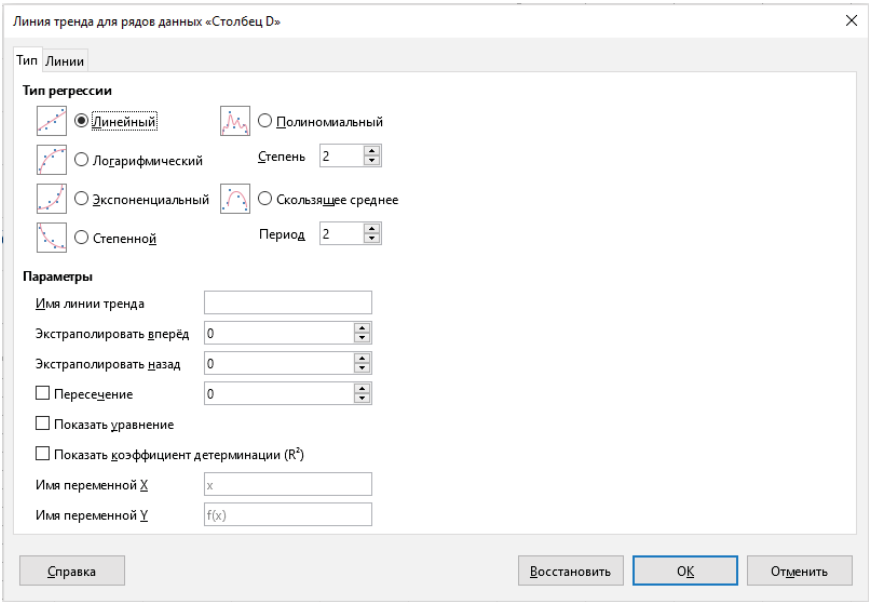


Рис. 2.29. Линия тренда для рядов данных

Как и при применении MS Excel, на диаграмме появятся линия тренда, уравнение и коэффициент детерминации (рис. 2.30).

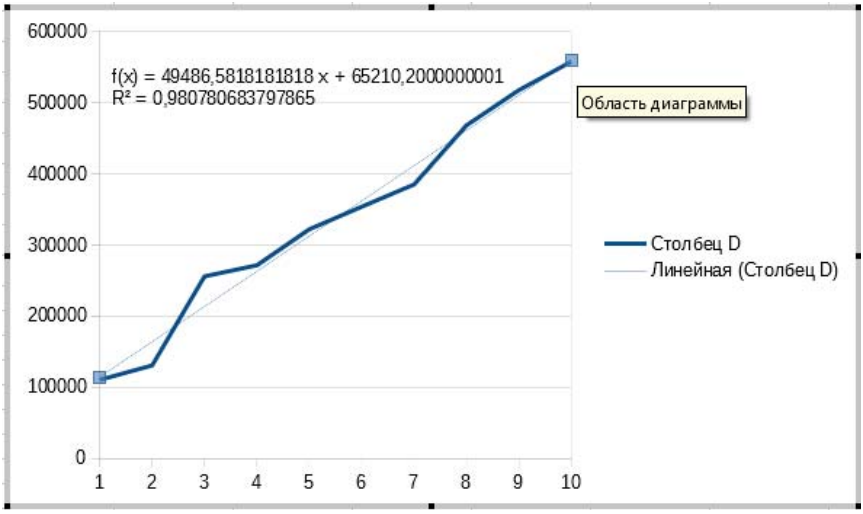


Рис. 2.30. Уравнение тренда

Полученные значения следующие: уравнение тренда y=49486,6*x+65210,2; R^2=0,98. Обратим внимание на то, что полученные результаты аналогичны тем, что получили при применении MS Excel.

2.3. Нелинейные модели и индексы сезонности

Наряду с линейным трендом можно построить и некоторые другие виды аппроксимирующих кривых[1], иногда их называют нелинейными[2]. Наиболее распространенными видами являются: экспоненциальная, логарифмическая, степенная и полиномиальная[3]. Например, при подборе линейного тренда для следующего графика (рис. 2.31) коэффициент аппроксимации R^2 составляет всего 0,59, что явно недостаточно для построения качественной модели.

[1] Иногда их еще называют сглаживающими кривыми.

[2] Здесь и далее нелинейные и криволинейные функции будут восприниматься как синонимы.

[3] Строго говоря, полином является степенной функцией, и степень полинома определяет количество пиков временного ряда.

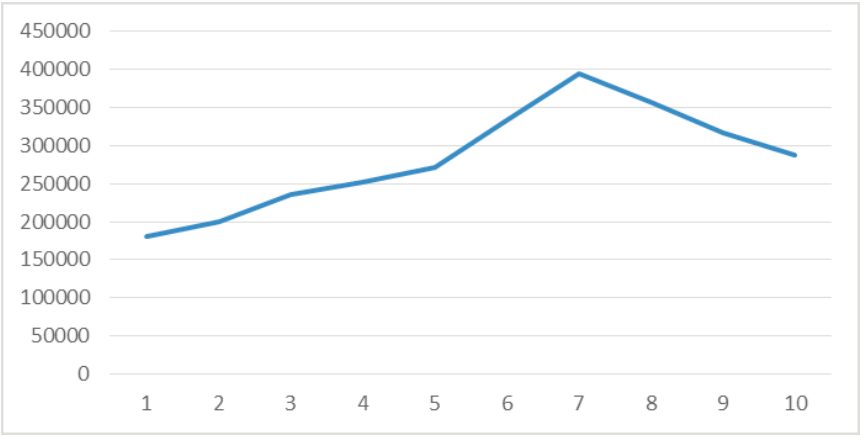


Рис. 2.31. График временного ряда

Кроме этого, при построении прогноза получается возрастающий тренд, хотя последние четыре отчетных периода наблюдалось снижение. В этой связи целесообразно попытаться подобрать иные формы зависимостей. В следующей таблице будут представлены различные виды уравнений и соответствующие им коэффициенты детерминации.

Таблица 2.3. Формы уравнений

Название	**Уравнение**	**R^2**
Линейная	$y = 17487x + 187210$	0,59
Экспоненциальная	$190738e^{0,0669x}$	0,65
Логарифмическая	$y = 78068\ln(x) + 165469$	0,69
Степенная	$y = 174546x^{0,3025}$	0,78
Полином 2-й степени	$y = -4213,4x^2 + 63833x + 94516$	0,81
Полином 3-й степени	$y = -1283x^3 + 16956x^2 - 33802x + 204597$	0,93

Очевидно, что для прогноза наилучшим образом подойдут полиномиальные формы уравнений. Однако высокие степени для практических расчетов применять сложно, поэтому мы ограничимся только полиномом 2-й степени. Кроме этого, используя полином, можно подобрать такую теоретическую функцию, которая будет иметь достаточно высокие значения коэффициента аппроксимации, однако интерпретировать полученные результаты будет зачастую невозможно.

Добавляем линию тренда, в параметрах выбираем полиномиальную и указываем степень 2, которая стоит по умолчанию (рис. 2.32).

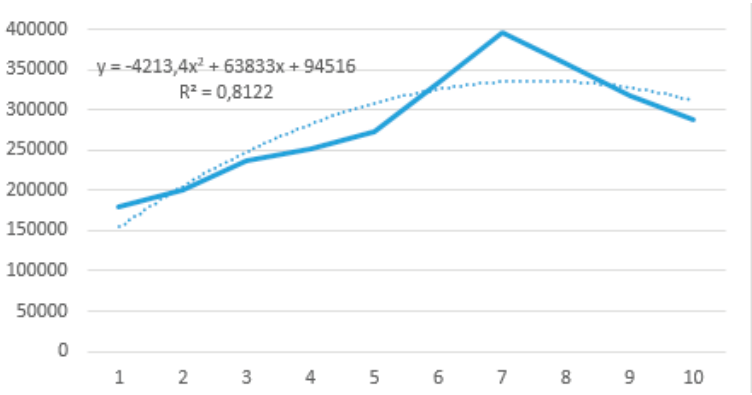


Рис. 2.32. График полинома 2-й степени

Вводим полученное уравнение в ячейку и фиксируем изменения нажатием клавиши «Enter»[1] (рис. 2.33).

СУММ | =-4213,4*B12^2+63833*B12+94516

	A	B	C	D	E	F	G	H
1								
2				Данные	Прогноз			
3		1	2009	180565				
4		2	2010	200848				
5		3	2011	236007				
6		4	2012	251631				
7		5	2013	272181				
8		6	2014	333879				
9		7	2015	395152				
10		8	2016	358235				
11		9	2017	317694				
12		10	2018	287672	=-4213,4*B12^2+63833*B12+94516			
13		11	2019					
14		12	2020					

Рис. 2.33. Ввод уравнения в таблицу

Распространим введенную формулу на следующие отчетные периоды и построим график (рис. 2.34). Аналогичным образом рассчитываются и остальные формы уравнений.

[1] Для возведения в квадрат можно просто умножить x на x либо использовать специальный знак «^». Таким образом, возведение в степень числа x будет обозначаться «x^2», где 2 – это показатель степени.

Рис. 2.34. График функции

Рассмотрим теперь методику построения модели при помощи Libre Office Calc. В целом алгоритм построения схож с MS Excel, но технология имеет некоторые отличительные особенности. Построим не полиномиальное уравнение, а логарифмическое. На первом этапе также строится диаграмма, для чего выделяется набор данных и активируется меню «Ставка» – «Диаграмма». В появившемся окне выбираем пункт «Линии» и 3-й значок «Только линии», и после нажатия кнопки «Готово» появится диаграмма. Переходим в режим редактирования двойным щелчком мыши, после нажатия правой кнопкой мыши на диаграмме активируется меню, в котором выбираем пункт «Вставить линию тренда». Далее переключаемся на закладку «Тип» и выбираем логарифмический тип регрессии (рис. 2.35).

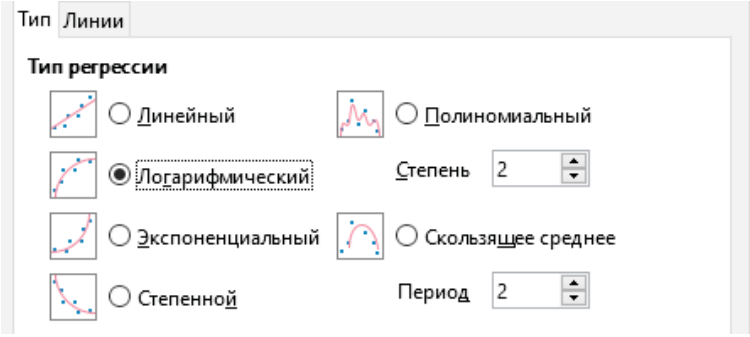


Рис. 2.35. Тип регрессии

Ставим галочки напротив пунктов «Показать уравнение» и «Показывать коэффициент детерминации (R^2)», и после нажатия кнопки «ОК» появится диаграмма следующего вида (рис. 2.36):

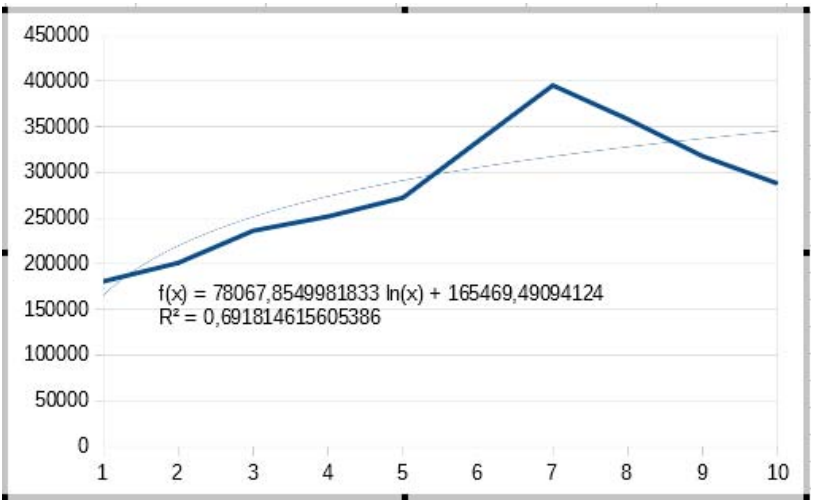
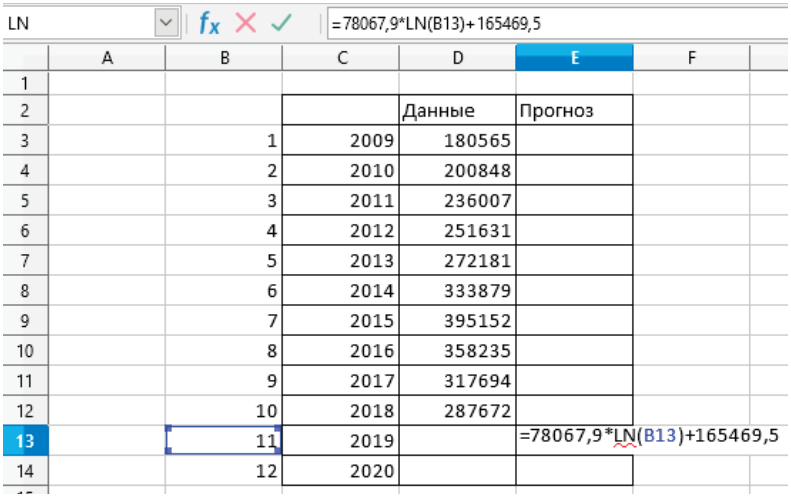


Рис. 2.36. Диаграмма с логарифмами

Введем полученное уравнение в ячейку и зафиксируем (рис. 2.37).



	A	B	C	D	E	F
1						
2				Данные	Прогноз	
3		1	2009	180565		
4		2	2010	200848		
5		3	2011	236007		
6		4	2012	251631		
7		5	2013	272181		
8		6	2014	333879		
9		7	2015	395152		
10		8	2016	358235		
11		9	2017	317694		
12		10	2018	287672		
13		11	2019		=78067,9*LN(B13)+165469,5	
14		12	2020			

Рис. 2.37. Ввод уравнения в ячейку

Распространим введенную формулу на все ячейки и построим график фактических и прогнозных значений (рис. 2.38).

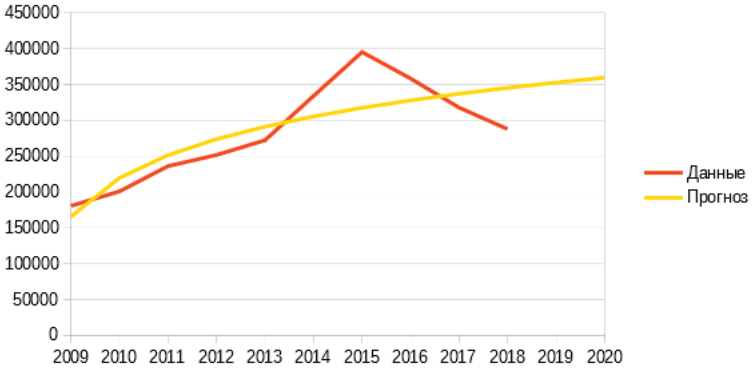


Рис. 2.38. График фактических и прогнозных значений

Для расчета неизвестных параметров уравнения также можно воспользоваться функцией «ПРЕДСКАЗ.ЛИНЕЙН», которая есть как в MS Excel, так и в LibreOffice Calc. Активируем нужную ячейку, куда нужно поместить прогноз, через «Мастер функций» находим функцию «ПРЕДСКАЗ.ЛИНЕЙН», выберем ее. В появившемся окне введем следующие параметры: «Значение» – это показатель x, то есть ячейка B13; «Данные Y» – это имеющиеся показатели y, то есть диапазон данных D3:D12; «Данные X» – это имеющиеся номера временного ряда, то есть диапазон номеров ряда B3:D12. Все должно получиться так, как указано на следующем рисунке (рис. 2.39).

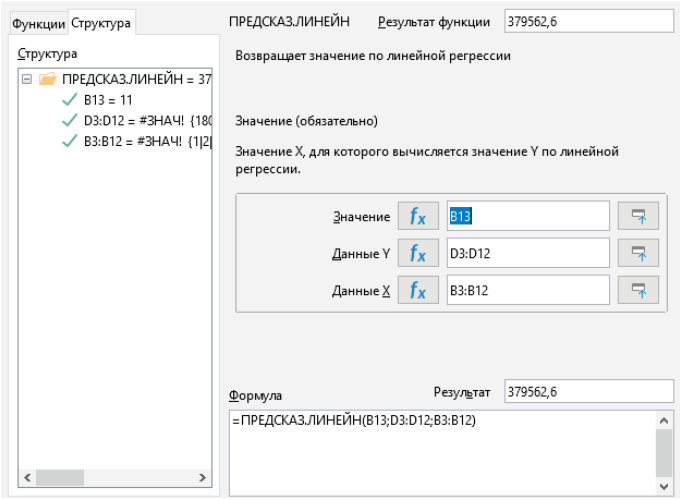


Рис. 2.39. Функция «ПРЕДСКАЗ.ЛИНЕЙН»

В результате в целевой ячейке мы получим прогнозное значение. Данный подход позволяет рассчитать только линейную форму уравнения, что не всегда целесообразно.

Индексы сезонности. Рассмотренный метод подбора аналитической функции предназначен для анализа годовых данных и не учитывает фактор сезонности. В случае если мы располагаем еженедельным или ежемесячным набором данных, то для прогнозирования необходимо учитывать фактор сезонности. Например, имеются ежемесячные данные о зарегистрированных преступлениях с января 2017 по декабрь 2019 г. (всего 36 точек). Необходимо построить прогноз на 2020 г. На первом этапе вставляем график и добавляем линию тренда (рис. 2.40). Нас интересует полученное уравнение $y = 11{,}947x + 14946$.

Рис. 2.40. Прогноз

На следующем этапе включим сезонные изменения в модель, для чего рассчитаем сезонные индексы по формуле:

$$\hat{y}_t = \frac{\sum y_t}{\sum y} \cdot 12,$$

где $\hat{y}_t$ – сезонный индекс t периода, y – набор всех данных, t – номер периода. Альтернативная формула для расчета индексов сезонности может быть представлена как отношение среднего арифметического за определенный период (месяц, квартал) к среднему арифметическому по всей совокупности данных:

$$\hat{y}_t = \frac{\overline{y_t}}{\bar{y}},$$

где $\overline{y_t}$ – среднее значение t периода, $\bar{y}$ – среднее значение всех данных. Рассчитаем сезонные индексы для каждого месяца (рис. 2.41).

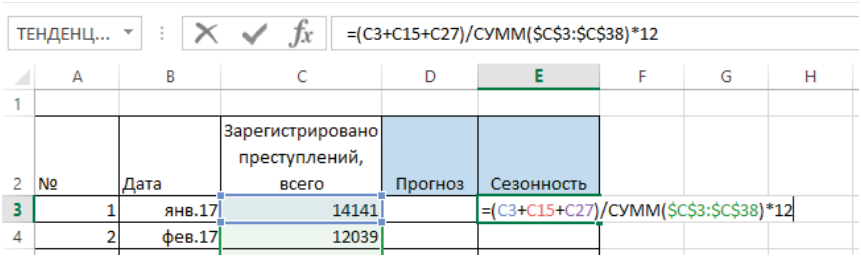


Рис. 2.41. Формула индексов сезонности

На следующем рисунке представлены рассчитанные индексы (рис. 2.42).

Сезонность
0,91
0,84
1,00
0,98
1,03
1,02
1,05
1,07
1,00
1,19
0,99
0,93

Рис. 2.42. Таблица рассчитанных индексов сезонности

На следующем этапе умножаем прогноз на соответствующий индекс (рис. 2.43).

38	36	дек.19	14379	
39	37	янв.20	15388	=C39*E3
40	38	фев.20	15400	

Рис. 2.43. Прогноз

И распространяем введенную формулу на все месяцы. Получившиеся значения представлены ниже (рис. 2.44).

39	37	янв.20	15388	14027
40	38	фев.20	15400	12922
41	39	мар.20	15412	15411
42	40	апр.20	15424	15059
43	41	май.20	15436	15887
44	42	июн.20	15448	15728
45	43	июл.20	15460	16254
46	44	авг.20	15472	16524
47	45	сен.20	15484	15469
48	46	окт.20	15496	18492
49	47	ноя.20	15508	15303
50	48	дек.20	15519	14389

Рис. 2.44. Прогнозная таблица

Приведем построенный график (рис. 2.45).

Рис. 2.45. Прогноз

Рассмотрим технологию прогнозирования с использованием LibreOffice Calc, но в отличие от предшествующего метода устраним влияние тренда. На первом этапе рассчитаем тренд при помощи функции «ПРЕДСКАЗ», как указано на следующем рисунке (рис. 2.46), и распространим формулу ниже.

	A	B	C	D	E	F
1		Месяц	Данные	Тренд		
2	1	Январь	14141	=ПРЕДСКАЗ(A2;$C$2:$C$37;$A$2:$A$37)		
3	2	Февраль	12039			
4	3	Март	16263			
5	4	Апрель	15125			
6	5	Май	15678			
7	6	Июнь	15693			

Рис. 2.46. Функция «Предсказ»

На следующем этапе рассчитаем индексы сезонности и также распространим введенную формулу на все ячейки (рис. 2.47).

	A	B	C	D	E	F	G	H
1		Месяц	Данные	Тренд	Индекс сезонности			
2	1	Январь	14141	14958	=СРЗНАЧ(C2;C14;C26)/СРЗНАЧ($C$2:$C$37)			
3	2	Февраль	12039	14970				

Рис. 2.47. Индекс сезонности

На следующем этапе умножаем полученные значения тренда на индекс сезонности и также распространяем на требуемый диапазон[1] (рис. 2.48).

38	37	Январь		15388	=D38*E2
39	38	Февраль		15400	12922
40	39	Март		15412	15411
41	40	Апрель		15424	15059

Рис. 2.48. Прогноз с сезонностью

Полученные прогнозные значения лучше всего отобразить на графике (рис. 2.49).

[1] В нашем примере на год.

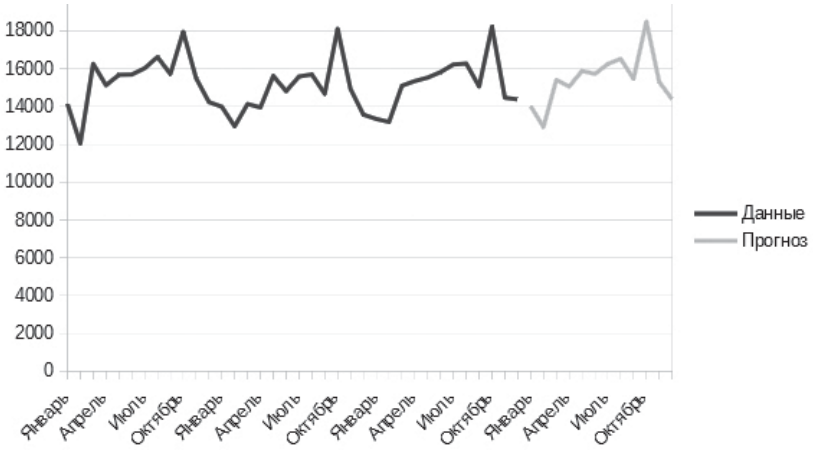


Рис. 2.49. График прогноза с сезонностью

Вторая группа методов анализа временных рядов основана на предположении, что каждый последующий показатель зависит от предыдущего. Эти методы основаны на сглаживании временного ряда, а также на экстраполяции.

Для расчета прогноза методом **скользящего среднего** необходимо определиться с периодом прогнозирования[1]. В нашем примере период прогнозирования представлен кварталом, то есть n=3.

Скользящее среднее рассчитывают по формуле:

$$S_t = \frac{y_1 + y_2 + \cdots + y_n}{n},$$

где S – прогнозные значения, n – период.

Например, для n=3

$$S_{\text{фев}} = \frac{14141 + 12039 + 16263}{3} \approx 14148,$$

$$S_{\text{мар}} = \frac{12039 + 16263 + 15125}{3} = 14476,$$

$$S_{\text{апр}} = \frac{16263 + 15125 + 15678}{3} = 15689$$

и т. д.

[1] Строгих рекомендаций по выбору периода расчета скользящего среднего не существует.

Введем формулу во вторую ячейку листа MS Excel следующим образом и распространим введенную формулу до предпоследней ячейки (рис. 2.50).

	A	B	C	D
1		Месяц	Данные	Скользящее среднее, n=3
2	1	Январь	14141	
3	2	Февраль	12039	=СРЗНАЧ(C2:C4)
4	3	Март	16263	14476
5	4	Апрель	15125	15689
6	5	Май	15678	15499

Рис. 2.50. Скользящее среднее

После того, как мы рассчитали скользящее среднее для всех периодов, нужно построить прогноз на январь 2020 г. по следующей формуле:

$$y_{t+1} = S_{t-1} + {}^{1}/_{k}\,(y_t - y_{t-1}),$$

где t – текущий отчетный период, y_{t+1} – прогнозируемый показатель, S_{t-1} – скользящее среднее за предыдущий период, t+1 – прогнозный период, k – интервал сглаживания, y_t – текущее значение показателя, y_{t-1} – предыдущее значение показателя. Подставляя значения в формулу, получим:

$$y_{\text{янв}} = 15689 + {}^{1}/_{3}\,(14379 - 14464) = 15661,$$

затем определим скользящую среднюю за декабрь:

$$S_{\text{дек}} = \frac{14141 + 12039 + 16263}{3} \approx 14148.$$

Повторяем процедуру. Строим прогноз на следующий месяц. Затем рассчитываем скользящее среднее и т. д.

MS Excel позволяет рассчитать скользящее среднее при помощи пакета анализа (рис. 2.51).

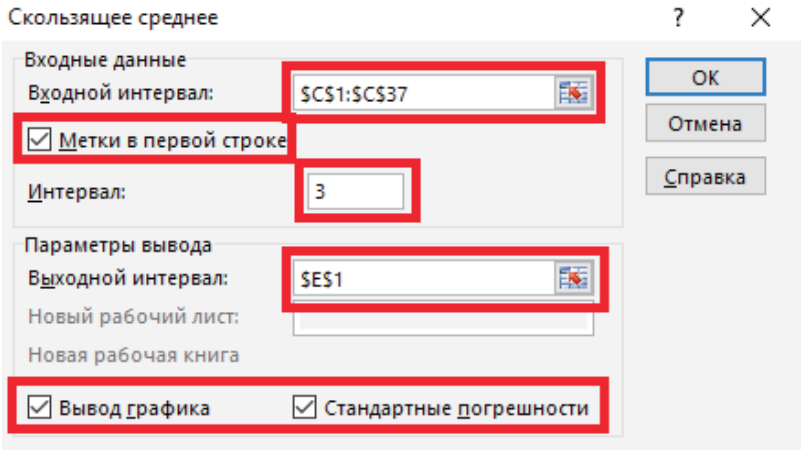


Рис. 2.51. Параметры скользящего среднего

В появившемся окне задаем параметры: «Входной интервал» – в нашем случае это имеющиеся данные в диапазоне C1:C37; «Метки в первой строке» – если выделяем заголовок данных; «Интервал» – в нашем примере это 3. Можно также отметить пункт «Вывод графика». Пункт «Стандартные погрешности» отмечать нет необходимости. После того, как задаем диапазон «Выходного интервала» (достаточно отметить одну ячейку, система автоматически расширит колонку в указанной ячейке до нужного размера) и нажимаем кнопку «ОК», мы получим следующие результаты[1] (рис. 2.52):

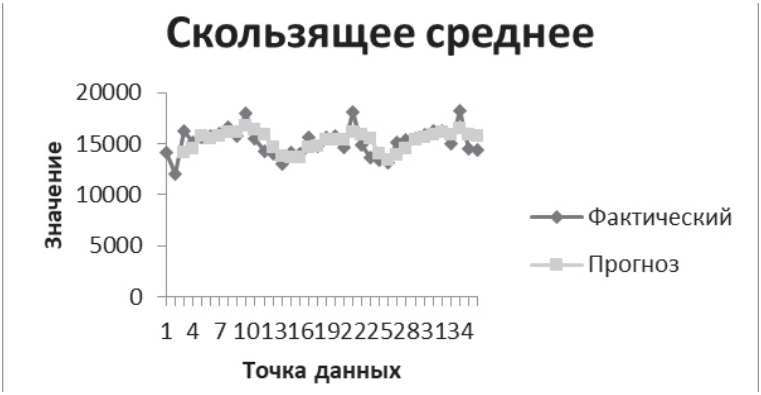


Рис. 2.52. Скользящее среднее

[1] Для удобства отображения данных нужно удалить значения «#Н/Д» в ячейках.

Второй метод основан на экспоненциальном сглаживании, отличительной особенностью которого является то, что прогнозировать с его помощью можно только на один период вперед. Кроме этого, простое экспоненциальное сглаживание не учитывает сезонность. Расчет осуществляется по формуле:

$$S_{t+1} = \alpha \cdot y_t + (1 - \alpha) \cdot S_t,$$

где t – текущий отчетный период, S_{t+1} – прогнозируемый показатель, S_t – сглаженный показатель за текущий период, y_t – текущее значение показателя, α – параметр сглаживания (фактор затухания) принимает значения от 0,1 до 0,9. Рассмотрим пример, когда параметр сглаживания (фактор затухания) α=0,5. Подставляя значения в формулу, получим:

$$S_{t+1} = 0{,}5 \cdot 14379 + 0{,}5 \cdot 15679{,}2 \approx 15029.$$

На следующем рисунке представлены параметры соответствующего инструмента в MS Excel (рис. 2.53).

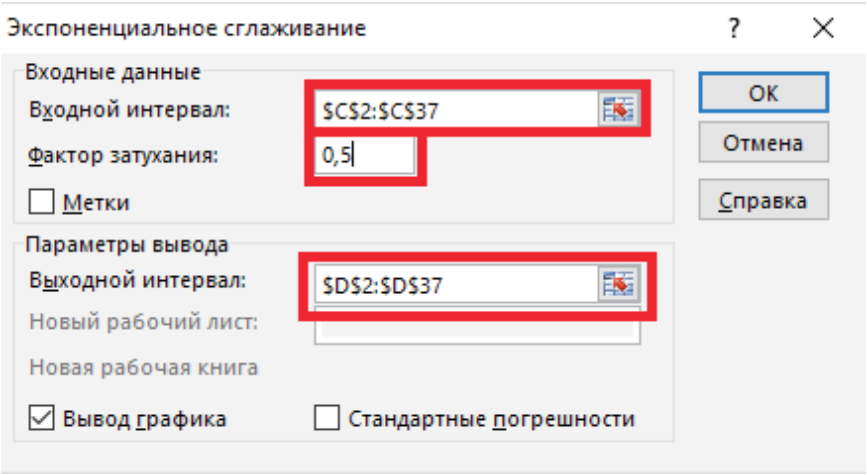


Рис. 2.53. Экспоненциальное сглаживание

На следующих двух рисунках (рис. 2.54 и рис. 2.55) представлены две диаграммы, в которых указан разный интервал сглаживания (0,5 и 0,05).

Рис. 2.54. Экспоненциальное сглаживание (α=0,5)

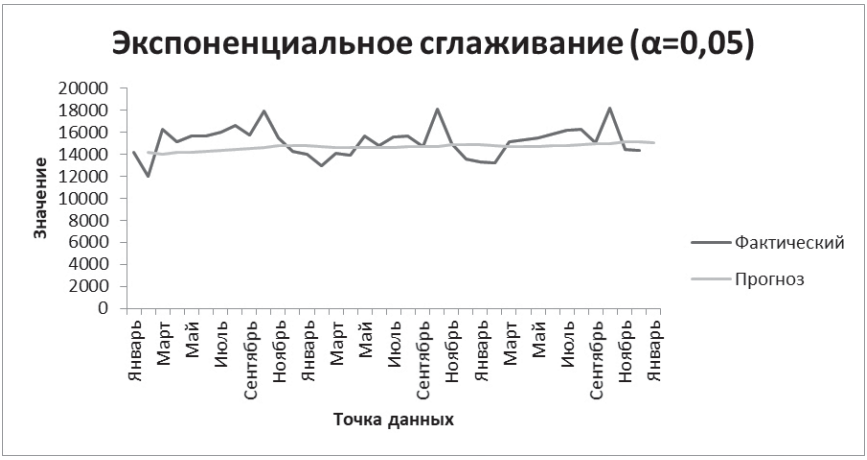


Рис. 2.54. Экспоненциальное сглаживание (α=0,05)

Отчетливо видна разница между двумя моделями с разными интервалами затухания.

На практике основной проблемой является выбор значения фактора затухания α. Один из подходов заключается в выборе параметра сглаживания по формуле: $\alpha = 2/(n+1)$, то есть чем длиннее временной ряд, тем ниже фактор затухания α.

Данные методы в качестве фактора, оказывающего влияние на исследуемый временной ряд, оперируют только одной переменной, это временная переменная *t*. В следующей главе рассмотрим

ситуацию, когда на исследуемый процесс (явление) оказывает влияние иной фактор (факторы).

С версии MS Excel 2016 появилось несколько новых функций, позволяющих осуществлять прогнозирование с использованием аддитивного (ПРЕДСКАЗ.ETS.ADD) или мультипликативного (ПРЕДСКАЗ.ETS.MULT) алгоритмов экспоненциального сглаживания (ETS). Функции «ПРЕДСКАЗ.ETS» позволяют рассчитать будущее значение на основе ретроспективных данных. Функции «ПРЕДСКАЗ.ETS.PI.ADD» и «ПРЕДСКАЗ.ETS.PI.MULT» позволяют рассчитать прогнозный интервал, а «ПРЕДСКАЗ.ETS.STAT.ADD» и «ПРЕДСКАЗ.ETS.STAT.MULT» позволяют рассчитать статистку, соответственно, для аддитивного и мультипликативного алгоритмов экспоненциального сглаживания. Кроме этого, функция «ПРЕДСКАЗ.ETS.СЕЗОННОСТЬ» позволяет определить количество элементов в периоде. Заметим, что данные функции поддерживаются Libre Offce Calc без каких-либо ограничений. Применение данных функций при решении задач анализа и прогнозирования существенно упрощает объемы «ручной» работы.

Следует иметь в виду, что набор функций чрезвычайно широк, и решить те или иные задачи возможно за счет различных способов. Например, для ручного расчета уравнения тренда можно использовать две функции «ОТРЕЗОК» и «НАКЛОН», которые возвращают значения коэффициентов *a* (отрезок, отсекаемый на оси линией регрессии) и *b* (наклон линии регрессии) линейной модели. Также наряду с функцией «ПРЕДСКАЗ» возможно использование функций «ТЕНДЕНЦИЯ» и «РОСТ».

Рассмотренные здесь методы анализа временных рядов не претендуют на полное их описание. Существует достаточно большой класс авторегрессионных моделей (*англ. AutoRegressive*, AR-модели), а также их различные комбинации. Например, сочетание скользящего среднего (*англ. moving average,* MA-модели) и авторегрессионной AR-модели называется ARMA (*англ. AutoRegressive Moving Average*). Расширение ARMA для нестационарных временных рядов называется ARIMA (*англ. AutoRegressive Integrated Moving Average*), эта же модель с учетом сезонности – SARIMA (*англ. Season AutoRegressive Integrated Moving Average*).

Глава 3. Регрессионный анализ

В данной главе рассматриваются линейные и нелинейные модели анализа данных, модели с конкретными переменными, методы оценки коэффициентов модели и компьютерные технологии для построения парных и множественных регрессионных моделей.

3.1. Модели линейной регрессии

Линейная регрессия – это статистически используемая регрессионная модель зависимости переменной (объявленной, зависимой) *y* от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) *x* с функцией линейной зависимости.

Основным методом оценки неизвестных параметров регрессионной модели является метод наименьших квадратов (*англ. ordinary least squares*, OLS). Модель линейной регрессии является наиболее широко используемой и изучаемой в эконометрике.

Основная цель регрессионного анализа – оценить функциональную связь между независимой переменной *X* и условным ожиданием зависимой переменной *Y*. Парная регрессия – это модель, в которой (среднее) теоретическое значение зависимой переменной *Y* принимается во внимание как функция независимой переменной *X*. Множественная регрессия – это модель, в которой рассматривается теоретическое (среднее) значение зависимой переменной *Y* как функции нескольких независимых переменных *X*1, *X*2, ... *X*m.

Спецификация модели – формулировка типа модели на основе соответствующей теории взаимосвязи между переменными. Определяется состав переменных и математическая функция, отражающая взаимосвязь между ними.

Мультиколлинеарность – это линейная связь между двумя или более независимыми переменными[1].

Криминологические модели являются характерными представителями факторных социальных моделей, с помощью которых они пытаются формализовать сложные социальные процессы, связанные с противоправным поведением, возникновением преступности и формированием механизмов правового регулирования.

Построение модели – это итеративный процесс поиска эффективных независимых переменных. Основная цель – попытаться объ-

[1] *Торопов Б. А., Гонов Ш. Х.* Статистические методы принятия управленческих решений: сборник задач (задачник). Москва: Академия управления МВД России, 2019. 76 с.

яснить зависимые переменные, которые необходимо моделировать и пересмотреть с помощью инструмента регрессии, и определить, какие величины являются эффективными предикторами. Затем удаляем и (или) добавляем независимые переменные, пока не найдется наиболее подходящая модель регрессии.

Наиболее распространенными моделями регрессии являются линейные модели, однако функциональная форма зависимости может быть различной. На следующем рисунке представлены наиболее распространенные функциональные формы (рис. 3.1).

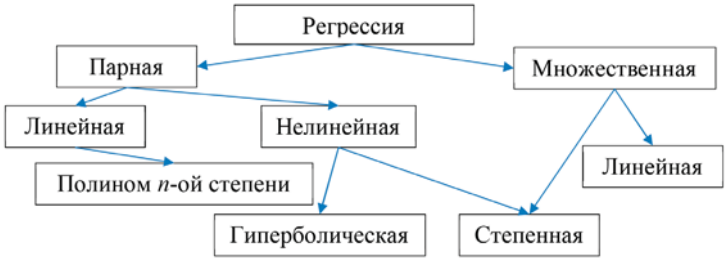


Рис. 3.1. Функциональные формы моделей

В графическом виде некоторые функции представлены ниже (рис. 3.2).

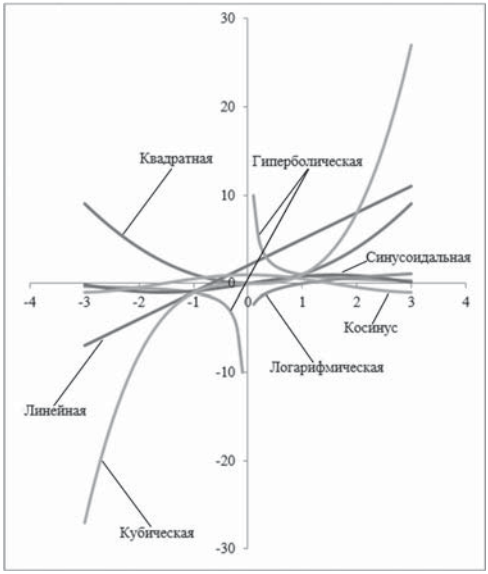


Рис. 3.2. Функции

Для оценки неизвестных параметров уравнения могут применяться различные методы, наиболее распространенным является МНК. В таблице представлены некоторые виды моделей и соответствующие им математические выражения (таблица 3.1).

Таблица 3.1. Некоторые виды моделей

Наименование	**Уравнение**
Линейная	$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon$
Полином k-ой степени	$y = \alpha + \beta_{11} x_1 + \beta_{12} x_1^2 + \cdots + \beta_{1k} x_1^k + \cdots$ $\ldots + \beta_{p1} x_p + \beta_{p2} x_{p2}^2 + \cdots + \beta_{pk} x_p^k + \varepsilon$
Обратная	$y = \dfrac{1}{\alpha + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon}$
Степенная	$y = \alpha \cdot x_1^{\beta_1} \cdot \ldots \cdot x_p^{\beta_p} \cdot \varepsilon$
Показательная	$y = \alpha \cdot \beta_1^{x_1} \cdot \ldots \cdot \beta_p^{x_p} \cdot \varepsilon$
Полулогарифмическая	$y = \alpha + \beta_1 \ln x_1 + \cdots + \beta_k \ln x_p + \varepsilon$

В общем виде модель парной регрессии может быть представлена как $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$, $i = 1, \ldots n$, где α и β – неизвестные параметры модели, y_i – зависимая переменная, x_i – зависимая переменная (регрессор) и ε_i – случайная компонента (случайная ошибка) модели, n – количество наблюдений, i – номер наблюдения.

В «Анализе данных» выбираем пункт «Регрессия» и в появившемся окне (рис. 3.3) в качестве входного интервала *Y* задаем значения зависимой переменной, а в качестве входного интервала *X* – значения независимой переменной (независимых переменных).

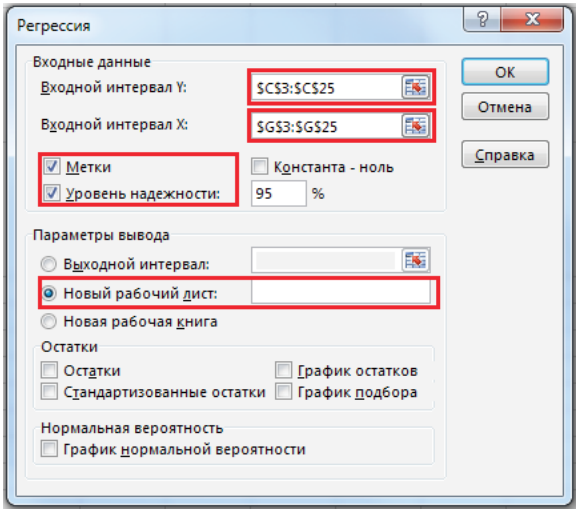


Рис. 3.3. Параметры расчета регрессии

Например, в качестве параметра *Y* будут выступать преступления, а в качестве *X* – численность сотрудников патрульно-постовой службы полиции (далее – ППСП). После нажатия кнопки «ОК» на новом листе появятся следующие результаты:

Регрессионная статистика	
Множественный R	0,85
R-квадрат	0,73
Нормированный R-квадрат	0,71
Стандартная ошибка	3064,8
Наблюдения	22

Дисперсионный анализ

	df	**SS**	**MS**	**F**	**Значимость F**
Регрессия	1	504026628	504026628	53,6612454	4,42E-07
Остаток	20	187854987	9392749,346		
Итого	21	691881615			

	Коэффициенты	**Стандартная ошибка**	**t-статистика**	**P-Значение**
Y-пересечение	10159,20099	4151,03493	2,44738991	0,02374833
ППСП	15,81146374	2,15844856	7,325383638	4,4218E-07

Проинтерпретируем некоторые полученные результаты.

R-квадрат. Полученная модель описывает 73 % вариации, оставшиеся 27 % – это другие неучтенные в модели факторы[1].

P-Значение. На 5 % уровне вероятности будем принимать данное значение не больше 0,05.

F-статистика *Фишера*. Должна быть больше, *чем значимость F.*

Полученная модель представлена в виде выражения, на основе которой рассчитаем модельные значения для следующей таблицы:

№	Преступления	ППСП	Модель	№	Преступления	ППСП	Модель
1	39662	1972	41317	12.	41735	2080	43023
2	44492	2013	41965	13.	41809	2252	45741
3	48214	1856	39484	14.	36126	1731	37509
4	48135	2199	44903	15.	34706	1483	33591
5	48023	2251	45725	16.	35584	1562	34839
6	38913	2062	42739	17.	33839	1638	36040
7	40991	2032	42265	18.	33476	1536	34428
8	44383	2194	44824	19.	36259	1409	32421
9	48088	2211	45093	20.	34336	1575	35044
10	45369	2327	46926	21.	31726	1569	34949
11	45968	2264	45930	22.	32303	1566	34902

Построим линейный график с фактическими и модельными данными (рис. 3.4).

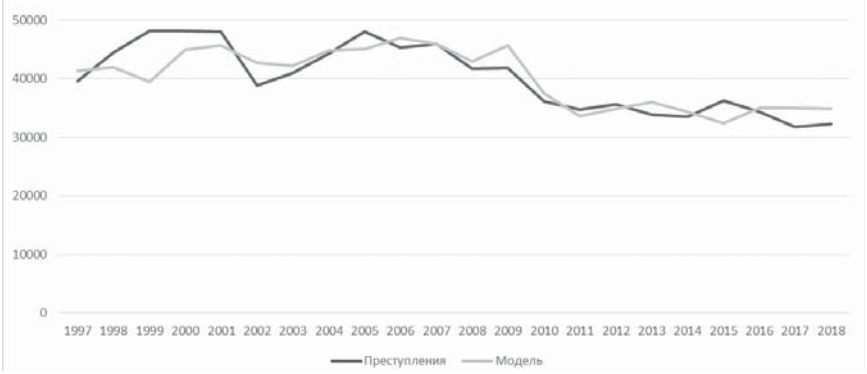


Рис. 3.4. График фактических и модельных значений

[1] Для практических целей моделирования значение R-квадрат принимается не менее 0,7.

На следующем этапе рассчитаем прогнозные значения уровня преступности на 2019 и 2020 г. при условии увеличения сотрудников ППСП ежегодно на 100 чел. В этом случае прогноз составит:

	ППСП	**Преступлений**
2019 г.	1666	36482
2020 г.	1766	38062

При исследовании организационных (социальных и экономических) систем парная регрессия не в полной мере может отразить степень влияния различных факторов, поэтому в анализе данных большее распространение получили модели множественной регрессии. В общем виде модель множественной регрессии может быть представлена как:

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \beta_k \cdot x_k + \varepsilon,$$

где a и β_k – неизвестные параметры модели, y – зависимая переменная, x_k – зависимая переменная (регрессор), ε – случайная компонента (случайная ошибка) модели и k – количество зависимых переменных (регрессоров).

В компактном виде данная модель записывается как:

$$y = a + \sum_{i=1}^{k} \beta_k x_k + \varepsilon.$$

На предыдущем этапе мы рассмотрели ситуацию, когда в качестве независимой (влияющей) переменной выступал только один фактор (численность сотрудников ППСП). Теперь рассмотрим ситуацию, когда в качестве независимых переменных выступают несколько факторов (множественная регрессия).

В инструменте «Анализ данных» выбираем пункт «Регрессия» и в появившемся окне задаем входной параметр *Y* и входной параметр *X*. В качестве *Y* будут выступать зарегистрированные преступления, а в качестве *X* факторы – безработица, численность сотрудников ППСП и численность сотрудников ДПС. После нажатия кнопки «ОК» на новом листе появятся следующие результаты:

Регрессионная статистика	
Множественный R	0,92
R-квадрат	0,84
Нормированный R-квадрат	0,81

Стандартная ошибка	2475,8
Наблюдения	22

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	3	5,82E+08	1,94E+08	31,62653	2,17E-07
Остаток	18	1,1E+08	6129378		
Итого	21	6,92E+08			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересечение	-3795,71	6326,487	-0,59997	0,556004
ППСП	10,77304	2,248609	4,790978	0,000146
ДПС	8,226943	3,725148	2,208488	0,040419
Численность безработных	299,2542	95,89231	3,120732	0,005905

Проинтерпретируем полученные результаты.

1. *R*-квадрат. Полученная модель описывает 84 % вариации, оставшиеся 16 % – это другие неучтенные в модели факторы.

2. *P*-Значение. На 5 % уровне вероятности будем принимать данное значение не больше 0,05.

3. *F*-статистика. Должна быть больше, чем значимость F.

Полученная модель выглядит как $y = -3795{,}7 + 10{,}8 \cdot x_1 + 8{,}2 \cdot x_2 + 299{,}2 \cdot x_3$, где x_1 – численность сотрудников ППСП, x_2 – численность сотрудников ДПС, x_3 – численность безработных. На основе полученной модели рассчитаем модельные значения для следующей таблицы:

Зарегистрировано преступлений	Численность сотрудников ППСП	Численность сотрудников ДПС	Численность безработных, тыс. чел.	Модель
39662	1972	1666	40,1	
44492	2013	1781	42,2	
48214	1856	1742	40,2	
48135	2199	1638	40,6	
48023	2251	1867	35,7	
38913	2062	1602	23,7	

40991	2032	1933	34,7	
44383	2194	1943	32,9	
48088	2211	1995	29,3	
45369	2327	1607	40,8	
45968	2264	1577	41,4	
41735	2080	1649	34,4	
41809	2252	1653	28,5	
36126	1731	1491	23,9	
34706	1483	1658	22,9	
35584	1562	1824	21,1	
33839	1638	1821	22,2	
33476	1536	1765	29,6	
36259	1409	1558	32,8	
34336	1575	1572	30,9	
31726	1569	1561	28,0	
32303	1566	1395	28,1	

В соответствующую ячейку введем следующее выражение (рис. 3.5):

Q	R	S	T	U	V
Преступления	ППС	Модель (I)	Модель (II)		
39662	1972	41316,8	=-3795,7+10,8*G4+8,2*H4+299,2*I4		
44492	2013	41964,6	45175,14		
48214	1856	39484	42561,34		
48135	2199	44903,4	45532,62		
48023	2251	45725	46505,94		

Рис. 3.5. Уравнение в ячейке по модели

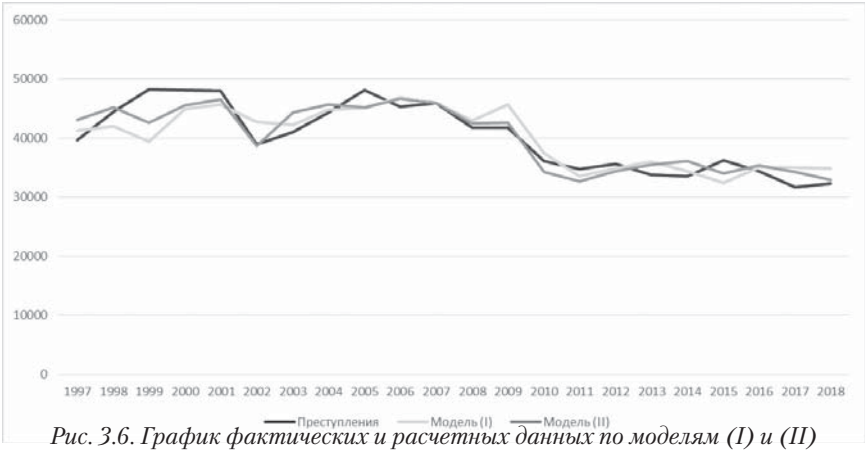


Рис. 3.6. График фактических и расчетных данных по моделям (I) и (II)

На следующем этапе рассчитаем прогнозные значения уровня преступности на 2019 и 2020 г. при условии увеличения сотрудников ППСП и ДПС ежегодно на 50 чел. и сокращения безработных на 1 тыс. чел. по сравнению с 2018 г. В этом случае прогноз составит:

	ППСП	**ДПС**	**Безработных**	**Преступлений**
2019 г.	1616	1445	27,1	33614,42
2020 г.	1666	1495	26,1	34265,22

Ранее мы рассмотрели процедуру построения линейной модели парной регрессии. Теперь рассмотрим процедуру построения регрессионной модели в виде полинома n-ой степени:

$$y = a + b_1 x_1 + b_2 {x_2}^2 + \cdots + b_n {x_n}^n.$$

На практике достаточным считается показатель 2-й и 3-й степени, применение более старших степеней связано с вычислительными сложностями, и рассматривать мы их не будем. В общем виде модель парной регрессии в виде полинома 2-й степени представлена ниже:

$$y = a + b_1 x_1 + b_2 {x_2}^2.$$

Вернемся к задаче, сформулированной ранее. На первом этапе произведем замену переменных. Так, X_1 – это численность ППСП, X_2 – это показатель X_1, возведенный в квадрат, Y – количество зарегистрированных преступлений. В ячейку D4 введем следующее значение (рис. 3.7):

	A	B	C	D
		Преступления	ППС	ППС^2
		Y	X_1	X_2
1997		39662	1972	=C4^2
1998		44492	2013	
1999		48214	1856	

Рис. 3.7. Возведение в квадрат показателя X_1

Таким образом, в ячейке появился показатель X, возведенный в квадрат. Распространим введенную формулу на все ячейки, для этого находим в правом нижнем углу ячейки значок + и, удерживая, протягиваем до конца. Формула автоматически распространится на все ячейки[1].

Далее рассчитаем неизвестные коэффициенты (a, b_1, b_2), для чего воспользуемся встроенной в MS Excel функцией «ЛИНЕЙН». Активируем свободную ячейку, в нашем примере это J4, и вставляем функцию «ЛИНЕЙН». В появившемся окне заполняем указанные поля (рис. 3.8).

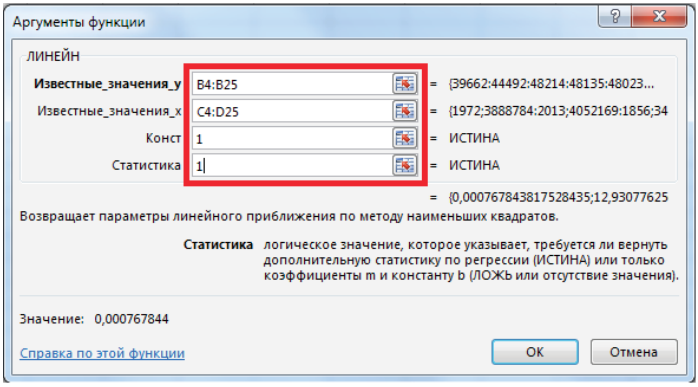


Рис. 3.8. Аргументы функции «ЛИНЕЙН»

В целевой ячейке появится некоторое значение. Проблема в том, что данная функция возвращает только одно значение, и это значение последнего коэффициента. Для того чтобы получить значения остальных коэффициентов, необходимо выполнить следующие действия: выделяем ячейки по следующей схеме: пять строк, начиная от уже рассчитанного коэффициента, и три столбца[2] (рис. 3.9).

[1] Можно использовать функции «Копировать» и «Вставить», результат будет тот же.
[2] Три (3) – это количество неизвестных коэффициентов.

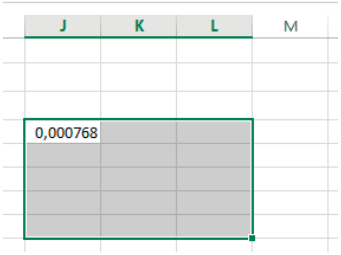


Рис. 3.9. Выделение целевого массива

Далее нажимаем клавишу «F2» и сочетание клавиш «Ctrl + Shift + Enter». Система заполнит массив данными следующего вида (рис. 3.10):

0,000767844	12,93077626	12790,26
0,011184943	42,02046109	38561,6
0,728554136	3143,986763	#Н/Д
25,49777034	19	#Н/Д
504073212,4	187808402,6	#Н/Д

Рис. 3.10. Массив данных

Неизвестные коэффициенты находятся в первой строке и располагаются справа налево. Таким образом, в нашем примере коэффициенты[1] равны:

$a \approx 12790{,}26;\ b_1 \approx 12{,}93;\ b_2 \approx 0{,}00077;$

$$y = 12790{,}26 + 12{,}93 \cdot x_1 + 0{,}00077 \cdot x_2{}^2.$$

Введем полученные значения в таблицу и распространим введенную формулу по всем ячейкам (рис. 3.11).

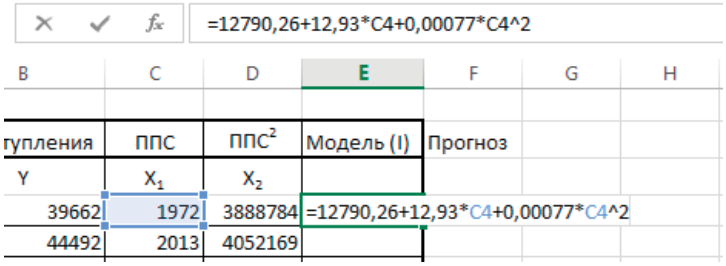


Рис. 3.11. Ввод формулы

[1] Коэффициенты даны с округлением по общим правилам.

На следующем этапе построим график фактических и модельных значений (рис. 3.12).

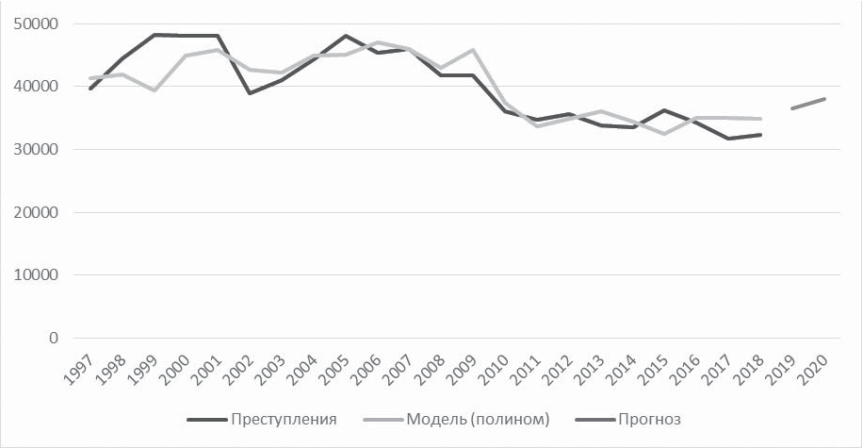


Рис. 3.12. График фактических и модельных значений. Полином (I)

На следующем этапе рассчитаем прогнозные значения уровня преступности на 2019 и 2020 г. при условии увеличения сотрудников ППСП ежегодно на 100 чел. Так, прогноз составит:

	ППСП	**Преступлений**
2019 г.	1666	36469
2020 г.	1766	38026

Чаще всего в качестве инструмента моделирования применяются линейные модели, но ограничиваться только их применением при исследовании не вполне корректно. Зачастую бывает так, что из-за сложности социально-правовых явлений и процессов их моделирование возможно только на основе нелинейных моделей.

3.2. Модели нелинейной регрессии

На предыдущем этапе мы рассмотрели ситуацию, когда в качестве зависимой (влияющей) переменной выступают несколько факторов. Зачастую бывает так, что линейные модели не в полной мере описывают социальное явление (процесс) либо полученные результаты свидетельствуют о низком качестве построенной модели.

Существуют несколько методов улучшения построенных моделей, в том числе когда линейные модели дают неудовлетворитель-

ные результаты. В этом случае можно перейти от линейной формы модели к нелинейной. Существует два класса таких моделей:

1. Нелинейные относительно независимых переменных, но линейные по параметрам (например, полиномиальная модель регрессии);

2. Нелинейные по независимым переменным и по оцениваемым параметрам (например, степенная модель регрессии).

Так, степенные функции применяются для исследования зависимости между объемом произведенной продукции и двух факторов производства – труда и капитала. Наряду с линейной одной из наиболее распространенных производственных функций является модель типа Кобба – Дугласа, имеющая форму степенной функции. В общем виде функцию этого типа можно представить так:

$$Y_i = AK_i^{\alpha_i} L_i^{\beta_i}, \alpha_i\, \beta_i \in [0,1], i = 1, \dots, N,$$

где α_i и β_i – оцениваемые параметры модели, Y_i – результаты, K_i – капитал и L_i – труд. Данную функцию удобно представить в логарифмическом виде[1]:

$$\ln Y_i = \ln A + \alpha \ln K_i + \beta \ln L_i + \ln \varepsilon_i.$$

Рассмотрим процедуру построения модели на основе исходных данных, полученных из Единой межведомственной информационно-статистической системы (ЕМИСС). На первом этапе из ЕМИСС (https://www.fedstat.ru) получим данные о валовом внутреннем продукте, основных средствах и численности занятых и составим следующую таблицу (таблица 3.1):

Таблица 3.1. Исходные данные для построения производственной функции Кобба – Дугласа

Год	Валовой внутренний продукт (в ценах 2016 г., млрд руб.)	Основные средства (млрд руб.)	Численность занятых, (млн чел.)
	Y	K	L
2011	81 750,6	238 100,9	70 856,6
2012	85 040,3	263 599,4	71 545,4

[1] Данная операция называется методом замены переменных.

2013	86 533,1	276 494,9	71 391,5
2014	87 170,2	286 081,4	71 539,0
2015	85 450,6	306 326,1	72 323,6
2016	85 616,1	340 098,5	72 392,6
2017	87 179,3	325 857,1	72 142,0
2018	89 626,6	348 492,7	72 354,4
2019	91 596,7	499 311,6	71 764,5
2020	89 138,9	507 269,2	70 460,8

На листе MS Excel эти данные будут располагаться в диапазоне данных [C3:E12]. Произведем логарифмирование переменных, для этого в соответствующие ячейки [F3:H12] введем значения, взятые с натуральными логарифмами. Для этого в ячейку [F3] введем следующее выражение: «=LN(C3)» (рис. 3.13).

Рис. 3.13. Ввод выражения

Фиксируем изменения нажатием клавиши «Enter» и, используя инструмент автозаполнения, распространяем введенное выражение по строкам и столбцам. Должны получиться следующие результаты (рис. 3.14):

ln(Y)	ln(K)	ln(L)
11,3114	12,3804	11,1684
11,3509	12,4822	11,1781
11,3683	12,5299	11,1759
11,3756	12,5640	11,1780
11,3557	12,6324	11,1889
11,3576	12,7370	11,1899
11,3757	12,6942	11,1864
11,4034	12,7614	11,1893
11,4252	13,1210	11,1811
11,3980	13,1368	11,1628

Рис. 3.14. Расчет значений

Далее рассчитаем неизвестные коэффициенты A, α, β методом наименьших квадратов с использованием функции MS Excel «ЛИНЕЙН». Данная функция возвращает параметры линейного приближения по МНК и имеет несколько аргументов:

1. Известные значения Y – зависимая переменная (валовой внутренний продукт);

2. Известные значения X – независимые переменные, или регрессоры (x_1 – основные средства и x_2 – численность занятых);

3. Логическое выражение, определяющее равенство нулю свободного члена (=1);

4. Логическое выражение, определяющее необходимость возврата дополнительной статистики (=1)[1].

Выделяем 5 строк и 3 (количество независимых переменных + 1) колонки пустых ячеек и вводим следующее выражение (рис. 3.15):

ln(Y)	ln(K)	ln(L)
11,3114	12,3804	11,1684
11,3509	12,4822	11,1781
11,3683	12,5299	11,1759
11,3756	12,5640	11,1780
11,3557	12,6324	11,1889
11,3576	12,7370	11,1899
11,3757	12,6942	11,1864
11,4034	12,7614	11,1893
11,4252	13,1210	11,1811
11,3980	13,1368	11,1628

=ЛИНЕЙН(F3:F12;G3:H12;1;1)

Рис. 3.15. Расчет неизвестных параметров

После нажатия клавиш «Ctrl+Shift+Enter» система заполнит выделенный диапазон следующими значениями (рис. 3.16):

[1] Строго говоря, для нашего примера возвращать дополнительную статистику не требуется, показатели качества модели и др. здесь не рассматриваются.

0,75131773	0,107053162	1,612531355
0,668755203	0,024314527	7,507181558
0,740680583	0,018373505	#Н/Д
9,996868219	7	#Н/Д
0,006749599	0,0023631	#Н/Д

Рис. 3.16. Результаты расчета

Значения коэффициентов располагаются слева направо, то есть A=1,61, α=0,107, β=0,751. Однако ввиду того, что при расчете мы приводили степенное выражение к натуральным логарифмам, нужно произвести обратное преобразование для коэффициента A. Для этого можно произвести следующие действия: число e (2,7182818284590452)[1] возвести в степень числа A:

$$2{,}7182818284590452^{1{,}61}.$$

Это действие можно произвести при помощи MS Excel (рис. 3.17):

	=I16^H16
1,612531355	2,718281828

Рис. 3.17. Расчет

Аналогичные результаты мы получим при использовании функции EXP, которая принимает только один аргумент – степень, в которую возводится основание e. Для этого в соседнюю ячейку введем следующее выражение (рис. 3.18):

=EXP(H16)	5,01549116
1,612531355	2,718281828

Рис. 3.18. Расчет

После нажатия клавиши «Enter» в ячейке появится число, идентичное рассчитанному ранее вручную.

Подставив рассчитанные значения в уравнение, получим следующее выражение: $Y = 5{,}01 \cdot K^{0{,}107} \cdot L^{0{,}751}$.

Рассчитаем модельные значения (рис. 3.19) и заполним следующую таблицу (таблица 3.2):

[1] Число e (число Эйлера) – это основание натурального логарифма, приблизительно равно 2,7182818284590452.

Рис. 3.19. Расчет модельных значений

Таблица 3.2. Исходные и модельные данные функции Кобба – Дугласа

	Валовой внутренний продукт (в ценах 2016 г., млрд руб.)	Модель
	Y	Y*
2011	81 750,6	83 193,5
2012	85 040,3	84 718,1
2013	86 533,1	85 014,7
2014	87 170,2	85 457,9
2015	85 450,6	86 794,1
2016	85 616,1	87 834,2
2017	87 179,3	87 205,4
2018	89 626,6	88 028,8
2019	91 596,7	90 922,9
2020	89 138,9	89 831,0

На следующем этапе построим диаграмму линейного типа с отображением фактических и модельных данных (рис. 3.20).

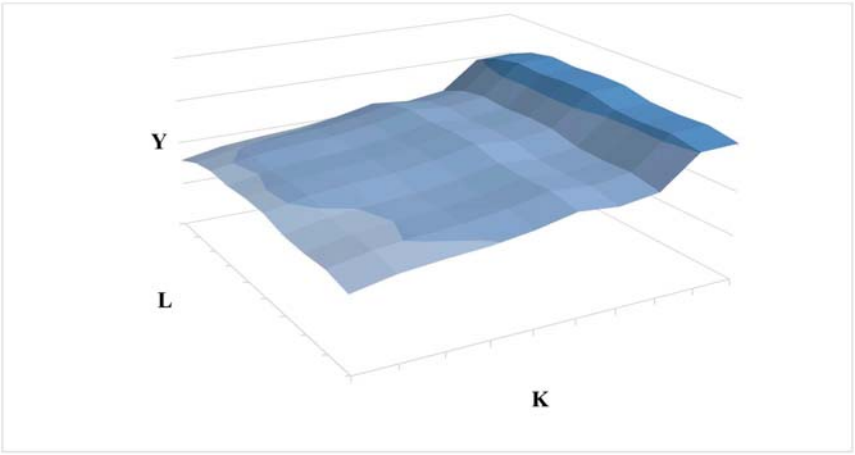


Рис. 3.20. Модель производственной функции Кобба – Дугласа

Рассмотрим процедуру построения полулогарифмической модели с использованием Calc LibreOffice, которая в общем виде выглядит как:

$$Y_i = \alpha + \beta \ln x_i + \varepsilon_i.$$

На первом этапе рассчитаем логарифм переменной x, для этого активируем первую ячейку [в нашем примере D2], и нажмем кнопку f_x. В появившемся окне «Мастера функций» находим в списке функцию LN, выбираем ее и нажимаем кнопку «Далее».

В следующем окне 

выбираем ячейку с переменной x [в нашем примере C2] и нажимаем «ОК». Используя функцию автозаполнения, аналогичную MS Office, заполняем диапазон данных [C2:C12] соответствующими значениями.

На следующем этапе нужно рассчитать неизвестные параметры модели α и β, для этого используем функцию «ЛИНЕЙН». В ячейках [B14] и [B15] будут располагаться, соответственно, значения α и β. Для того чтобы минимизировать выводимую информацию, используем функцию «ИНДЕКС» для извлечения из возвращаемого массива значений необходимые значения. В ячейку [B14] введем выражение «=ИНДЕКС(ЛИНЕЙН(B2:B12;D2:D12;1;0);1;2)», а в ячейку [B15] «=ИНДЕКС(ЛИНЕЙН(B2:B12;D2:D12;1;0);1;1)». Система рассчитает неизвестные значения параметров модели МНК α=-39745,5 и β=7585,1.

В первую ячейку [E2] диапазона ячеек [E2:E12] введем выражение «=$B$14+$B$15*D2». Напомним, что в ячейках [B14] и [B15] располагаются значения α и β. Используя функцию автозаполнения, заполняем диапазон данных [E2:E12] соответствующими значениями. Затем рассчитаем разницу между модельными и фактическими значениями в диапазоне ячеек [F2:F12], то есть из модельных значений [E2:E12] вычитаем [B2:B12]. Для этого введем в ячейку [F2] выражение «=E2-B2» и распространим введенное значение на весь диапазон. Должны получиться следующие данные, представленные в таблице 3.3.

Таблица 3.3. Данные полулогарифмической модели

№ п/п	y	x	ln (x)	Модель	Остаток
	1	**2**	**3**	**4**	**5**
1	4 473	354	5,86929691	4774,00852	301,00852

2	5 134	402	5,99645209	5738,49891	604,49891
3	5 632	413	6,02344759	5943,26371	311,26371
4	6 137	413	6,02344759	5943,26371	-193,73629
5	6 355	418	6,03548143	6034,54213	-320,45787
6	6 688	420	6,04025471	6070,74814	-617,25186
7	6 749	424	6,04973346	6142,64578	-606,35422
8	6 805	461	6,13339804	6777,25377	-27,74623
9	6 840	465	6,14203741	6842,78459	2,78459
10	6 979	488	6,19031541	7208,98020	229,98020
11	7 001	495	6,20455776	7317,01053	316,01053

На последнем этапе построим линейный график[1] с отображением исходных и модельных данных нашей модели. Рассмотрим подробную процедуру построения диаграммы с использованием Calc LibreOffice. Выбираем весь диапазон данных с заголовками [A1:F12], и после выбора пункта меню «Вставка-Диаграмма» или после нажатия кнопки  запустится окно следующего вида (рис. 3.21):

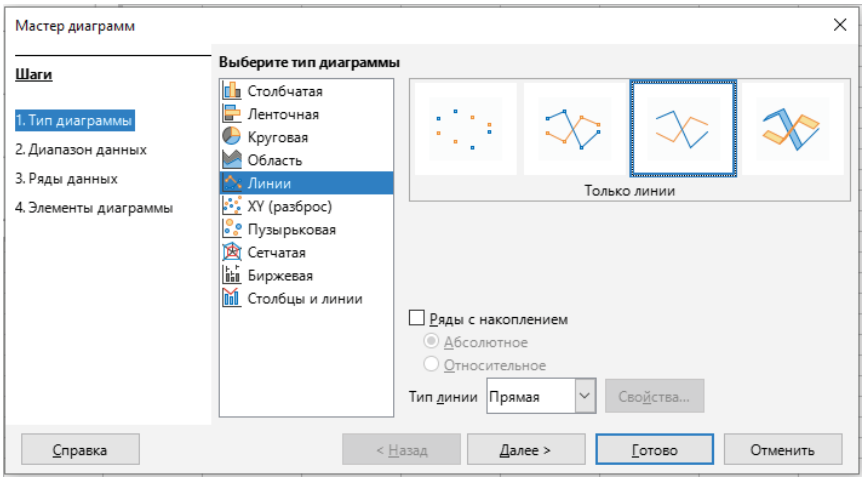


Рис. 3.21. Мастер диаграмм (тип диаграммы)

Выберем тип диаграммы, как указанно на рисунке, и нажмем кнопку «Далее». Появится окно следующего вида (рис. 3.22), где зададим диапазон данных, расположение ряда данных и наличие подписей.

[1] Calc LibreOffice (версия 7.3) не позволяет строить диаграммы поверхностного типа.

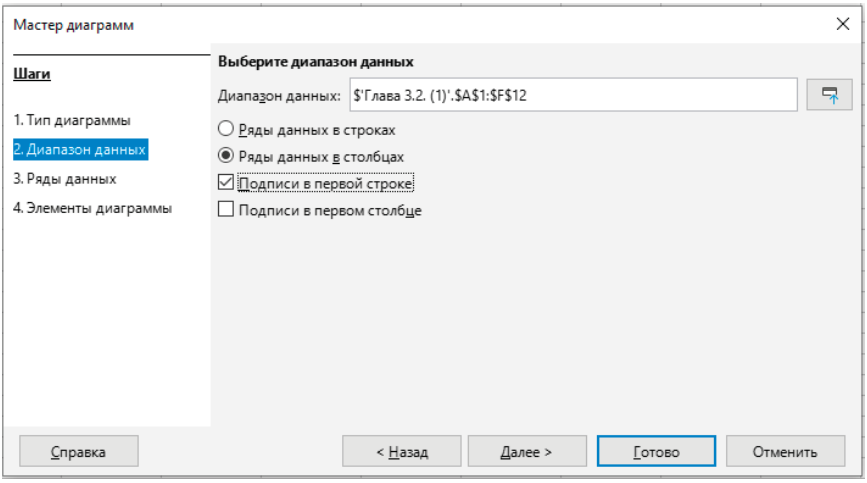


Рис. 3.22. Мастер диаграмм (диапазон данных)

После нажатия кнопки «Далее» появится окно, в котором можно настроить диапазоны данных для каждого ряда данных (рис. 3.23).

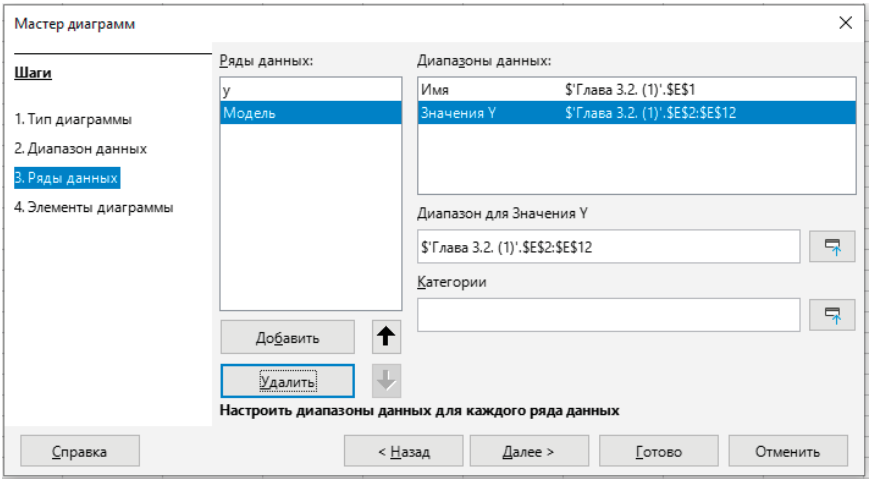


Рис. 3.23. Мастер диаграмм (ряды данных)

В данном окне можно удалить ненужные наборы данных, добавить ряды данных, добавить категории (подписи данных). После нажатия кнопки «Далее» запустится последнее окно «Мастера диаграмм» (рис. 3.24).

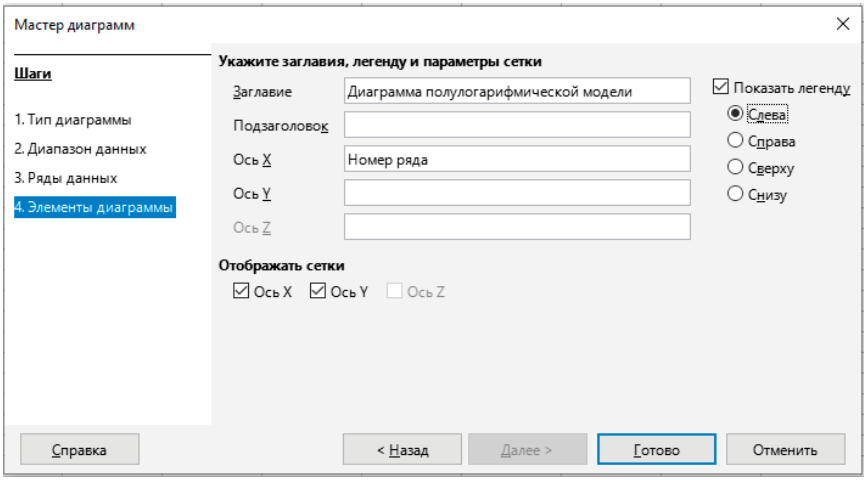


Рис. 3.24. Мастер диаграмм (элементы диаграммы)

Вводим данные, как показано на рисунке, и после нажатия кнопки «Готово» появится диаграмма линейного типа с отображением фактических и модельных данных (рис. 3.25).

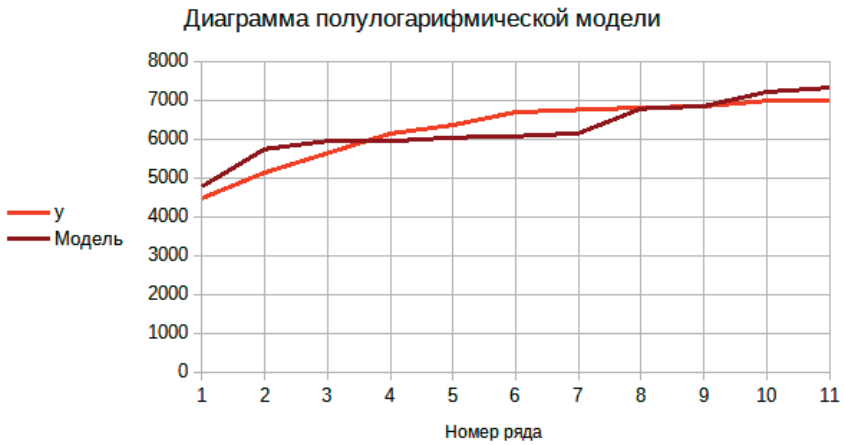


Рис. 3.25. Диаграмма полулогарифмической модели

Аналогичным образом можно построить регрессионные модели других видов. На следующем рисунке представлено сравнение

линейной, степенной и логарифмической моделей с фактическими данными (рис. 3.26).

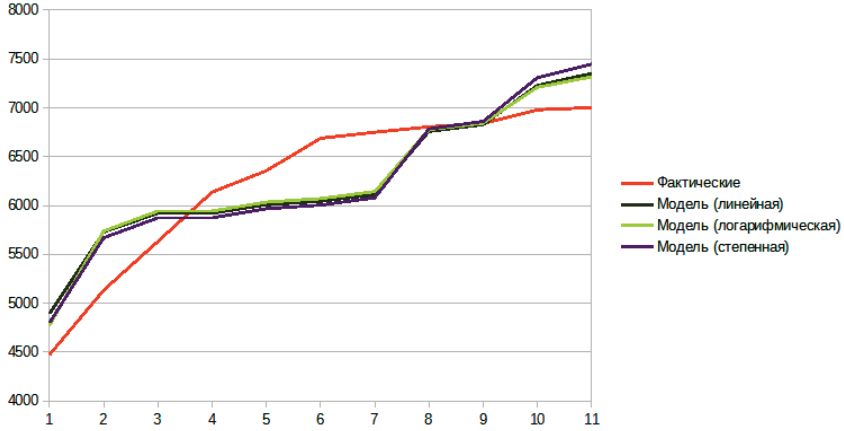


Рис. 3.26. Диаграмма моделей

Далее рассмотрим процедуру построения моделей регрессии со специфическими переменными (переменными, измеряемыми по дихотомической шкале).

3.3. Модели регрессии со специфическими переменными

В классическом регрессионном анализе значения представлены в основном в виде переменных, измеряемых по интервальной шкале. Однако зачастую возникает необходимость применения переменных, измеряемых по дискретной шкале. К данной категории и относятся номинальные переменные, которые могут принимать конечное число значений (например, Тяжесть преступления = {небольшой, средней, тяжкое, особо тяжкое}, Направленность преступления = {общеуголовная, экономическая} или Пол = {мужской, женский}). Причем переменная, принимающая два значения {0, 1}, называется дихотомической.

Модели регрессии с фиктивными переменными

Набор данных, находящихся у аналитика, чаще всего обладает несколькими характерными особенностями:

– различия, возникающие за счет влияния отдельных факторов при неизменном влиянии других;

– структурные различия объясняющих переменных, измеренных по интервальной или порядковой шкале.

Выявить их возможно за счет оценки влияния номинальной переменной на моделируемый показатель путем введения фиктивных переменных непосредственно в математическую модель[1]. Другими словами, они вводятся для исследования структуры данных и характера взаимосвязи переменных с другим уровнем измерения. Так, например, можно ожидать, что динамика преступности будет по-разному проявляться при изменении нормативной правовой базы, действующей в системе учета преступлений, а также варьироваться по регионам с разным уровнем социально-экономического развития либо природно-климатических особенностей.

Модель с фиктивной переменной выглядит следующим образом:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{i=1}^{m} c_i z_i + \varepsilon ,$$

где y – зависимая переменная; x – независимая переменная; z – фиктивная переменная, принимающая два значения [0, 1]. В графическом виде данную модель можно представить в виде следующего рисунка (рис. 3.27):

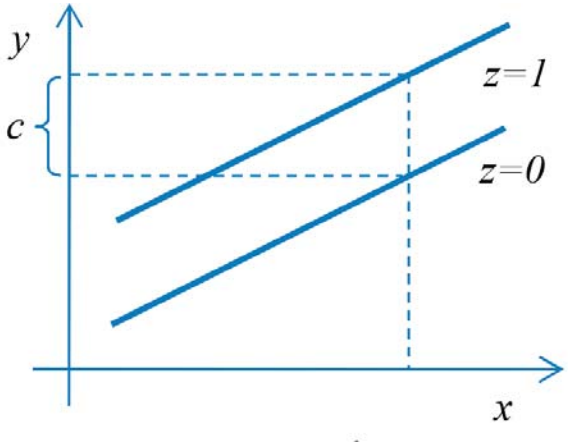


Рис. 3.27. Модель регрессии с фиктивной переменной сдвига

Рассмотрим следующий пример. В качестве независимой переменной (y) будут выступать преступления, совершенные в общественных местах, β_0, β_j, c_i – коэффициенты модели, z_i – фиктивная

[1] *Елисеева И. И., Курышева С. В.* Фиктивные переменные в анализе данных // Социология: 4М. 2010. № 30. С. 43–63.

переменная, а в качестве зависимых переменных (x_i) – численность сотрудников дорожно-патрульной службы полиции. Модель множественной регрессии, учитывающей два регрессора, будет представлена в виде:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z + \varepsilon,$$

где ε – случайная компонента, z – переменная, отражающая период до вступления в силу Федерального закона «О полиции» и после [0 – период с 1997 г. по 2010 г.; 1 – период с 2011 г. по 2019 г.]. Заполненная таблица представлена ниже (рис. 3.28).

	A	B	C	D
1		Совершено в общественных местах	Численность ДПС	Фиктивная переменная
2	1997	4544	1666	0
3	1998	3796	1781	0
4	1999	3588	1742	0
5	2000	3427	1638	0
6	2001	3424	1867	0
7	2002	3797	1602	0
8	2003	3186	1933	0
9	2004	4137	1943	0
10	2005	5062	1995	0
11	2006	5652	1607	0
12	2007	5702	1577	0
13	2008	6164	1649	0
14	2009	6160	1653	0
15	2010	7078	1491	0
16	2011	8122	1658	1
17	2012	8336	1824	1
18	2013	8208	1821	1
19	2014	10860	1765	1
20	2015	11021	1558	1
21	2016	9822	1572	1
22	2017	9000	1561	1
23	2018	9910	1395	1

Рис. 3.28. Таблица данных

Рассчитаем регрессию[1] и получим следующие результаты (рис. 3.29):

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,917636
R-квадрат	0,842056
Нормированный R-квадрат	0,82543
Стандартная ошибка	1092,123
Наблюдения	22

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*значимость F*
Регрессия	2	1,21E+08	60409364	50,64791	2,43E-08
Остаток	19	22661900	1192732		
Итого	21	1,43E+08			

	Коэффициенты	*артная ошибка*	*t-статистика*	*Значение*	*ижние 95%*	*рхние 95%*	*ижние 95,0*	*рхние 95,0*
Y-пересечение	11593,48	2748,236	4,218517	0,000465	5841,357	17345,61	5841,357	17345,61
Численность ДПС	-4,00065	1,584563	-2,52477	0,020633	-7,31718	-0,68412	-7,31718	-0,68412
Фиктивная переменная	4394,466	500,4851	8,780412	4,09E-08	3346,938	5441,993	3346,938	5441,993

Рис. 3.29. Регрессия

Полученное уравнение регрессии представим в виде $y = 11593.5 - 4 \cdot x + 4394.5$ и на ее основе построим прогноз на 2019–2020 гг. (рис. 3.30).

23	2018	9910	1395	1	10408
24	2019		1395	1	=11593,5-4*C24+4394,5*D24
25	2020		1395	1	10408

Рис. 3.30. Прогноз

Полученная диаграмма прогноза представлена на рис. 3.31.

Рис. 3.31. Диаграмма прогноза

[1] См. параграф 3.1. главы 3.

Модели дискретного выбора

В предыдущем разделе была рассмотрена технология построения регрессионных моделей с фиктивными переменными. В случае если дискретные переменные выступают в качестве зависимых, их называют моделям дискретного выбора. Зависимая переменная в них является дискретной, то есть может принимать бинарные и множественные значения.

Одной из наиболее распространенных моделей является Logit-модель, или модель логистической регрессии. Построение модели логистической регрессии в MS Excel включает в себя 7 последовательных этапов:

1. Упорядочение данных. Используя инструмент MS Excel «Сортировка», производится первичная сортировка данных по зависимой переменной (рис. 3.32).

y	x_1	x_2
0	536	116
0	766	230
0	963	102
0	301	470
0	593	980
0	221	824
0	630	133
1	367	149
1	235	177
1	787	539
1	634	207
1	150	392
1	539	733
1	120	342

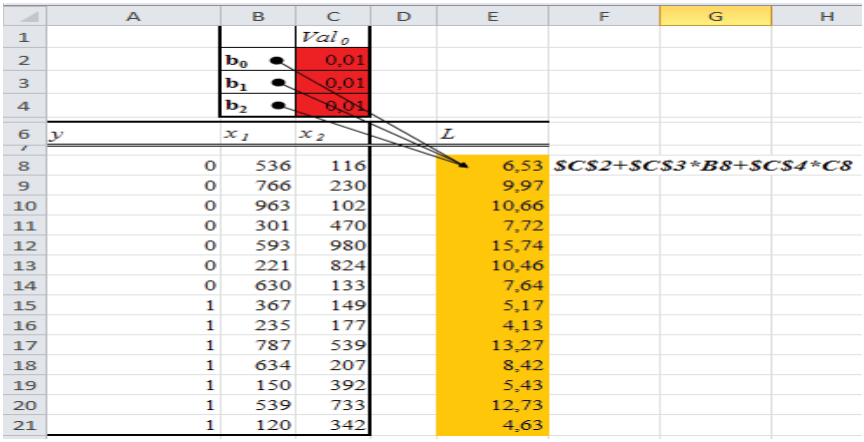
Рис. 3.32. Первичная сортировка данных

2. Расчет значений Logit-модели для набора входящих параметров $x_1,x_2,...,x_k$:

$$LogReg=L=b_0+b_1x_1+ b_2x_2+... +b_kx_k.$$

При помощи встроенного инструмента MS Excel «Поиск решения» оптимизируются коэффициенты $b_0,b_1,b_2,...,b_k$, которые произвольно устанавливаются со значением 0,01 (рис. 3.33).

Для примера рассмотрим процедуру построения Logit-модели с двумя независимыми переменными x_1, x_2.

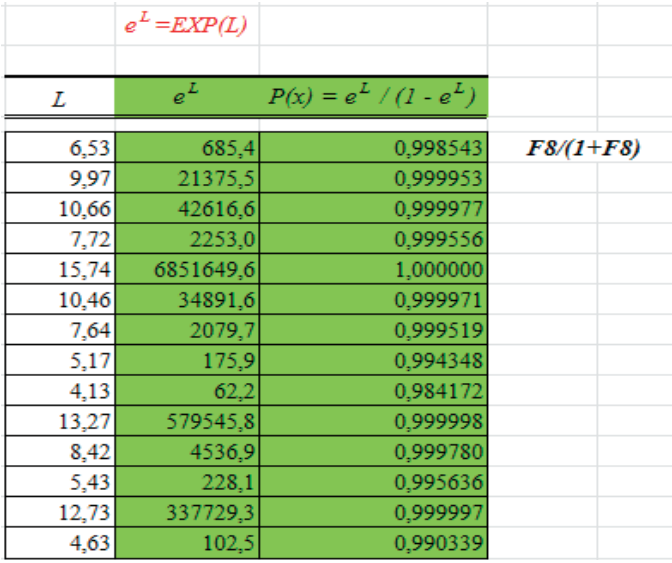


	A	B	C	D	E	F	G	H
1			Val_0					
2		b_0	0,01					
3		b_1	0,01					
4		b_2	0,01					
6	y	x_1	x_2		L			
8	0	536	116		6,53	SCS2+SCS3*B8+SCS4*C8		
9	0	766	230		9,97			
10	0	963	102		10,66			
11	0	301	470		7,72			
12	0	593	980		15,74			
13	0	221	824		10,46			
14	0	630	133		7,64			
15	1	367	149		5,17			
16	1	235	177		4,13			
17	1	787	539		13,27			
18	1	634	207		8,42			
19	1	150	392		5,43			
20	1	539	733		12,73			
21	1	120	342		4,63			

Рис. 3.33. Расчет первичных значений модели

3. Расчет значения e^L. Число e является основанием натурального логарифма и приблизительно равно 2,71828. Данное число рассчитывается для каждого значения данных e^L (рис 3.34).

4. Расчет вероятности события $P(x)$ по формуле $P(x) = \frac{e^L}{1+e^L}$ представлен на следующем рисунке (рис. 3.34):



	e^L =EXP(L)		
L	e^L	$P(x) = e^L / (1 - e^L)$	
6,53	685,4	0,998543	F8/(1+F8)
9,97	21375,5	0,999953	
10,66	42616,6	0,999977	
7,72	2253,0	0,999556	
15,74	6851649,6	1,000000	
10,46	34891,6	0,999971	
7,64	2079,7	0,999519	
5,17	175,9	0,994348	
4,13	62,2	0,984172	
13,27	579545,8	0,999998	
8,42	4536,9	0,999780	
5,43	228,1	0,995636	
12,73	337729,3	0,999997	
4,63	102,5	0,990339	

Рис. 3.34. Расчет значения eL и вероятности события P(x)

5. Расчет функции правдоподобия *PP*. Вероятность $Pr\ (Y_i=y_i|X_{1i}, X_{2i},..., X_{ki})$ – вероятность того, что предсказанная зависимая переменная y_i равна значению Y_i с учетом значений независимых переменных $X_{1i}, X_{2i},..., X_{ki}$. В сокращенном виде данное выражение записывается как $Pr(Y=y|X)$ и рассчитывается по формуле: $Pr\ (Y=y|X)=P(X)^{Y*}[1-P(X)](1-Y)$. Прологарифмировав обе части, получим: $ln[Pr(Y=y|X)]=y*ln[P(X)]*(1-y)*ln[[1-P(X)]]$. Функция правдоподобия (*PP*) представляет собой сумму значений $ln[Pr(Y=y|X)]$ и рассчитывается так:

$$PP = \sum Y_i \cdot P(X_i) + (1 - Y_i)(1 - P(X_i)) \quad .$$

Расчет *PP* в MS Excel представлен на следующем рисунке:

L	e^L	$P(x) = e^L / (1 - e^L)$	$y*ln[P(X)]*(1-y)*ln[[1-P(X)]]$	
6,53	685,3982115	0,99854312	-6,531457943	*A8*LN(G8)+(1-A8)*LN(1-G8)*
9,97	21375,48535	0,99995322	-9,970046781	
10,66	42616,63717	0,999976536	-10,66002346	
7,72	2252,959581	0,999556336	-7,720443762	
15,74	6851649,608	0,999999854	-15,74000015	
10,46	34891,55145	0,999971341	-10,46002866	
7,64	2079,743817	0,999519403	-7,640480713	
5,17	175,9148375	0,994347563	-0,005668473	
4,13	62,17792293	0,984171686	-0,015954919	
13,27	579545,8161	0,999998275	-1,72549E-06	
8,42	4536,903455	0,999779634	-0,00022039	
5,43	228,1492454	0,995636032	-0,004373518	
12,73	337729,3115	0,999997039	-2,96095E-06	
4,63	102,5140641	0,990339477	-0,009707489	
		СУММА	-68,75841095	*СУММ(H8:H21)*

Рис. 3.35. Расчет функции правдоподобия

6. Расчет функции максимального правдоподобия. Используется «Поиск решения» MS Excel для нахождения коэффициентов $b_0,b_1,b_2,...,b_k$, который максимизирует функцию PP в ячейке «СУММА». Данный инструмент настраивает числа в определенных ячейках, в которых находятся значения коэффициентов $b_0,b_1,b_2,...,b_k$ для оптимизации (максимизации или минимизации) целевой функции. Эти ячейки будут скорректированы таким образом, чтобы максимизировать *PP*, который находится в ячейке «СУММА» (рис. 3.36). Для оптимизации целевой функции используем «Поиск решения» MS Excel для нелинейных и гладких задач методом обобщенного приведенного градиента (ОПГ).

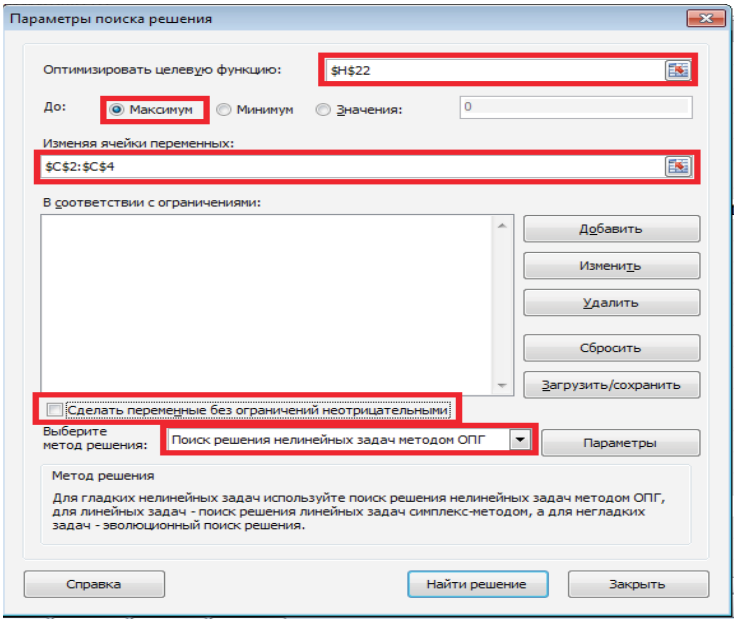


Рис. 3.36. Оптимизация целевой функции

Результаты работы приведены на следующем рисунке (рис. 3.37).

	A	B	C	D	E	F	G	H
1			Val_0					
2		b_0	0,009984089					
3		b_1	-0,000847362					
4		b_2	0,000443785					
5								
6	y	x_1	x_2		L	e^L	$P(x) = e^L / (1 - e^L)$	$y*\ln[P(X)]+(1-y)*\ln[1-P(X)]$
8	0	536	116		-0,39272	0,675215698	0,403061945	-0,515941932
9	0	766	230		-0,53702	0,584484529	0,368879922	-0,460259136
10	0	963	102		-0,76076	0,46731124	0,318481333	-0,383431637
11	0	301	470		-0,03649	0,964164672	0,490877718	-0,675067052
12	0	593	980		-0,05759	0,944034283	0,485605779	-0,664765341
13	0	221	824		0,188396	1,207311003	0,54696008	-0,791775034
14	0	630	133		-0,46483	0,628241392	0,38584045	-0,487500532
15	1	367	149		-0,23487	0,790670497	0,441549966	-0,817464092
16	1	235	177		-0,1106	0,895300226	0,472379106	-0,749973425
17	1	787	539		-0,41769	0,658566275	0,397069617	-0,923643657
18	1	634	207		-0,43538	0,647018631	0,392842326	-0,934346952
19	1	150	392		0,056843	1,058489961	0,514207007	-0,665129358
20	1	539	733		-0,12145	0,885635297	0,469674755	-0,755714834
21	1	120	342		0,060075	1,061916149	0,515014226	-0,663560756
22							СУММА	-9,488573738

Рис. 3.37. Результаты оптимизации

Значения функции максимального правдоподобия и коэффициентов $b_0, b_1, ..., b_k$ равны: $MPP=-9,488$; $b_0=0,009$; $b_1=-0,000847362$; $b_2=0,000443785$.

7. Тестирование результатов (рис. 3.38).

	Coeff
b_0	0,009984089
b_1	-0,000847362
b_2	0,000443785

x_1	161
x_2	949
$L=$	0,294710395
$P(x)=$	57%

Рис. 3.38. Тестирование произвольных данных

Задавая произвольные значения x_1 и x_2, получаем соответствующее значение $P(x)$.

Глава 4. Методы анализа анкетных данных

В главе 4 рассматриваются особенности формирования анкет в процессе исследования и методика создания анкет с учетом особенностей их подачи в электронном виде, а также технология создания сводных таблиц. Рассматриваются методы использования компьютерных технологий для анализа результатов анкетирования[1].

4.1. Методика сбора результатов анкетных опросов

Первичная компьютерная обработка результатов опроса – это последовательное заполнение каждого документа (анкеты, формы и т. д.), заполнение всех ответов на все вопросы в единую матрицу. Результатом являются возможность получения данных (как в абсолютных числах, так и в процентах) для каждой позиции, в каждой группе и для всей таблицы (обычно это называется линейным распределением); данные о связях между определенными ответами и той или иной парой (произвольной) или даже несколькими вопросами (эти данные принято называть корреляциями); разные типы коэффициентов и др.

При редактировании исследовательских работ вручную особенно важно знать (и лучше подумать об этом при создании исследовательской программы), нужна ли Вам просто информация о том, как все респонденты ответят на заданный вопрос, или Вас интересуют подробные ответы от представители той или иной группы.

Однако логика поиска часто требует дальнейшего исследования, установления связи между двумя или более характеристиками респондента. В общем случае мы говорим о том, что именно зависит от появления или распространения такого-то ответа на такой-то вопрос. Практика показывает, что определяющая и объяснительная характеристика в основном социально-демографическая – возраст, пол, образование, место работы или учебы, трудовой стаж и т. д.

На первом этапе исследования необходимо выбрать метод сбора информации от респондентов, наиболее часто используемым является анкетирование. Для этого этапа можно использовать открытые инструменты Yandex Forms или Google Forms. Рассмотрим процедуру создания анкеты при помощи Yandex Forms. Переходим

[1] Математические методы исследования социальных систем: курс лекций / И. В. Горошко, Б. А. Торопов, Ш. Х. Гонов. Москва: Академия управления МВД России, 2019. 80 с.

по ссылке (https://forms.yandex.ru/)[1], на появившейся странице мы можем создать форму из шаблонов (рис. 4.1) либо выбираем пункт «Создать форму» и создаем форму без заданных полей.

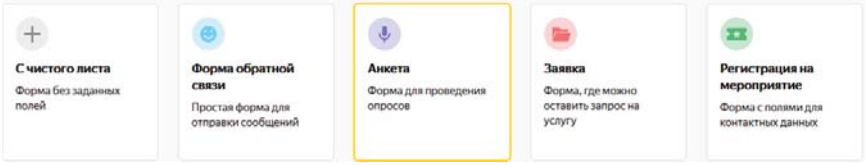


Рис. 4.1. Выбор шаблона

После создания формы в меню Yandex Forms имеется несколько разделов. Раздел «Конструктор» позволяет создавать, редактировать анкеты (рис. 4.2).

Рис. 4.2. Настройки теста

В разделе «Темы» можно выбирать как готовую тему, так и создавать свою, выбирая цветовую схему, шрифты, отображение текста и т. п. В разделе «Интеграция» можно задавать действие (перечень действий), выполняющееся всегда или при выполнении определенного условия (группы условий). В разделе «Настройки» можно задавать параметры, относящиеся ко всей форме. Например, задавать события после отправки или ограничить время доступа к форме (рис. 4.3).

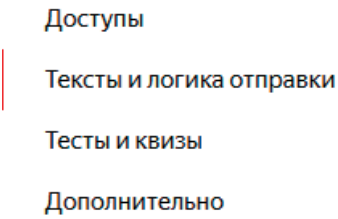


Рис. 4.3. Настройки формы

Подчеркнем, что, на наш взгляд, цель анкетирования заключается, прежде всего, в получении нового знания, установлении скрытых взаимосвязей в изучаемом процессе, объекте или явлении, взаимосвя-

[1] Для работы с опросными формами потребуется аккаунт в поисковых системах «Yandex»или «Google».

зей, которые, несомненно, должны содержаться в ответах на вопросы, сформулированные в анкете[1].

Поставленная цель достигается, если в основе проводимого анкетирования лежит определенная технология, базирующаяся на общих и специальных принципах. Общие принципы характерны для любого вида социологического исследования, в том числе и прикладного, специальные – относятся только к анкетированию.

Общими принципами опросных методов социологического исследования предлагаем считать принципы квантификации, репрезентативности и необходимой верификации[2].

Квантификация предполагает обязательное преобразование результатов опросов в цифровые данные для анализа и сравнения.

Принцип репрезентативности означает обязательность сопоставимости характеристик выборки (группы опрашиваемых респондентов) и характеристик изучаемой генеральной совокупности (например, определенной категории прокурорских работников).

Соблюдение принципа необходимой верификации обеспечивается в ходе итерационного проведения процедуры анкетирования, на каждом этапе которой возможна корректировка вопросов, уточнение параметров выборочной совокупности, переформулирование гипотез в зависимости от полученных результатов.

Что касается специальных принципов, то к ним относятся:

– обязательный учет специфических особенностей опрашиваемой аудитории (возраста, стажа работы, уровня образования, направленности деятельности и т. п.);

– обеспечение точности и корректности формулируемых вопросов, логической стройности и непротиворечивости построения анкеты;

– обязательное использование уточняющих и контрольных вопросов, ответы на которые должны подтверждать ответственное отношение опрашиваемого к заполнению анкеты;

– обеспечение толерантности сформулированных вопросов: респонденты, понимая общую постановку задачи, не должны догадываться о собственной позиции анкетера;

– чередование уровня сложности вопросов с постепенным его повышением;

---

[1] *Горошко И. В.* Технология проведения анкетных опросов // Вестник Университета прокуратуры Российской Федерации. № 3 (77). 2020. С.72–77.

[2] *Ядов В. А.* Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности. Москва: Омега-Л, 2007. 567 с.

– разумное ограничение объема анкеты: заполнение анкеты и нахождение ответов на поставленные в ней вопросы не должны казаться утомительной процедурой и отнимать слишком много времени.

Анкета чаще всего начинается с вступительного обращения к респондентам. Здесь указываются цель исследования, выходные результаты, общая методика опроса и т. п. В качестве примера ниже приведен фрагмент анкеты, разработанной авторским коллективом Академии управления МВД России в рамках научно-исследовательской работы по теме: «Разработка проекта концепции использования искусственного интеллекта в деятельности подразделений МВД России».

Уважаемые коллеги!

Авторский коллектив Академии управления МВД России в рамках научно-исследовательской работы по теме: «Разработка проекта концепции использования искусственного интеллекта в деятельности подразделений МВД России» проводит опрос сотрудников МВД России. Нам важно знать Ваше мнение по ряду вопросов...

1. Укажите Ваш пол.

– мужской ☐

– женский ☐

2. К какой возрастной группе Вы относитесь?

– Менее 28 лет ☐

– 29–38 лет ☐

– 39– 48 лет ☐

– Более 48 лет ☐

5. Насколько, на Ваш взгляд, целесообразно внедрение систем искусственного интеллекта в различные направления деятельности органов внутренних дел? (оцените каждый вариант по шкале от 1 до 5, где 1 – абсолютно нецелесообразно, 5 – необходимо)

– Расследование преступлений ☐

– Оперативно-розыскная деятельность ☐

– Экспертно-криминалистическая деятельность ☐

– Охрана общественного порядка ☐

– Организационно-аналитическая деятельность

– Информационно-аналитическая деятельность

– Кадровая работа

– Материально-техническое обеспечение

– Другое (укажите)

6. С каким родом деятельности связано Ваше место службы?

– Расследование преступлений

– Оперативно-розыскная деятельность

– Экспертно-криминалистическая деятельность

– Охрана общественного порядка

– Организационно-аналитическая деятельность

– Информационно-аналитическая деятельность

– Кадровая работа

– Материально-техническое обеспечение

– Другое (укажите)

9. На Ваш взгляд, готовы ли сотрудники и работники подразделения, где Вы проходите службу, по своей подготовке и квалификации для использования систем искусственного интеллекта в своей оперативно-служебной деятельности?

Оцените по шкале от 1 до 5, где 1 – абсолютно не готовы, а 5 – полностью готовы

Рассмотрим технологию создания указанной выше анкеты при помощи Yandex Forms. Вначале добавим вопросы. Для этого выбираем пункт «Один вариант»[1] (рис. 4.4).

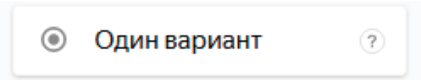


Рис. 4.4. Выбор варианта

[1] Данный элемент пользовательского интерфейса называется переключателем, или радиокнопкой (*англ. RadioButton*).

В появившейся форме (рис. 4.5) вводим информацию о первом вопросе и вариантах ответа.

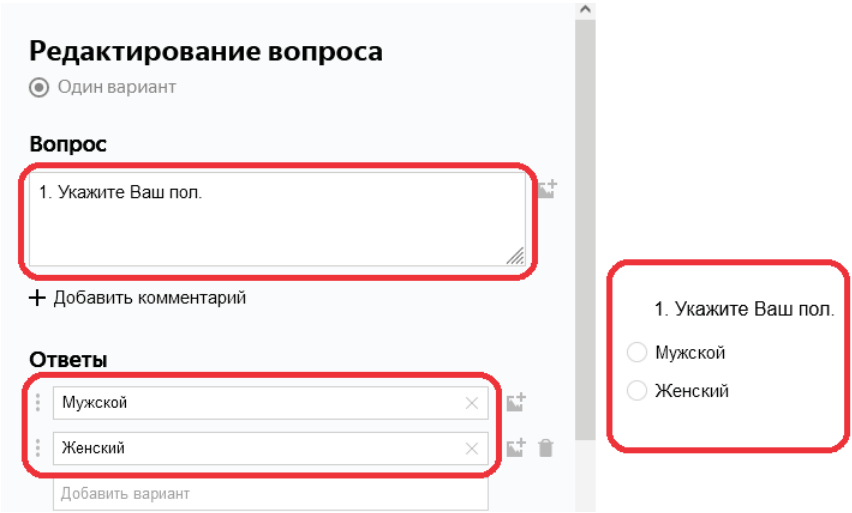


Рис. 4.5. Редактирование вопроса

Ниже в форме (рис. 4.6) можно задать дополнительные настройки вопроса (обязательность[1], скрытость вопроса, варианты сортировки ответов).

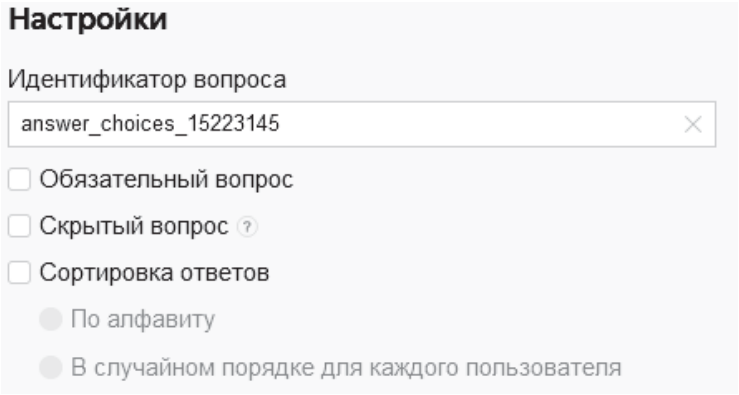


Рис. 4.6. Настройки вопроса

[1] При включении данной опции рядом с вопросом появляется красная звездочка, которая и сигнализирует об обязательности ответа на данный вопрос.

После нажатия кнопки «Сохранить» вопрос добавляется на страницу (рис. 4.7), в которой можно: 1) перемещать; 2) копировать; 3) удалять и 4) создавать условия показа.

Рис. 4.7. Добавление вопроса

Рассмотрим, как создаются условия показа вопросов. Например, следующий вопрос: «К какой возрастной группе Вы относитесь?», не задается респондентам, выбравшим в первом вопросе ответ «Женский». После нажатия кнопки 4) активируется следующее окно (рис. 4.8), в котором активируем пункт «При условии», указываем вопрос, математический оператор (равно, не равно) и варианты вопроса. Теперь если респондент выберет вариант ответа «Женский», то вопрос о возрасте будет ему недоступен.

Для обработки вопроса № 5 анкеты нам потребуется вариант «Оценка по шкале» (рис. 4.9), в котором задаем критерии и ответы (рис. 4.10).

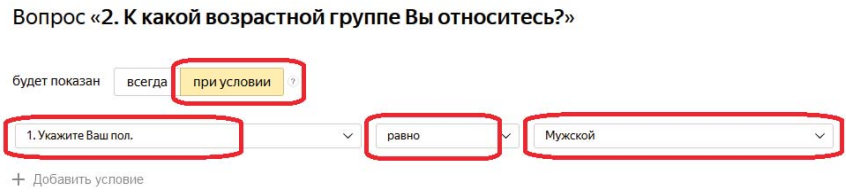


Рис. 4.8. Условия показа

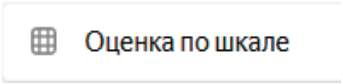


Рис. 4.9. Оценка по шкале

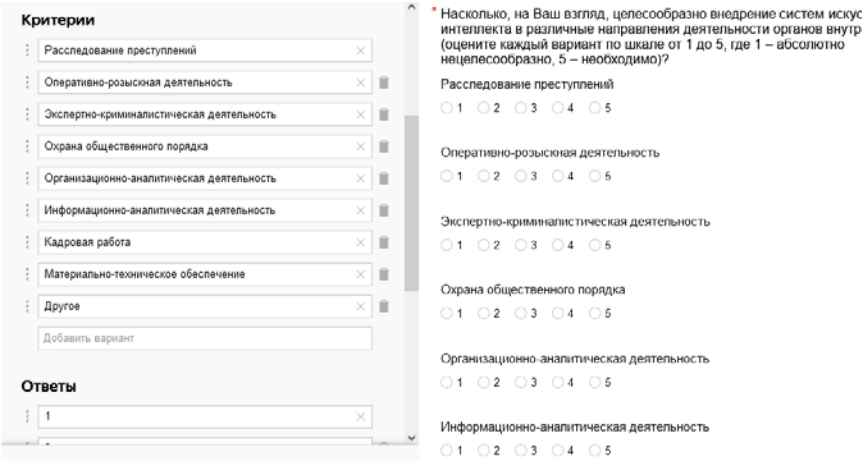


Рис. 4.10. Оценка по шкале

Следующий вопрос анкеты (№ 6) предполагает множественный выбор ответов[1] (рис. 4.11).

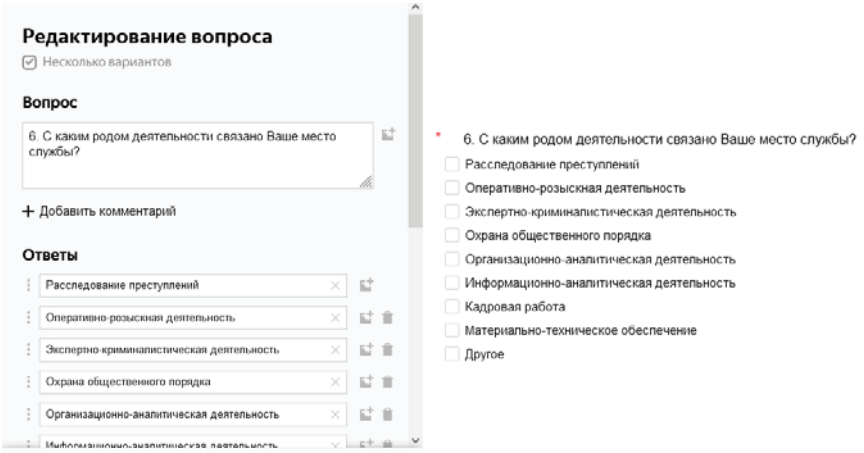


Рис. 4.11. Множественный выбор

Для вопроса № 11 анкеты указываем вариант «Длинный текст» или «Короткий текст».

После того, как анкета создана, можно посмотреть, как она выглядит за счет кнопки «Предпросмотр», и нажатием кнопки

[1] Данный элемент пользовательского интерфейса называется флажком, галочкой или чекбоксом *(англ. CheckBox)*.

«Публикация» дать к ней доступ для заполнения респондентами[1]. В окне, которое появляется после (рис. 4.12), можно поделиться ссылкой, отправить через соцсети и т. д., также можно снять форму с публикации.

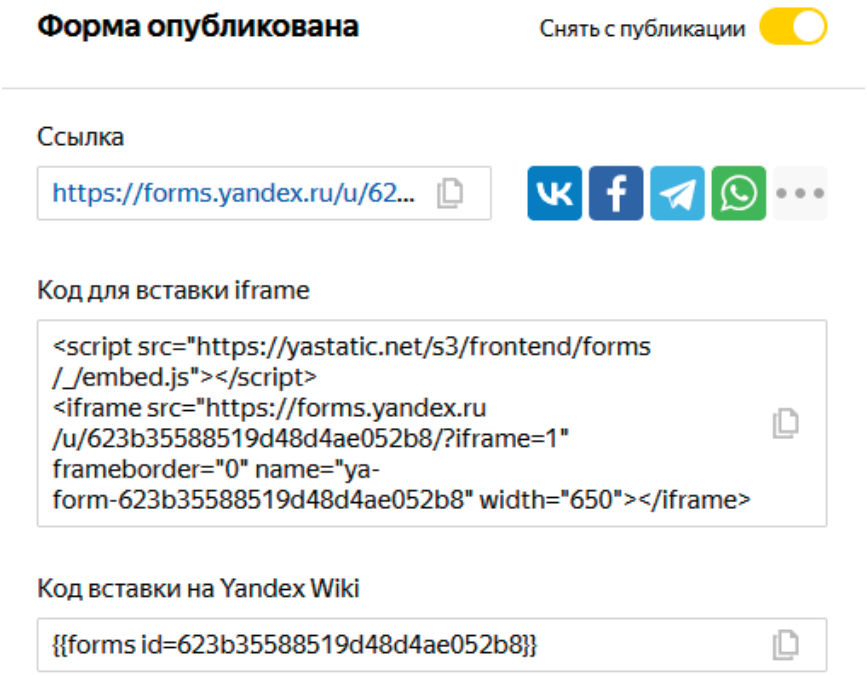


Рис. 4.12. Публикация формы

После публикации форма становится доступной для заполнения. После завершения процедуры сбора анкет автор может просматривать и обрабатывать результаты в разделе «Ответы», где будет указано количество поступивших ответов, а ниже в подменю «По ответам» можно просмотреть каждый ответ по отдельности. Более расширенные возможности предоставляет подменю «Сводка», где можно обрабатывать агрегированные данные: указывать интересующие вопросы, скачивать в разных форматах (xlsx, csv, json), фильтровать по дате ответа, а также включать дополнительные реквизиты в ответ (даты создания, обновления и т. п.). Кроме этого, в дан-

[1] Ссылка для созданной формы: https://forms.yandex.ru/u/623b35588519d48d4ae052b8/.

ном пункте можно просмотреть визуальные данные о статистике ответов (рис. 4.13).

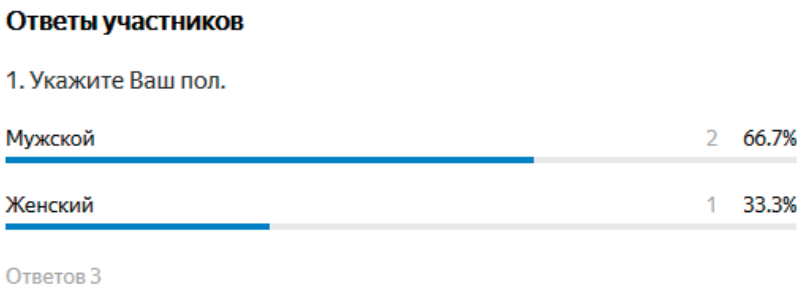


Рис. 4.13. Статистика ответов

Выбрав формат (xlsx) и нажав кнопку «Скачать», мы получим файл с ответами, который в дальнейшем можно обработать. Ответы в файле представлены в виде текстовых значений, которые нужно преобразовать в числовые значения, соответствующие номеру варианта[1].

Для этого создаем в файле лист («Описание»), в котором будет располагаться анкета с кодами ответов (рис. 4.14), также переименуем лист с ответами в («Данные»).

	A	B
1	**1. К какой возрастной группе Вы относитесь?**	
2	Мужской	**1**
3	Женский	**2**
5	**2. К какой возрастной группе Вы относитесь?**	
6	Менее 28 лет	**1**
7	29 - 38 лет	**2**
8	39 - 48 лет	**3**
9	Более 48 лет	**4**
11	**3. К какой группе Вы относитесь по стажу службы в органах внутренних дел?**	
12	Менее 5 лет	**1**
13	5 – 10 лет	**2**
14	11 – 15 лет	**3**
15	16-20 лет	**4**
16	Более 20 лет	**5**

Рис. 4.14. Фрагмент описания анкеты

[1] Для расчета корреляций текстовые значения не подходят, их нужно преобразовать в числовые.

Причем варианты должны строго соответствовать ответам, указанным в самой форме.

Создадим еще один лист, на котором и будут располагаться обработанные данные («Сборка»). В ячейку [A2] на листе «Сборка» введем следующее выражение: (=ВПР(Данные!A2;Описание!$A$2:$B$3;2;ЛОЖЬ)). Функция ВПР позволяет найти искомое значение [A2] в крайнем левом столбце таблицы [Описание.A2:B5] и возвращает значение, находящееся во [2-й] ячейке. Аналогичным образом вводим выражение для второго вопроса (=ВПР(Данные!B2;Описание!$A$6:$B$9;2;ЛОЖЬ)) и т. д.

4.2. Технология обработки результатов анкетных опросов

Корреляция (*от лат. correlatio* – отношение) или корреляционная зависимость – это статистическая связь двух или более случайных величин (или величин, которые могут рассматриваться как таковые с некоторой приемлемой степенью точности). В этом случае изменения значений одной или нескольких из этих переменных сопровождаются систематическим изменением значений той или иной переменной.

Выбор респондентами определенных ответов на разные вопросы также может быть взаимозависимым, то есть коррелированным. Числовая характеристика такой корреляции выражается коэффициентом корреляции (непараметрический коэффициент корреляции Фехнера, коэффициент ранговой корреляции Кендалла и Спирмена).

Для проверки статистических гипотез Карл Пирсон в 1900 г. предложил простой, универсальный и эффективный способ проверки соответствия между предсказаниями модели и экспериментальными данными. Его критерий хи-квадрат является наиболее важным и широко используемым статистическим тестом. С его помощью можно решить большинство проблем, связанных с оценкой неизвестных параметров модели и проверкой согласованности модели и экспериментальных данных.

Рассмотрим практическую реализацию авторской технологии обработки результатов анкетных опросов с использованием инструментария MS Excel.

Имеются данные о результатах анкетного опроса 343 респондентов по анкете[1], состоящей из трех вопросов: Предпочтение внедорожников? («Да»|«Нет»); Размер семьи («Больше 2 детей»|«Не больше 2

[1] Данные взяты с сайта Финансового университета при Правительстве Российской Федерации (http://www.fa.ru/org/dep/findata/Pages/Stat-book.aspx). Анализ данных

детей»); Доходы («Высокие»|«Низкие»). Необходимо выяснить, имеется ли взаимообусловленность между ответами респондентов:

а) на вопрос о количестве детей и на вопрос о предпочитаемом автомобиле;

б) на вопросы о количестве детей и о доходах.

На первом этапе формулируется нулевая гипотеза (H0) о том, что ответы респондентов не взаимообусловлены между собой. Также формулируется альтернативная гипотеза (H1) о том, что имеется взаимообусловленность между ответами респондентов. Задача исследования заключается в том, чтобы подтвердить или опровергнуть гипотезу.

На следующем этапе рассчитаем описательную статистику. Удобнее всего это сделать при помощи функций MS Excel СЧЁТ или в случае подсчета непустых записей СЧЁТЗ. Пример расчета и соответствующая формула представлены в таблице.

Таблица 4.1. Пример расчета описательной статистики

ВСЕГО	343	СЧЁТЗ($A$2:$A$344)
Предпочитают внедорожники		
- Да	148	СЧЁТЕСЛИ($A$2:$A$344;"=Да")
- Нет	195	СЧЁТЕСЛИ($A$2:$A$344;"=Нет")
Размер семьи		
- Больше 2 детей	186	СЧЁТЕСЛИ($B$2:$B$344;"=Больше 2 детей")
- Меньше 2 детей	157	СЧЁТЕСЛИ($B$2:$B$344;"=Не больше 2 детей")
Доходы		
- Высокие	246	СЧЁТЕСЛИ($C$2:$C$344;"=Высокие")
- Низкие	97	СЧЁТЕСЛИ($C$2:$C$344;"=Низкие")

Однако проще и быстрее реализовать данную функцию через инструмент MS Excel «Сводная таблица». Для этого открываем меню «Вставка» и выбираем элемент «Сводная таблица». Появляется следующее окошко (рис. 4.15), в котором выбираем «Таблица или диапазон»[1] и указываем, куда нужно поместить отчет сводной таблицы. Это может быть или новый лист (установлен по умолча-

в экономике: теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel: учебник / В. И. Соловьев. Москва: КноРус, 2018. 500 с.

[1] Чаще всего MS Excel правильно определяет диапазон данных, который необходимо включить в сводную таблицу, но иногда лучше выделить необходимый диапазон самостоятельно.

нию), или существующий лист, в этом случае необходимо указать диапазон ячеек, куда будет помещен отчет.

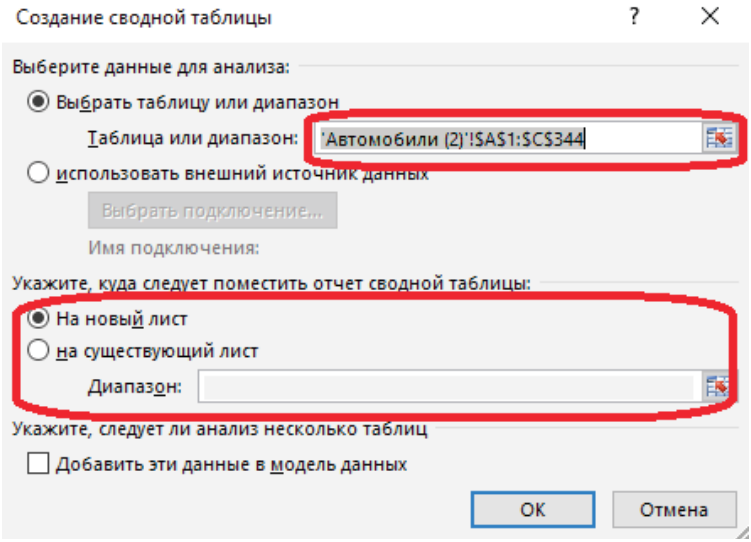


Рис. 4.15. Создание сводной таблицы MS Excel

Для удобства работы выберем ячейку на том же листе, на котором и находятся наши исходные данные. После нажатия кнопки «ОК» MS Excel поместит отчет сводной таблицы на лист с результатами анкетных опросов[1].

Для построения отчета нужно выбрать поля из списка полей сводной таблицы, который появляется справа (рис. 4.16).

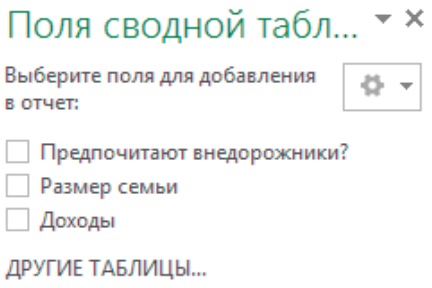


Рис. 4.16. Поля сводной таблицы

[1] На этом этапе отчет сводной таблицы не содержит данных.

Для этого нужное поле «перетаскиваем»[1] мышью в соответствующую область. Например, поле «Размер семьи» «перетаскиваем» в область «Строки», а поле «Предпочитают внедорожники» в область «Колонны». Кроме этого нам нужно указать и итоговые значения. Для этого одно из выбранных полей «перетаскиваем» в область «∑ Значения»[2], например, поле «Размер семьи», как это представлено на следующем рисунке (рис. 4.17).

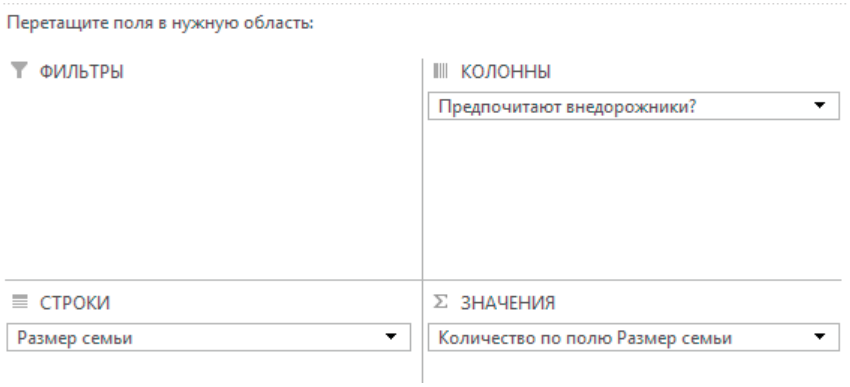


Рис. 4.17. Итоговые значения

На листе MS Excel появится таблица следующего вида (рис. 4.18):

Количество по полю Размер семьи	Названия столбцов		
Названия строк	Да	Нет	Общий итог
Больше 2 детей	138	48	186
Не больше 2 детей	10	147	157
Общий итог	**148**	**195**	**343**

Рис. 4.18. Сводная таблица

Для удобства дальнейших расчетов скопируем данные в отдельные ячейки листа. Примечание: вставлять нужно только значения[3], в противном случае скопируется макет сводной таблицы. Таким образом, у нас получается таблица с эмпирическим (наблюдаемым) распределением ответов.

[1] Более корректное выражение на английском языке звучит как «Drag&Drop» (бери-и-брось).

[2] Знак ∑ в математике означает сумму.

[3] Вставка – только значения.

На следующем этапе рассчитаем ожидаемые частоты (теоретические частоты), то есть теоретическое распределение ответов при полной независимости ответов на два вопроса. Для расчета теоретических частот можно использовать следующее выражение:

$$\sum \text{Col} * \sum \text{Row} / \sum \text{All},$$

где ∑ Col – количество по колонке, ∑ Row – количество по строке, ∑ All – всего анкет (респондентов).

Для нашего примера: Количество предпочитающих внедорожники*Количество «Больше 2 детей» / Общее количество респондентов. Подставляя полученные значения, получаем: 148*186/343.

Данные расчеты проще реализовать в MS Excel, где введенная формула будет представлена в следующем виде (рис. 4.19):

Наблюдаемые частоты			
	Да	Нет	Общий итог
Больше 2 детей	138	48	186
Не больше 2 детей	10	147	157
Общий итог	148	195	343
Ожидаемые частоты			
	Да	Нет	
Больше 2 детей	=G23*I21/I23		
Не больше 2 детей			

Рис. 4.19. Расчет ожидаемых частот

Введенное в ячейку выражение лучше написать следующим образом: =G$23*$I21/$I$23. То есть зафиксировать ячейки для того, чтобы можно было использовать функцию автозаполнения.

Далее рассчитаем наблюдаемое значение Хи-квадрат Пирсона по следующей формуле:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4.1),$$

где χ^2 – Хи-квадрат, O_i – наблюдаемые частоты, E_i – ожидаемые частоты.

Данный расчет реализуем в MS Excel. Введенное выражение будет выглядеть следующим образом (рис. 4.20):

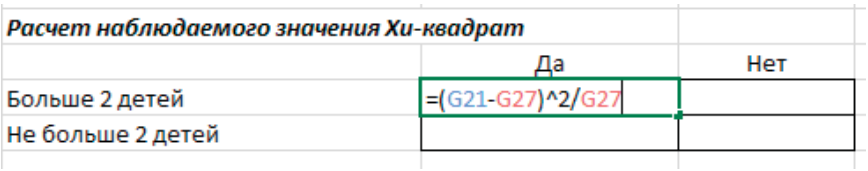


Рис. 4.20. Расчет наблюдаемого значения Хи-квадрат

Сумма полученных результатов и будет являться наблюдаемым значением Хи-квадрат (рис. 4.21).

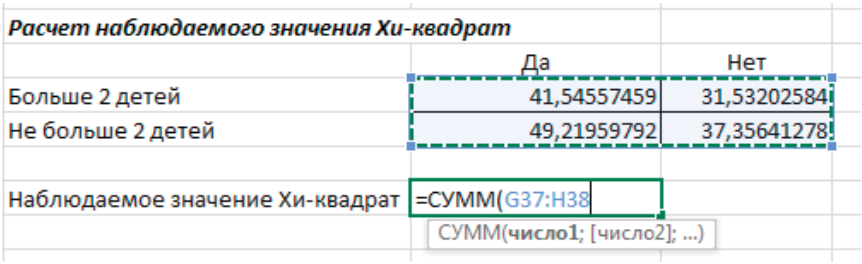


Рис. 4.21. Хи-квадрат

На последнем этапе сравнивается наблюдаемое значение χ^2 с критическим, для расчета которого используется функция MS Excel ХИ2.РАСП.ПХ, которая возвращает правостороннюю вероятность распределения χ^2. Данная функция принимает два аргумента, каждый из которых является обязательным:

1) уровень значимости. Для гуманитарных исследований обычно принимается равным 0,05;

2) число степеней свободы рассчитывается как df=(Row-1)*(Col-1).

Для нашего примера первый аргумент равняется 0,05, а второй – (2-1)*(2-1)=1. Таким образом, ХИ2.РАСП.ПХ(0,05;1) (рис. 4.22).

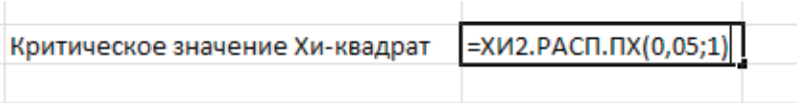


Рис. 4.22. Хи-квадрат

Так как критическое значение Хи-квадрат меньше, чем расчетное, нулевая гипотеза о невзаимозависимости ответов отвергается.

Рассмотрим процедуру создания сводной таблицы средствами LibreOffice Calc. Она очень похожа на методику, рассмотренную ранее с помощью MS Excel. Здесь также выбирается диапазон данных и пункт меню «Вставка» – «Сводная таблица» (рис. 4.23), но указывать на данном этапе, куда поместить отчет сводной таблицы, нельзя.

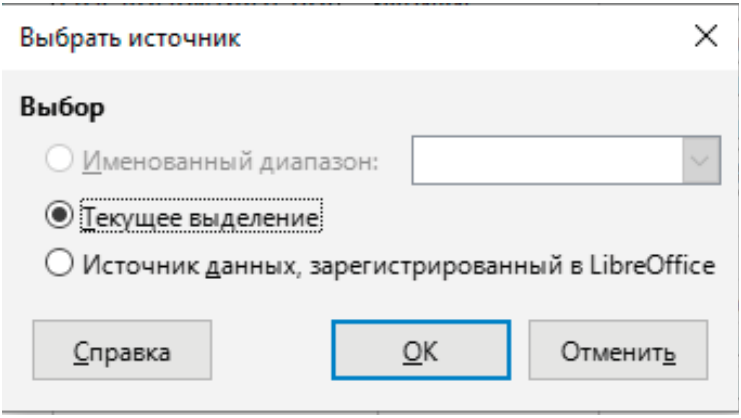


Рис. 4.23 Создание сводной таблицы LibreOffice Calc

После нажатия кнопки «ОК» активируется следующая форма (рис. 4.24). Для построения отчета также нужно выбрать поля из списка полей сводной таблицы, который находится справа. На рис. 4.24 номерами обозначены примерные этапы:

1) доступное поле «Предпочитают внедорожники» разместим в полях столбцов;

2) поле «Размер семьи» разместим в полях строк;

3) одно из доступных полей (в нашем примере – «Размер семьи») разместим в полях данных. По умолчанию в данном поле активна функция «Сумма». Это значит, что в итогах посчитается сумма значений данного поля. Для того чтобы переключиться на другую функцию, на поле данных на значении «Сумма – размер семьи» производим двойной щелчок мышью[1]. Активируется окно с дополнительными настройками для поля данных (рис. 4.25), в котором выбираем функцию «Количество»;

[1] Или при активном поле нажать кнопку «Enter».

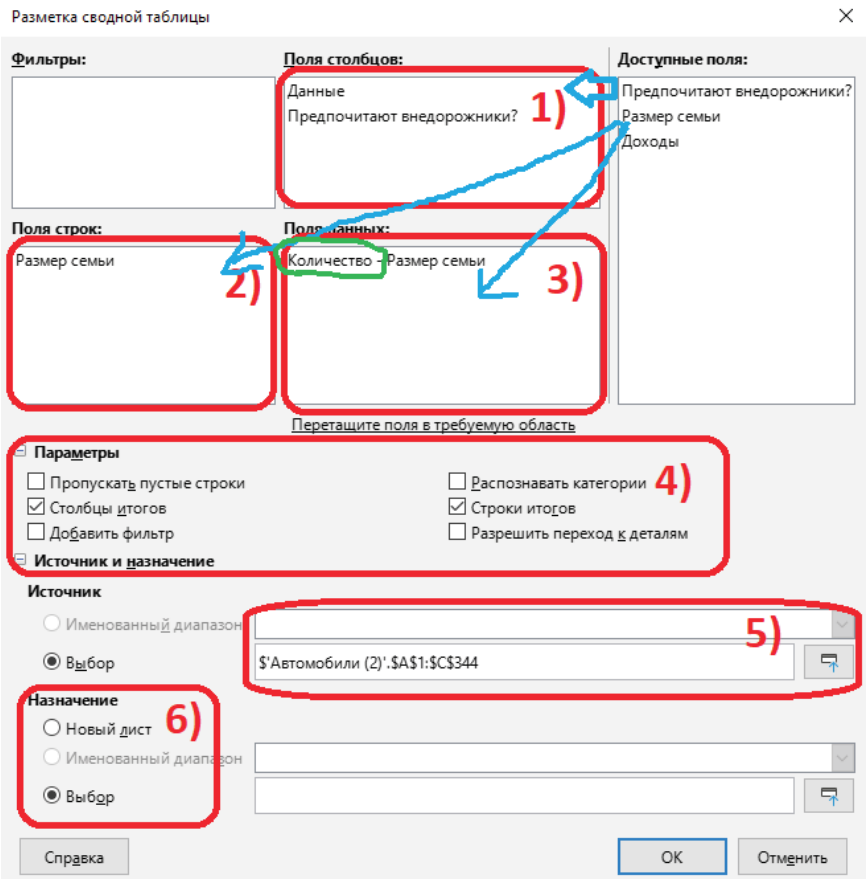


Рис. 4.24. Разметка сводной таблицы LibreOffice Calc

4) после нажатия на значок «+» в пункте «Параметры» активируются дополнительные параметры, в которых можно установить «птички» напротив нужных пунктов;

5) после нажатия на значок «+» в пункте «Источник» активируются параметры выбора источника данных. Здесь можно поменять источник;

6) кроме этого, можно задать параметры назначения, то есть куда будут выводиться результаты (сводная таблица). Есть возможность вывести на новый лист, именованный диапазон[1] и определенный диапазон (пункт «Выбор»).

[1] Здесь не рассматривается.

После нажатия кнопки «ОК» система разместит в указанном в п. 6 месте сводную таблицу следующего вида (таблица 4.2):

Таблица 4.2. Сводная таблица

Количество и размер семьи	Данные		
Размер семьи	Да	Нет	**Итог Результат**
Больше 2 детей	138	48	**186**
Не больше 2 детей	10	147	**157**
Итог Результат	**148**	**195**	**343**

Дальнейшая процедура расчета идентична расчету с использованием MS Excel, и на ней мы останавливаться не будем.

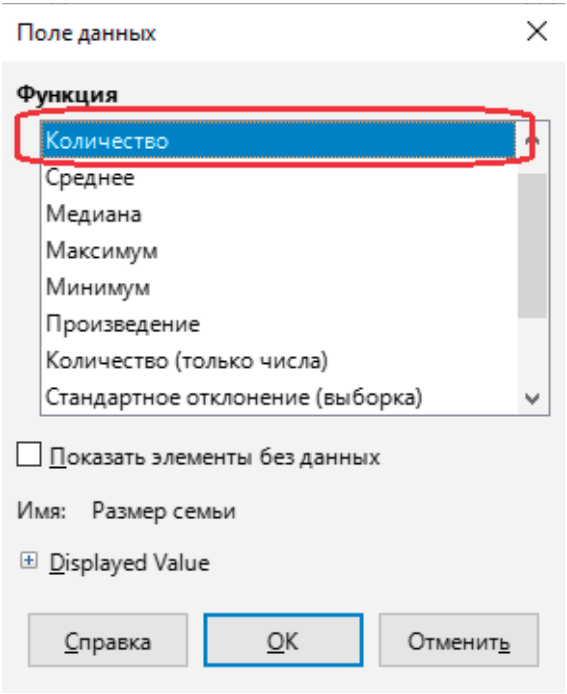


Рис. 4.25. Поля данных сводной таблицы LibreOffice Calc

В следующем параграфе рассмотрим вопросы оценки согласованности мнений респондентов (экспертов).

4.3. Методы оценки согласованности мнений респондентов

Прикладное социологическое исследование в области правоохранительной деятельности направлено на решение задач совершенствования системы управления органами внутренних дел, создания стабильного правоохранительного пространства, оценки и улучшения результатов деятельности органов внутренних дел.

Одной из задач научного исследования является оценка согласованности мнений респондентов. Данная оценка чаще всего проводится по вопросам, предполагающим ранжирование. Наиболее простым и распространенным методом выступает оценка за счет расчета коэффициента конкордации Кендалла (W), показывающего степень согласованности мнений экспертов (респондентов) и рассчитываемого как:

$$W = \frac{S}{\frac{1}{12}m^2(n^3 - n) - m\sum T_i} \quad (4.2),$$

где S – разность между суммой квадратов рангов по каждому признаку и средним квадратом суммы рангов по каждому признаку, m – число респондентов, n – число признаков. S и T_i – рассчитываются как:

$$S = \sum P^2 - \frac{(\sum P)^2}{n} \quad (4.3),$$

$$T_i = \frac{1}{12}\sum(t_i^3 - t_i) \quad (4.4),$$

где P – ранги, t_i – число повторений каждого ранга в i-ом ряду.

Для оценки статистической значимости коэффициента конкордации применяется критерий χ-квадрат по формуле:

$$\chi^2 = \frac{S}{\frac{1}{12}mn(n+1) + \frac{1}{n-1}\sum T_i} \quad (4.5).$$

Рассмотрим технологию оценки ответов экспертной группы на вопросы анкеты, предложенной в параграфе 1 главы 4 настоящего пособия. В данной анкете вопрос № 5: «Насколько, на Ваш взгляд, целесообразно внедрение систем искусственного интеллекта в различные направления деятельности органов внутренних дел», предполагает ранжирование.

Приведем фрагмент листа с данными (рис. 4.26), которые располагаются на листе «Данные».

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		1. Пол	2. Возрастная группа	3. Стаж службы в органах внутренних дел	4. Познания о технологиях и системах искусственного интеллекта	5 - Расследование преступлений	5 - Оперативно-розыскная деятельность	5 - Экспертно-криминалистическая деятельность	5 - Охрана общественного порядка	5- Организационно-аналитическая деятельность	5 -Информационно-аналитическая деятельность	5 - Кадровая работа	5 - Материально-техническое обеспечение
2	1	1	3	5	3	5	5	5	5	4	5	3	1
3	2	2	3	4	2	5	5	5	5	5	5	4	5
4	3	2	3	5	2	5	5	5	5	3	4	3	3
5	4	1	3	5	2	5	5	5	3	3	3	4	4
6	5	1	3	5	2	5	5	5	3	4	4	1	4
7	6	2	3	5	2	5	5	5	5	5	5	5	5
8	7	2	3	4	2	5	4	4	3	4	4	2	2

Рис. 4.26 Фрагмент листа с данными

На отдельном листе создадим таблицу (рис. 4.27) с названием «Расчет».

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		АНК5_1	АНК5_2	АНК5_3	АНК5_4	АНК5_5	АНК5_6	АНК5_7	АНК5_8	ОБР5_1	ОБР5_2	ОБР5_3	ОБР5_4	ОБР5_5	ОБР5_6	ОБР5_7	ОБР5_8	T.(ОБР5)
2	1																	
3	2																	
4	3																	
5	4																	
6	5																	
7	6																	
8	7																	
9	8																	
10	9																	

Рис. 4.27 Фрагмент листа с расчетами

В колонке [A], начиная со строки [2], будут располагаться номера анкет, а в первой строке наименования переменных (АНК5_1... АНК5_8) в диапазоне колонок [B1...I1] – это варианты ответа на соответствующие пункты вопроса № 5 анкеты. Переменные в первой строке (ОБР5_1...ОБР5_8), соответственно, диапазон колонок [J1...Q1] – это расчетные значения.

На первом этапе необходимо рассчитать средний ранг оценки по конкретному вопросу в диапазоне рангов. Для это-

го в ячейку [B2] вводим следующее выражение: «=РАНГ.СР(Данные!F2;Данные!$F2:$M2)»[1], и распространяем введенное значение на весь диапазон ячеек (рис. 4.28).

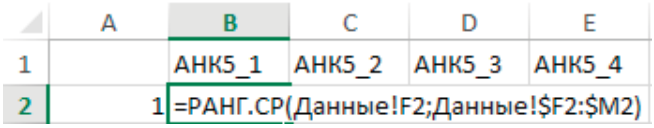


Рис. 4.28 Ввод формулы

Функция «РАНГ.СР» возвращает ранг оценки конкретного вопроса в диапазоне оценок, выставленных респондентом, то есть его величину относительно других значений в списке. В отличие от «РАНГ» функция «РАНГ.СР» возвращает среднее, если несколько оценок имеют одинаковый ранг[2].

На втором этапе рассчитаем сумму рангов. Для этого в ячейку под рассчитанными значениями [B618] введем следующее выражение: «=СУММ(B2:B617)» (рис. 4.29).

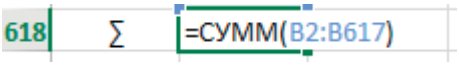


Рис. 4.29 Ввод формулы

Распространим введенную формулу на остальные ячейки [C618:I618], в результате получим следующие значения:

3103,5	2830	2238,5	2676,5	2279	2047	3590	3411,5

В следующей ячейке рассчитаем сумму квадратов. Самый простой способ – предыдущий рассчитанный показатель возвести в квадрат. Для этого в ячейке [B619] введем следующее выражение: «=B618^2» (рис. 4.30).

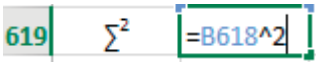


Рис. 4.30 Ввод формулы

[1] Чтобы иметь возможность использовать автозаполнение, как по вертикали, так и по горизонтали фиксируем диапазон ячеек только для колонок, то есть знак $ стоит только перед буквой.

[2] Используем функцию «РАНГ.СР» вместо «РАНГ», так как респондент может выставить одинаковые оценки для двух и более вариантов.

Далее рассчитаем показатель S по формуле (4.3). Для этого в ячейку [B621] введем следующее выражение: «=СУММ(B619:I619)-((СУММ(B618:I618)^2)/8)». Однако существует более простой способ расчета с использованием функции «КВАДРОТКЛ». Для этого в ячейку [B621] введем следующее выражение: «КВАДРОТКЛ(B618:I618)». На рисунке ниже представлены ячейки с расчетом и формулами (рис. 4.31).

620	S=	2253757	СУММ(B619:I619)-((СУММ(B618:I618)^2)/8)
621	S=	2253757	КВАДРОТКЛ(B618:I618)

Рис. 4.31 Ввод формулы

Обратим внимание, что результаты абсолютно идентичны. Таким образом, для расчета целесообразно использовать наиболее быстрый второй метод.

На третьем этапе рассчитаем показатель t_i. Это число повторений каждого ранга в *i*-ом ряду. Для этого в ячейку [J2] вводим выражение «=СЧЁТЕСЛИ(Данные!$F2:$M2;1)», в ячейку [K2] «=СЧЁТЕСЛИ(Данные!$F2:$M2;2)» и т. д. Данная операция призвана рассчитать, какую оценку для каких вариантов указал респондент. Распространяем введенное значение на весь диапазон ячеек (рис. 4.32).

J	K	L	
ОБР5_1	ОБР5_2	ОБР5_3	О
=СЧЁТЕСЛИ(Данные!$F2:$M2;1)			

Рис. 4.32 Ввод формулы

Далее рассчитаем T_i по формуле (4.4). Для этого в ячейку [R2] введем следующее выражение: «=((J2^3-J2)+(K2^3-K2)+(L2^3-L2)+(M2^3-M2)+(N2^3-N2)+(O2^3-O2)+(P2^3-P2)+(Q2^3-Q2))/12». А в ячейке [R618] рассчитаем сумму. Для этого введем следующее выражение: «=СУММ(R2:R617)».

На четвертом этапе рассчитаем коэффициент конкордации (W) по формуле (4.2). Для этого необходимо определить значения m – число респондентов (617), n – число признаков (8). Подставляя полученные значения в формулу (4.2), получаем:

$$W = \frac{2253757}{\frac{1}{12}616^2(8^3 - 8) - 616 \cdot 9808{,}5} \approx 0{,}228.$$

Эти же расчеты можно произвести в MS Excel. Для этого в ячейку [B622] введем следующее выражение: «=B621/((1/12*(616^2)*((8^3)-8))-616*R618)».

Однако для повышения универсальности наших расчетов произведем следующие действия. В ячейку [B623] поместим значение переменной m, а в ячейку [B624] значение переменной n. Для этого в соответствующие ячейки введем следующие выражения [B624] («=СЧЁТ(B2:B617)») и [B624] («=СЧЁТ(B617:I617)»).

Попробуем теперь ввести формулу со ссылкой на соответствующие ячейки (рис. 4.33).

622	W=	0,227764586	B621/((1/12*(616^2)*((8^3)-8))-616*R618)
623	m=	616	СЧЁТ(B2:B617)
624	n=	8	СЧЁТ(B617:I617)
625	W=	0,227764586	B621/((1/12*(B623^2)*((B624^3)-B624))-B623*R618)

Рис. 4.33 Ввод формулы

Обратите внимание, что результаты, рассчитанные двумя подходами, абсолютно идентичны.

На последнем этапе произведем оценку статистической значимости полученного коэффициента конкордации. Для этого рассчитаем значение критерия χ-квадрат по формуле (4.5). Введем в ячейку [B626] следующее выражение: «B621/((1/12*B623*B624*(B624+1))+((1/(B624-1))*R618))», а в ячейку [B627] критическое значение критерия χ-квадрат «ХИ2.ОБР.ПХ(0,05;B624-1)» (рис. 4.34).

626	$\chi^2_{расч}$	442,1546503	B621/((1/12*B623*B624*(B624+1))+((1/(B624-1))*R618))
627	$\chi^2_{крит}$	14,06714045	ХИ2.ОБР.ПХ(0,05;B624-1)

Рис. 4.34 Ввод формулы

Так как расчетное значение критерия χ-квадрат (442,2) больше критического (14,1), то полученные результаты статистически значимы. Однако само значение W = 0,228 говорит о слабой степени согласованности мнений экспертов.

Заключение

В учебном пособии авторами предпринята попытка дать обобщенное описание основных методов и моделей анализа статистических данных, характеризующих результаты научно-исследовательской работы. Очевидно, что успех применения математических моделей в аналитической работе во многом зависит от правильного выбора метода исследования. Для обоснования необходимости разработки и использования моделей того или иного типа необходимо обладать соответствующими знаниями в области теории управления, системного анализа и эконометрики. Изучение состояния процесса (объекта, явления), его динамики и структурных сдвигов должно производиться с применением современных методов визуализации и последующего моделирования. Так, рассмотренные способы и методы визуализации играют важную роль не только на этапе исследования данных, но и на этапе презентации данных.

Предлагаемое пособие не претендует на полное описание инструментария моделирования социально-правовых и экономических процессов (явлений), но при этом рассматриваемые методы и модели являются эффективным практическим инструментом как в научно-исследовательской так и в информационно-аналитической работе штабных и информационных подразделений.

Рассмотренные методы и модели могут быть интересны не только адъюнктам, слушателям и курсантам образовательных организаций системы МВД России, но и сотрудникам информационно-аналитических и штабных подразделений.

Список использованной литературы

Горошко И. В. Технология проведения анкетных опросов // Вестник Университета прокуратуры Российской Федерации. № 3 (77). 2020. С. 72–77.

Елисеева И. И., Курышева С. В. Фиктивные переменные в анализе данных // Социология: 4М. 2010. № 30. С. 43–63.

Инструменты для качественной визуализации данных: искусство использования диаграмм. Копенгаген: Европейское региональное бюро ВОЗ, 2021. Лицензия: CC BY-NC-SA 3.0 IGO.

Информационные технологии в науке и образовании: учебное пособие / И. В. Горошко, Б. А. Торопов. Москва: Академия управления МВД России, 2021. 76 с.

Информационные технологии управления и организация защиты информации: учебник / В. В. Баранов и др. Москва: Академия управления МВД России, 2018. 456 с.

Кравченко Ю. А. Работа полицейского в MS Excel 2013: практическое пособие. Москва, 2016. 320 с.

Математические методы исследования социальных систем: курс лекций / И. В. Горошко, Б. А. Торопов, Ш. Х. Гонов. Москва: Академия управления МВД России, 2019. 80 с.

Новиков Д. А., Новочадов В. В. Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: ВолГМУ, 2005. 84 с.

Новиков Д. А. Статистические методы в педагогических исследованиях (типовые случаи). Москва: МЗ-Пресс, 2004. 67 с.

Носко В. П. Эконометрика для начинающих. Москва, 2005. 379 с.

Торопов Б. А., Гонов Ш. Х. Статистические методы принятия управленческих решений: сборник задач (задачник). Москва: Академия управления МВД России, 2019. 76 с.

Фадеева Л. Н. Теория вероятностей и математическая статистика: учебное пособие / Л. Н. Фадеева, А. В. Лебедев. 2-е изд., перераб. и доп. Москва: Эксмо, 2010. 496 с.

Шеффе Г. Дисперсионный анализ. Москва: Наука; Главная редакция физико-математической литературы, 1980. 512 с.

Ядов В. А. Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности. Москва: Омега-Л, 2007. 567 с.

Яу Н. Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами. Москва: Манн, Иванов и Фербер, 2013. 352 с.

ДЛЯ ЗАМЕТОК

ДЛЯ ЗАМЕТОК

Учебное издание

Шамиль Хасанович Гонов
кандидат технических наук
(Академия управления МВД России)

Игорь Владимирович Горошко
доктор технических наук, профессор
(Академия управления МВД России,
Университет прокуратуры Российской Федерации)

АКТУАЛЬНЫЕ ВОПРОСЫ АНАЛИЗА ДАННЫХ, ХАРАКТЕРИЗУЮЩИХ РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ВНУТРЕННИХ ДЕЛ

Учебное пособие

Редактор *Я. В. Артемьева*
Верстка *С. Н. Портновой*

Подписано в печать _._.2022. Формат 60 × 84 $^{1}/_{16}$.
Усл. печ. л. 7,44. Уч.-изд. л. 3,34. Тираж 65 экз. Заказ № 37у

Отделение полиграфической и оперативной печати РИО
Академии управления МВД России
125171, Москва, ул. Зои и Александра Космодемьянских, д. 8

ISBN 978-5-907530-05-8

Академия управления МВД России

Ш. Х. Гонов, И. В. Горошко

АКТУАЛЬНЫЕ ВОПРОСЫ АНАЛИЗА ДАННЫХ, ХАРАКТЕРИЗУЮЩИХ РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ВНУТРЕННИХ ДЕЛ

Учебное пособие

Москва • 2022

УДК 004.03
ББК 32.97
Г65

Одобрено редакционно-издательским советом Академии управления МВД России

Рецензенты: *И. А. Кубасов,* главный научный сотрудник НИИСТ ФКУ НПО «СТиС» МВД России, доктор технических наук, доцент; *М. Ю. Паклянченко,* доцент кафедры специальных информационных технологий учебно-научного комплекса информационных технологий Московского университета МВД России имени В. Я. Кикотя, кандидат технических наук.

Гонов Ш. Х., Горошко И. В.

Актуальные вопросы анализа данных, характеризующих результаты деятельности органов внутренних дел : учебное пособие / Ш. Х. Гонов, И. В. Горошко. – Москва : Академия управления МВД России, 2022. – 128 с.

Г65

ISBN 978-5-907530-05-8

В учебном пособии рассматривается система информационно-аналитического обеспечения научных исследований и образовательного процесса в органах внутренних дел Российской Федерации. В системном и обобщенном виде описываются методы и модели исследования социальных и экономических систем, а также механизмы управления сложными организационными системами.

Учебное пособие может использоваться в образовательном процессе по дисциплинам, преподаваемым по программе подготовки научно-педагогических кадров по направлениям подготовки 09.07.01 – Информатика и вычислительная техника, 37.07.01 – Психологические науки, 38.07.01 – Экономика, 40.07.01 – Юриспруденция, 44.07.01 – Образование и педагогические науки, а также по программам магистратуры и дополнительным профессиональным программам повышения квалификации. Материалы учебного пособия могут быть полезны практическим работникам штабных и информационных подразделений, а также соискателям ученых степеней и специалистам, интересующимся актуальными вопросами использования компьютерных технологий в научных исследованиях и образовательном процессе.

УДК 004.03
ББК 32.97

ISBN 978-5-907530-05-8

Оглавление

Введение

Для изучения объекта исследования необходимо иметь ясное представление не только о внутренних (эндогенных), но и о внешних (экзогенных) переменных, характеризующих закономерности функционирования организационной системы. Моделируя состояние объекта, развитие явления или структуру процесса, исследователь должен анализировать и прогнозировать возможные воздействия на них со стороны внешней среды. Модели такого вида можно получить с помощью современных методов статистического анализа данных, в основе которых лежат основные положения теории управления, законы теории вероятностей и математической статистики, теории систем и т. д.

Существует множество различных подходов, методов и моделей исследования организационных систем, которые были созданы для их описания и решения практических задач. И сами модели, и информационные технологии, созданные для их реализации, представляют собой очень разнообразное сочетание различных математических подходов и методов. Это обстоятельство затрудняет систематизацию задач исследования организационных систем и отрицательно сказывается на ресурсном обеспечении.

Одна из задач пособия – систематизировать существующие математические подходы и методы описания организационных систем, а также дать эффективный инструментарий молодым исследователям для решения задач моделирования в первую очередь социально-правовых систем.

Актуальность данного учебного пособия обусловлена необходимостью обобщения и систематизации знаний и подходов к современным аспектам построения математических моделей описания организационных систем на основе компьютерных технологий.

Пособие призвано помочь реализации одной из главных задач комплекса мероприятий по совершенствованию системы подготовки кадров для органов внутренних дел Российской Федерации – повышению качества и усилению практической направленности системы подготовки высококвалифицированных научных и научно-педагогических кадров.

Глава 1. Методы визуализации данных

В настоящей главе рассматриваются основные методы отображения статистических данных при решении научно-исследовательских задач, а также способы визуализации данных с помощью гистограмм, графиков и других типов диаграмм и технологии их создания[1].

1.1. Виды данных и их визуализация

Визуализация данных – это набор методов, которые позволяют использовать визуальное представление для изучения, анализа и коммуникации количественных данных[2]. Это помогает изучать тенденции и закономерности в имеющемся у исследователя наборе данных. Конечная цель визуализации данных – способствовать принятию более эффективных решений и мер.

Чем больше становятся объемы доступных нам данных, тем важнее иметь возможность интерпретировать постоянно увеличивающиеся массивы информации, и визуализация данных позволяет эффективно решить эту задачу. Сказанное актуально не только для специалистов по обработке данных и аналитиков, владение методами визуализации данных необходимо в научной и образовательной сферах.

Данные можно визуализировать на различных этапах анализа и коммуникации. Во-первых, на ***этапе исследования данных*** производить анализ, визуализируя их с помощью инструментов статистического анализа и электронных таблиц, гораздо проще. С помощью инструментов визуализации можно выявлять различные связи, изучать распределения и сравнивать данные. На этапе исследования данных не столь важно, какой тип диаграммы выбрать, какие пояснительные надписи разместить и как оформить иллюстративный материал, главное, чтобы визуализация позволяла аналитику получить новую информацию.

После того, как удалось достаточно глубоко разобраться в наблюдаемых тенденциях и закономерностях, начинается ***этап презентации данных***. Цель презентации данных (иногда называемой представлением данных) – ознакомить целевую аудиторию с конечными результатами исследования. Они могут быть представ-

---

[1] Информационные технологии в науке и образовании: учебное пособие / И. В. Горошко, Б. А. Торопов. Москва: Академия управления МВД России, 2021. 76 с.

[2] Инструменты для качественной визуализации данных: искусство использования диаграмм. Копенгаген: Европейское региональное бюро ВОЗ, 2021. Лицензия: CC BY-NC-SA 3.0 IGO.

лены в отчете, справке, презентации или в диссертационном исследовании. На этом этапе большое значение приобретают визуальное оформление материала, тип выбранной диаграммы, пояснительные подписи и т. д. Переработав данные в информацию, мы должны помочь широкой аудитории усвоить сделанные нами выводы.

Силу визуализации можно проиллюстрировать на примере, который известен как квартет Энскомба. Это четыре набора данных, которые почти идентичны по описательным характеристикам, но имеют разное распределение и при графическом представлении дают совершенно разную картину. Каждый набор данных состоит из 11 точек (*x*, *y*). Эти наборы данных были разработаны в 1973 г. специалистом по статистике Фрэнсисом Энскомбом (*англ. Francis John Anscombe*), чтобы продемонстрировать, как важно перед анализом данных представить их в виде диаграммы (графика).

Таблица 1.1. Квартет Энскомба

	Набор I		**Набор II**		**Набор III**		**Набор IV**	
	x1	*y1*	*x2*	*y2*	*x3*	*y3*	*x4*	*y4*
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
Среднее	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
Дисперсия	10,00	3,75	10,00	3,75	10,00	3,75	10,00	3,75
Корреляция	0,82		0,82		0,82		0,82	
Уравнение регрессии	*y=0,5*x+3*		*y=0,5*x+4*		*y=0,5*x+5*		*y=0,5*x+6*	
Коэффициент детерминации	0,67		0,67		0,67		0,67	

В таблице 1.1 представлены наборы данных, у первых трех значения x одинаковы. Заметим, что приведенные в конце таблицы некоторые описательные статистические характеристики одинаковы, поэтому можно предположить, что эти наборы данных идентичны, но если представить их в виде диаграмм, то различия становятся очевидны (рис. 1.1–1.4).

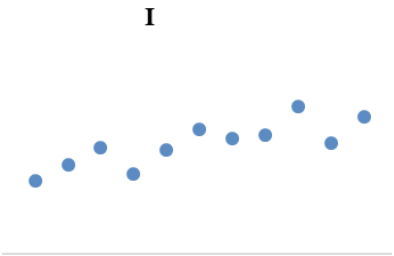


Рис. 1.1. Квартет Энскомба (I)

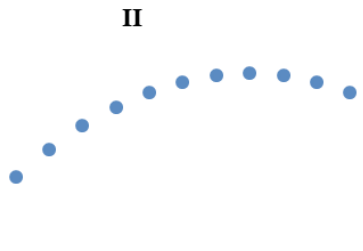


Рис. 1.2. Квартет Энскомба (II)

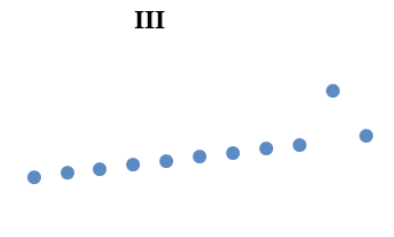


Рис. 1.3. Квартет Энскомба (III)

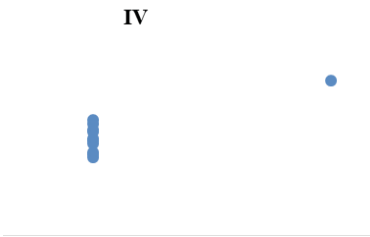


Рис. 1.4. Квартет Энскомба (IV)

Таким образом, визуализация данных может дать информацию, которую не всегда дают табличные данные и описательные статистики. Без визуального представления информации зачастую трудно понять истинное значение полученных результатов.

Способы визуализации данных

Рассмотрим основные способы представления данных. Визуализация данных имеет множество применений, и каждый ее вид можно использовать по-разному[1]. Самым распространенным способом применения визуализации данных являются временные изменения, или изменения в динамике. Этот способ наиболее распро-

[1] Более подробно различные типы диаграмм рассматриваются в следующем параграфе.

странен, поскольку в большинстве наборов данных присутствует временной фактор *t*. Для этого типа данных лучше всего подходят линейные графики. Однако нужно иметь в виду, что, если необходимо отобразить много линий тренда сразу, линейные графики часто оказываются перегружены и образуют диаграммы типа «спагетти» (рис. 1.5).

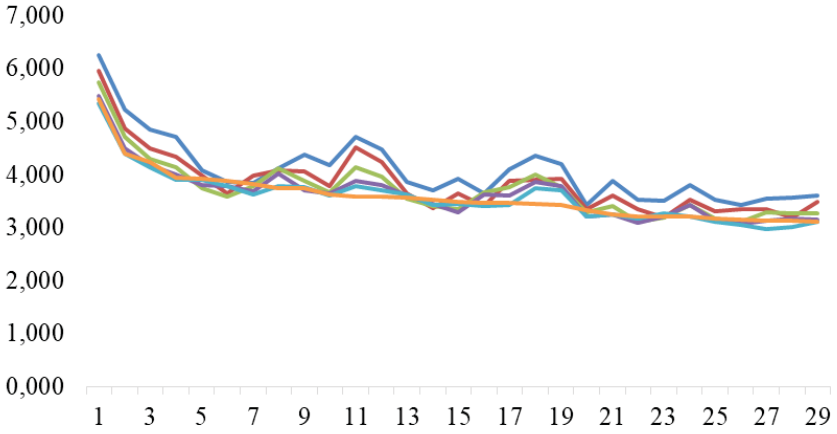


Рис. 1.5. «Спагетти» диаграмма

В качестве дополнительной графической переменной можно использовать цвет. Визуализация данных в данном случае столь же проста, как цветовое кодирование; такой прием может помочь при отображении больших наборов данных с большим количеством линий. Цветовая маркировка просто добавляет дополнительную переменную к координатам по осям *X* и *Y*.

Отображение частот – еще одно частое применение визуализации данных. Оно возможно, когда данные разделены на классы, которые могут быть порядковыми либо номинальными. Для этого типа визуализации лучше всего подходят столбчатые диаграммы, линейчатые диаграммы и гистограммы.

Выявление корреляций (определение отношений) – чрезвычайно ценное применение визуализации данных. Без визуализации трудно определить отношения между двумя переменными, однако знать о взаимосвязях в данных крайне важно. Для этого очень полезны точечные диаграммы и пузырьковые диаграммы. Визуализация наборов данных Энскомба (рис. 1.1–1.4) является примером точечной диаграммы. Это замечательный пример, пока-

зывающий ценность визуализации при анализе данных. Становится очевидно, что для понимания корреляций и выбросов одной описательной статистики недостаточно.

При изучении **пространственных закономерностей** (наличии пространственного компонента) хорошим способом визуализации распределения по территориальному признаку (например, по району, городу, региону или стране) являются картографические решения. Этот способ может быть удобен, например, для визуализации различий в уровне преступности между регионами страны.

Для определения сложных показателей, таких как значения с доверительными интервалами, необходимо учитывать множество различных переменных, что делает адекватный просмотр данных с помощью простой электронной таблицы почти невозможным. Для визуализации значений со степенью доверительной вероятности удобно использовать диаграмму **диапазонов с областями**.

Иногда может потребоваться сочетать разные способы визуализации данных. Если, например, необходимо визуализировать изменение частотного значения с течением времени, можно поместить рядом две линейчатых диаграммы, однако наилучшим вариантом будет **диаграмма наклона**. Сравнение частот с другой популяцией или целевым ориентиром может быть выполнено с помощью диаграммы-шкалы.

Любой набор данных имеет ряд характеристик. Рассмотрим две характеристики, которые важно учитывать при выборе типа диаграммы:

1) содержит ли набор данных индивидуальные (отдельные) единицы данных или агрегированные (сгруппированные);

2) какая используется шкала измерения.

Первой характеристикой набора данных является то, содержит ли он индивидуальные или сгруппированные данные. Так, индивидуальные данные могут быть визуализированы в виде точечной диаграммы или представлены в обобщенном виде как частоты на гистограмме.

Основная характеристика сгруппированных данных заключается в том, что отдельные наблюдения объединяются в группы. Примеры:

- количество совершенных преступлений за месяц;
- количество новых случаев COVID-19 за сутки;
- количество погибших в ДТП.

Сгруппированные данные лучше всего визуализировать с помощью линейных графиков, столбчатых диаграмм или линейчатых диаграмм.

Вторая характеристика набора данных – это шкала измерения. Выбранная шкала определяет тип данных и операции, которые можно осуществлять с этими данными. Состояние объекта исследования оценивается по критериям, а оценки измеряются по определенной шкале. В 1940-х гг. американский психолог Стэнли Смит Стивенс (*англ. Stanley Smith Stevens*) выделил четыре шкалы измерения: номинальную (наименований), порядковую (ранговую), интервальную и шкалу отношений. Эти шкалы активно используются в научных и прикладных исследованиях для описания характеристик переменных. Определение шкалы измерения переменных очень важно для выбора корректного метода исследования и правильного способа визуализации данных.

Номинальная шкала (шкала наименований) характеризует переменную по принадлежности к определенной категории. Иными словами, это качественная шкала, компоненты которой представляют собой связанные между собой отдельные элементы, которые не предполагают какого-либо строгого порядка. Номинальные переменные можно кодировать числами, но порядок присвоения этих чисел будет произвольным, и любые вычисления с ними – некорректными.

Примерами номинальных переменных могут быть пол, цвет автомобиля, список субъектов Российской Федерации и др. Для этой шкалы измерения наилучшим образом подойдет линейчатая диаграмма. Поскольку качественные компоненты не имеют собственного порядка, их можно произвольно переставлять, чтобы выявлять закономерности в данных.

Порядковая (ранговая) шкала – это шкала, в которой важен порядок следования уровней, но не разница между значениями. Ее компоненты представляют собой элементы, которым свойственна некая естественная последовательность, например, «холодно – тепло – горячо» или «белый – серый – черный». Примерами порядковых переменных могут быть уровень образования (начальное, среднее, высшее), уровень дохода (высокий, средний, низкий), удовлетворенность качеством оказанной государственной (муниципальной) услуги (удовлетворительно, неудовлетворительно).

Для этой шкалы измерения лучше всего подходит столбчатая диаграмма.

Интервальная шкала – это шкала разностей, в которой уровни упорядочены, а интервалы между ними равны. Ее компоненты представляют собой элементы с постоянным числовым соотношени-

ем между собой. Примером интервальной шкалы может быть измерение времени (последовательность секунд, минут), температуры (по Фаренгейту, по Цельсию).

Переменная на ***шкале отношений*** (абсолютной шкале) имеет все свойства интервальной, но для нее также четко определен абсолютный ноль. Когда переменная равна нулю, означаемая ею сущность отсутствует. Примеры переменных, для которых используется шкала отношений: продолжительность, вес, длина, стоимость (цена).

При работе с переменными на шкале отношений, в отличие от интервальных переменных, можно получить содержательную интерпретацию, оценив соотношение двух значений. Как для интервальных переменных, так и для переменных на шкале отношений лучше всего использовать линейные графики.

Иногда интервальные данные могут отображаться при помощи довольно специфических инструментов технического анализа. Ниже приведена схема биржевой диаграммы (рис. 1.6), которая называется «Японские свечи» (*англ. candlestick chart / Japanese Candlesticks*), применяемой в первую очередь для отображения изменений биржевых котировок акций, цен на валюту и т. д. Кроме этого, на рисунке отображена схема диаграммы ящик с усами (*англ. box and whiskers plot*) с отображением квартилей (Q0-Q4).

Рис. 1.6. Японские свечи и ящик с усами

Для построения ***биржевой диаграммы*** (блочной, ящичной) (*от англ. box chart*) необходимо создать таблицу, в которой последовательно расположить следующие данные: цена открытия, максимальная цена, минимальная цена и цена закрытия. В качестве подписей обычно используются даты или наименования индикаторов (рис. 1.7).

Рис. 1.7. Биржевая диаграмма[1]

«Японские свечи» сочетают в себе свойства линейного графика и интервальной диаграммы и активно используются как эффективный инструмент теханализа, например, на основе чисел Фибоначчи (веер, дуги, зоны и др.).

Кроме этого, в визуализации данных активно применяются ***комбинированные диаграммы*** (*англ. combo chart*), сочетающие в себе разные типы. Ниже представлено сочетание линейчатой и столбчатой диаграмм (рис. 1.8).

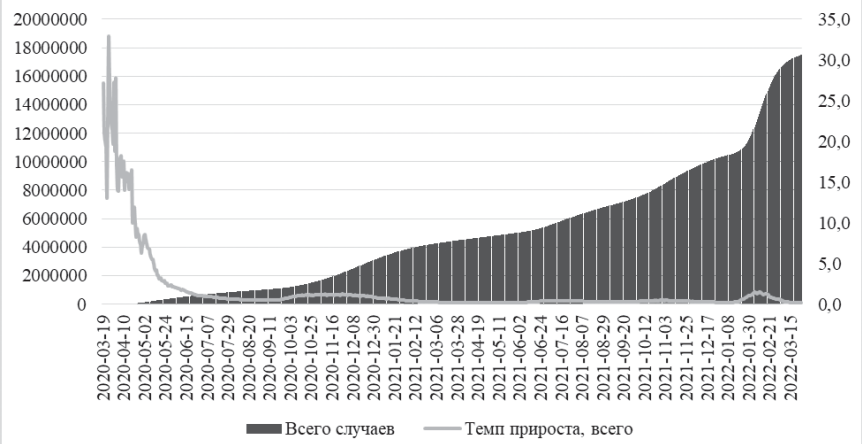


Рис. 1.8. Комбинированные типы диаграмм

[1] Данные для построения диаграммы взяты с портала «ProFinance» (https://www.profinance.ru).

На данном графике представлен временной ряд заболеваемости коронавирусом COVID-19 в Российской Федерации по состоянию на 28 марта 2022 г. Для сравнения на графике также отражен темп прироста[1].

Еще один инструмент визуализации данных – ***поверхностные диаграммы*** (*англ. surface chart*), которые могут использоваться для отображения трендов в значениях по двум измерениям (рис. 1.9).

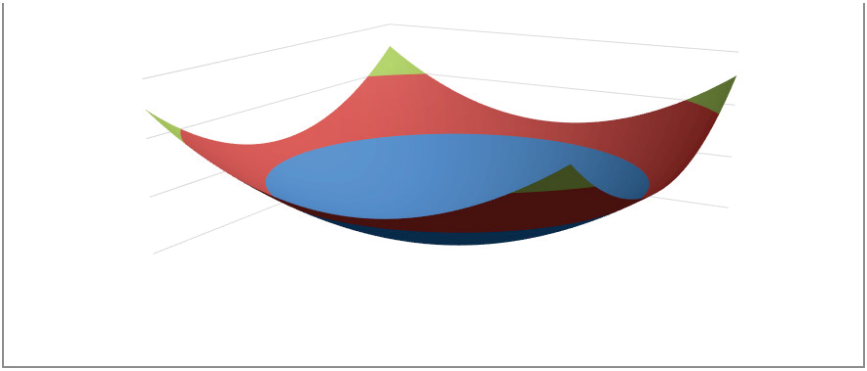


Рис. 1.9. График функции $z=x^2+y^2$

В исследовании экономических и социальных систем такие диаграммы могут использоваться для отображения построенных моделей, например, производственной функции Кобба – Дугласа. Методику построения регрессионной модели данного типа и ее визуализации рассмотрим в главе 3 настоящего учебного пособия.

Средством визуализации данных о численности населения может выступать возрастно-половая пирамида (*англ. population pyramid*). На следующем рисунке представлена ***возрастно-половая пирамида***, построенная на основе официальных данных Федеральной службы государственной статистики (Росстата) о численности населения Российской Федерации по полу и возрасту на 1 января 2021 г.[2] (рис. 1.10).

[1] Данные с портала «Our World in Data» (https://ourworldindata.org/covid-vaccinations?country=OWID_WRL).

[2] Данные с официального сайта Росстата (https://rosstat.gov.ru/compendium/document/13284).

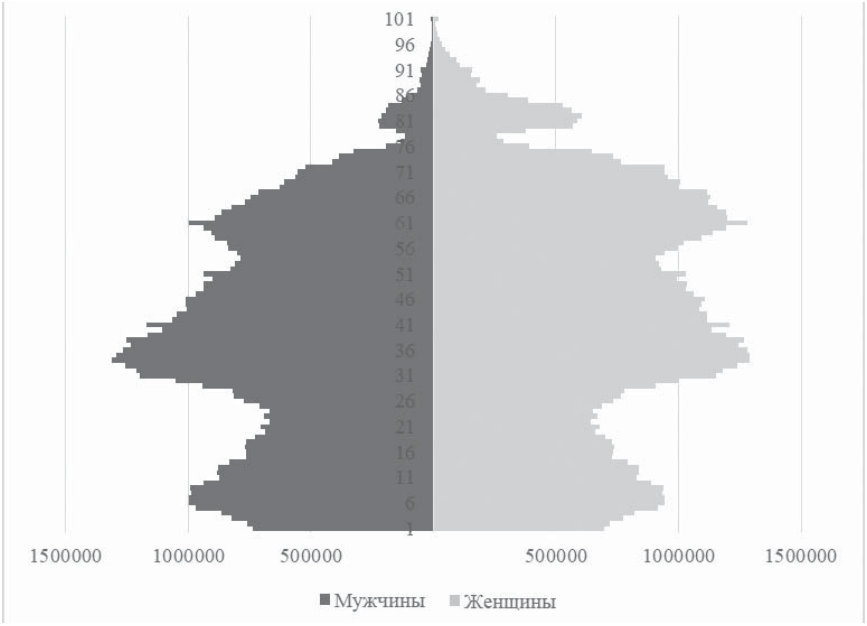


Рис. 1.10. Возрастно-половая пирамида по РФ за 2021 г.

Данный тип диаграммы хорошо иллюстрирует социально-экономические, социально-демографические и политические сдвиги в общественной жизни. На пирамиде хорошо заметны демографические волны снижения числа родившихся в середине 40-х, конце 70-х и 90-х гг., а также рост числа родившихся в конце 80-х и начале 60-х гг. Кроме этого, в верхней части пирамиды видно значительное превышение численности женщин над мужчинами, особенно старших возрастов.

1.2. Основные типы диаграмм

Линейный график *(англ. line chart)* используется для отображения интервальной шкалы или шкалы отношений по оси абсцисс *X* и количественного показателя по оси ординат *Y*. Часто по оси *X* расположена шкала времени, но могут использоваться и другие непрерывные шкалы. Линейный график хорошо подходит для иллюстрации развития процесса (явления) в динамике, например, количества зарегистрированных в отчетном периоде преступлений, квалифицируемых по статье 158 Уголовного кодекса Российской Федерации (далее – УК РФ) (рис. 1.11).

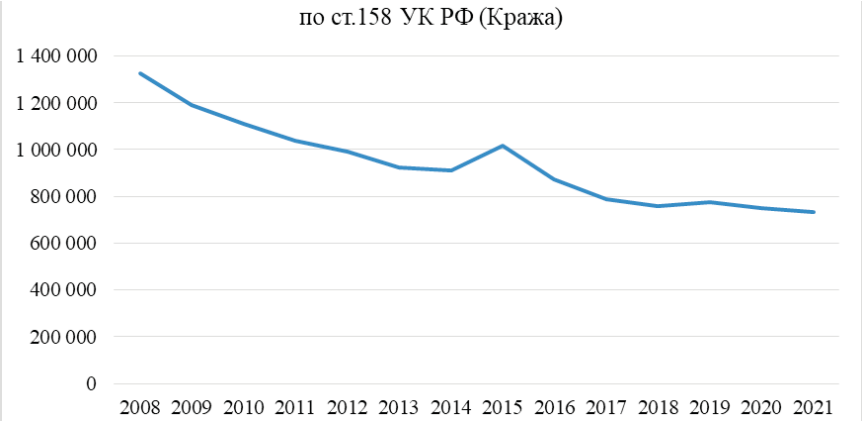


Рис. 1.11. Динамика краж, зарегистрированных в отчетном периоде

Диаграмма с областями (*англ. stacked area chart*) – это вариант линейного графика, где область под линией закрашена, чтобы подчеркнуть ее значимость. Если на графике отображено более одной переменной, диаграмму с областями можно использовать как линейный график с накоплением (рис. 1.12). Значения отдельных категорий, показанных на диаграмме, складываются в общую сумму. На диаграммах этого типа нет необходимости отображать строку «Всего количество преступлений», так как отдельные показатели сами складываются в итоговое значение.

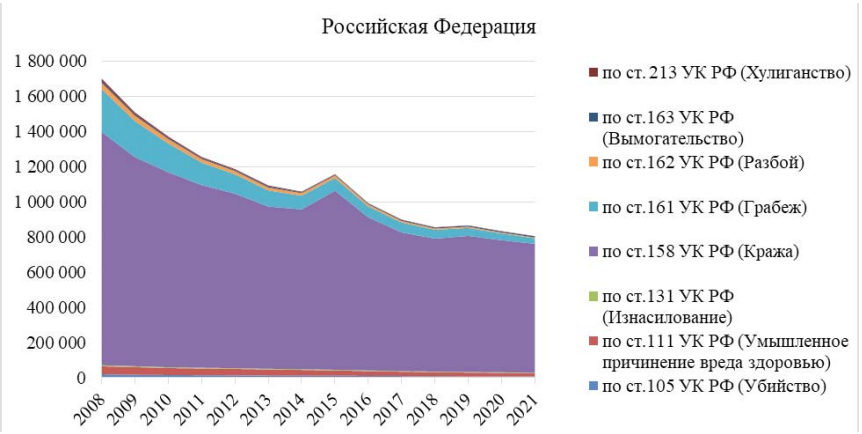


Рис. 1.12. Всего преступлений, зарегистрированных в отчетном периоде

Диаграмма диапазонов с областями – это диаграмма с областями, где область определяется двумя значениями: верхним и нижним. Диа-

грамма диапазонов с областями в основном используется для отображения предполагаемого диапазона конкретного показателя. В следующем примере в качестве диапазона используется стандартное отклонение (рис. 1.13).

Рис. 1.13. Всего зарегистрировано преступлений по России

Использование ***столбчатой диаграммы*** (*англ. column chart)* – лучший способ отобразить распределение значений порядковых переменных. На рис. 1.14 показан пример распределения безработных по возрастным группам по Российской Федерации.

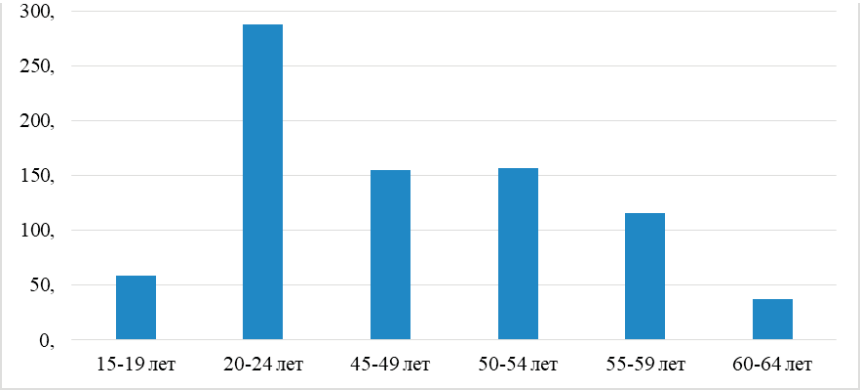


Рис. 1.14. Численность безработных по возрастным группам

Этот же вид диаграммы подходит для сравнения двух и более показателей. Например, для сравнения соотношений численности безработных по возрастным группам и полу (рис. 1.15).

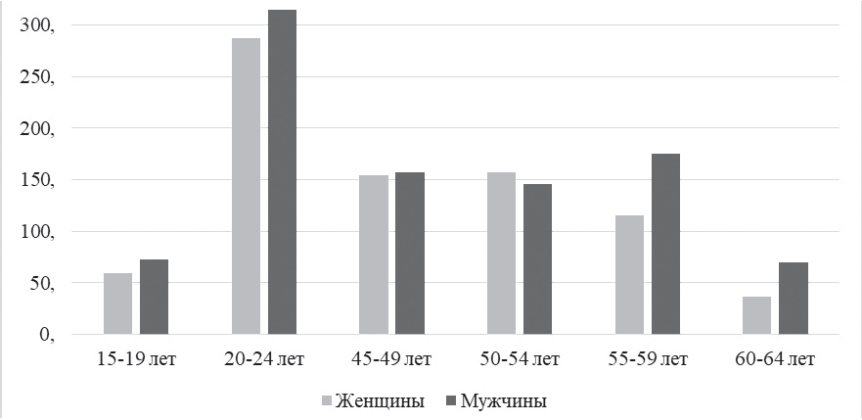


Рис. 1.15. Численность безработных по возрастным группам и полу

Линейчатая диаграмма *(англ. bar chart)* лучше всего подходит для отображения распределения значений номинальных переменных. Особенностью отображения номинальных переменных является обязательность подписи горизонтальной оси. Поскольку подписи, обозначающие номинальные переменные, могут быть длинными, линейчатая диаграмма в данном случае удобнее, чем столбчатая.

Если для отображения номинальных переменных используется столбчатая диаграмма, подписи, идущие вдоль оси X, могут стать слишком длинными для размещения по горизонтали. Поэтому для номинальных данных более оптимальным решением является линейчатая диаграмма (рис. 1.16). С технической точки зрения на линейчатой диаграмме оси X и Y расположены не так, как на столбчатой: ось Y ориентирована горизонтально, а ось X – вертикально.

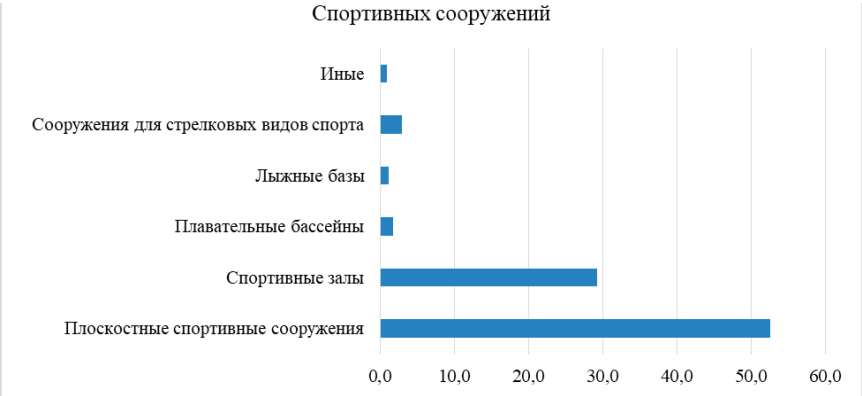


Рис. 1.16. Количество спортивных сооружений

Диаграмма-шкала с маркером – это линейчатая диаграмма с дополнительным маркером (засечкой, отметкой) на каждой линии (полосе). Такой маркер, например, может отмечать значение аналогичного показателя для другой группы или целевое значение для сравнения (рис. 1.17).

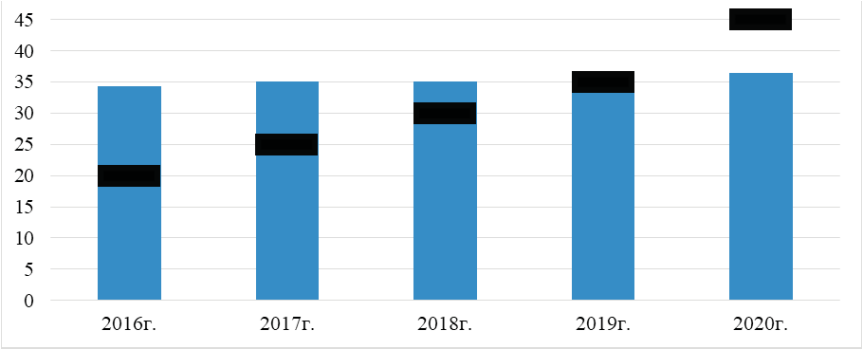


Рис. 1.17. Целевые индикаторы

Гистограмма *(англ. histogram)* выглядит как линейчатая диаграмма, но отражает распределение частот, а не тренд на порядковой шкале. По оси X гистограммы перечислены разряды (интервалы) переменной; по оси Y отсчитывается частота, поэтому каждая полоса пропорциональна частоте соответствующего разряда (рис. 1.18).

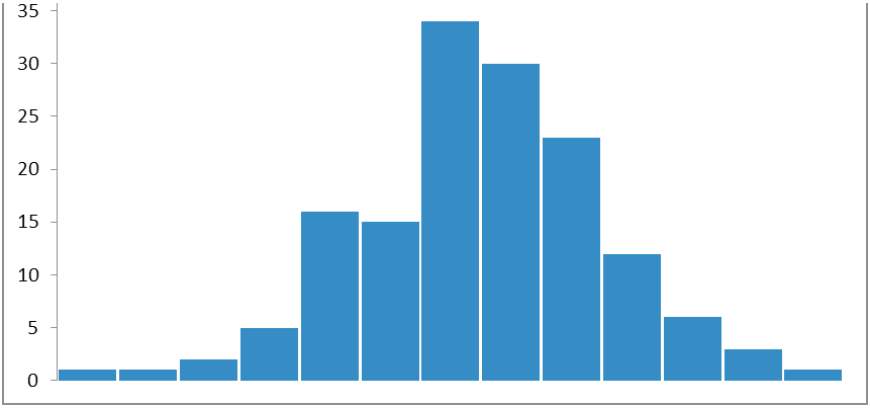


Рис. 1.18. Гистограмма

Гистограмма – это представление нормального распределения (распределения Гаусса) числовых данных. Впервые была предложена Карлом Пирсоном (*англ. Carl Pearson*) в 1895 г. Для того чтобы

построить гистограмму, необходимо сначала выделить серию разрядов (интервалов)[1], а следом подсчитать, сколько значений попадает в каждый разряд. Обычно разряды имеют одинаковый размер и располагаются по соседству, так как гистограмма показывает частоту.

Разряды могут выделяться разной ширины и в разном количестве. Самый распространенный подход к выбору количества разрядов – это метод квадратного корня. Для этого извлекается квадратный корень из числа элементов данных в выборке и округляется до следующего целого числа[2].

Точечные диаграммы *(англ. scatter chart)* (диаграммы рассеяния, разброса) чаще всего используются для поиска статистической связи (корреляций). Каждая точка на точечной диаграмме имеет координаты по оси абсцисс (x) и оси ординат (y) $A_1(x_1,y_1)$.. $A_n(x_n,y_n)$. Таким образом, если в распределении точек наблюдается определенный тренд (повышающий, понижающий и т. д.), между ними существует связь. Если точки полностью рассеяны и трендов не наблюдается, можно заключить, что переменные вообще не влияют друг на друга. Однако для того, чтобы точно сказать, есть связь или нет, используется корреляционный анализ, который мы рассмотрим в следующей главе.

Точечная диаграмма на рис. 1.19 показывает взаимосвязь между весом и ростом 5 000 человек с разбивкой по полу[3].

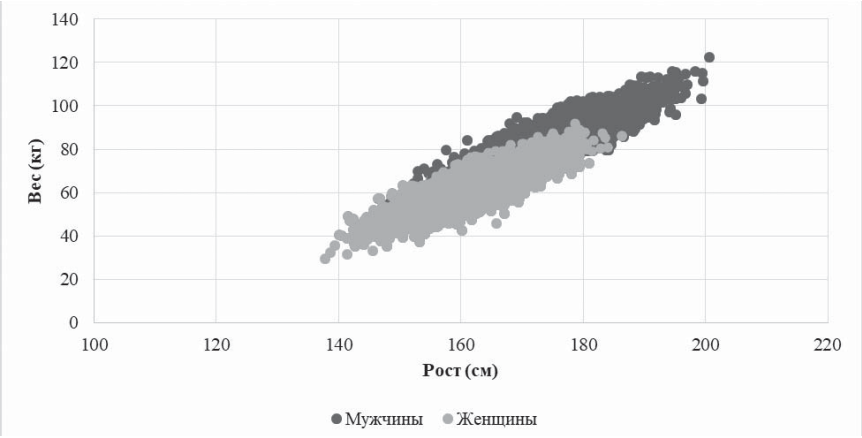


Рис. 1.19. Диаграмма разброса

[1] Иногда их называют карманами.

[2] Этот метод используется для построения гистограмм в MS Excel.

[3] Диаграмма построена на основе информации из набора данных, опубликованных на веб-платформе по адресу: https://www.kaggle.com/mustafaali96/weight-height.

Пузырьковая диаграмма *(англ. bubble chart)* – это вариант точечной диаграммы, где каждая точка изображена в виде «пузырька», площадь которого также несет определенную информацию в дополнение к положению точки по координатным осям (рис. 1.20). Трудность, связанная с пузырьковыми диаграммами, заключается в том, что пузырьки могут не помещаться на осях; поэтому не все данные подходят для этого типа визуализации. На следующей диаграмме размер пузырьков отражает данные о численности населения указанных стран[1].

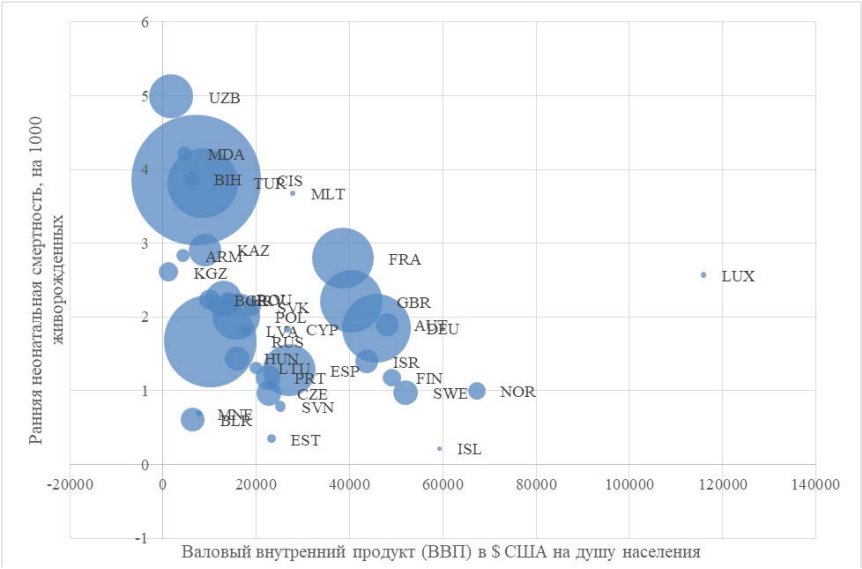


Рис. 1.20. Пузырьковая диаграмма

Круговая диаграмма *(англ. pie chart)* – это способ иллюстрации процентных долей, поскольку она показывает каждый элемент как часть целого. Ее основное преимущество заключается в том, что идея об «отношении части и целого» отражена в ней самым непосредственным образом (рис. 1.21).

[1] Диаграмма построена на основе данных с Европейского портала информации здравоохранения (https://gateway.euro.who.int/ru/hfa-explorer/).

Рис. 1.21. Круговая диаграмма

Существенным недостатком таких диаграмм является то, что размер любых других долей на круговой диаграмме оценить трудно. Причем такие диаграммы могут быть как объемными (*англ. 3D pie chart*) (рис. 1.22), так и кольцевыми (*англ. donut chart*) (рис. 1.23).

Рис. 1.22. Объемная круговая диаграмма

Однако, несмотря на очевидность содержимого круговой диаграммы, линейчатые диаграммы гораздо лучше приспособлены для сравнения величин каждой из частей. Круговые диаграммы позволяют легко

оценить величину сегмента, только когда она близка к 0 %, 25 %, 50 %, 75 % или 100 %.

Рис. 1.23. Кольцевая диаграмма

На круговой диаграмме необходимо использовать разные цвета для разных сегментов, а это означает, что для визуализации данных может потребоваться множество цветов. На линейчатой диаграмме разные элементы могут иметь одинаковые цвета, хотя при необходимости можно выделить один элемент, назначив его полосе другой цвет. Причем выбор конкретного типа диаграммы зависит от фантазии исследователя или аналитика, а также от конкретной задачи и целевой аудитории.

Для решения различных задач могут применяться разные виды диаграмм. Например, при решении задач планирования и управления проектами могут применяться ***диаграммы Ганта***[1], представляющие собой метод визуального планирования задач. Для ее построения используются две оси: по вертикали находится перечень задач, а по горизонтали время для их выполнения. Сначала создается таблица с исходными данными, с указанием перечня задач и продолжительности их выполнения, времени начала и окончания работ и др. (таблица 1.2).

Таблица 1.2. Исходные данные для построения диаграммы Ганта

№ п/п	Задача	Начало (план)	Конец (план)	Длительность (план)	Начало (факт)	Длительность (факт)	% выполнения

[1] Разработаны американским инженером Генри Гантом во время Первой мировой войны.

1	Задача 1	01.01.2021	02.01.2021	1			
2	Задача 2	01.01.2021	03.01.2021	2			
3	Задача 3	02.01.2021	04.01.2021	2			
4	Подзадача 3.1	02.01.2021	03.01.2021	1			
5	Задача 4	02.01.2021	04.01.2021	2			
6	Задача 5	03.01.2021	05.01.2021	2			
7	Задача 6	04.01.2021	08.01.2021	4			
8	Задача 7	05.01.2021	09.01.2021	4			
9	Задача 8	06.01.2021	08.01.2021	2			
10	Подзадача 8.1	06.01.2021	07.01.2021	1			
11	Задача 9	07.01.2021	10.01.2021	3			

Чаще всего для построения диаграмм Ганта используется специализированное программное обеспечение, такое как dotProject, ProjectLibre, GanttProject и др[1]. Но для решения простейших задач могут использоваться и табличные процессоры. На следующем рисунке представлена диаграмма Ганта в виде линейчатой диаграммы, построенной стандартными средствами MS Excel (рис. 1.24).

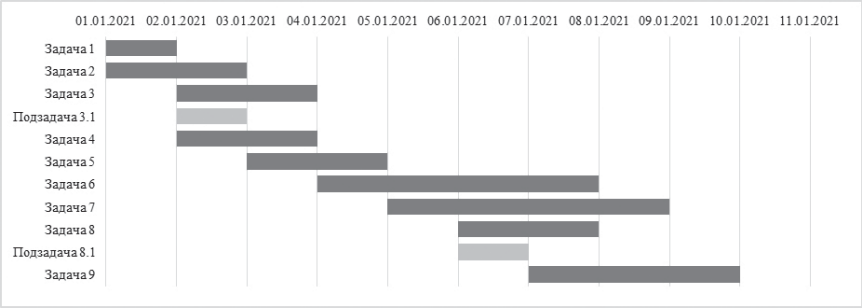


Рис. 1.24. Диаграмма Ганта

Диаграмма наклона является одной из модификаций линейного графика. На линейном графике отображаются три или более точки времени, а на диаграмме наклона – ровно две (рис. 1.25). Диаграммы наклона могут быть эффективным инструментом для визуализации

[1] В данном списке представлено программное обеспечение, распространяемое по лицензиям General Public License (GPL) или Common Public Attribution License Version (CPAL).

изменений значений для одного объекта и сравнения его с другими. Этот тип диаграммы может быть полезен для демонстрации:

– иерархических отношений набора объектов в два момента времени;

– процентной доли каждого из объектов в определенный момент времени;

– темпов изменения доли объекта в сравнении с другими объектами;

– любых отклонений в общем тренде.

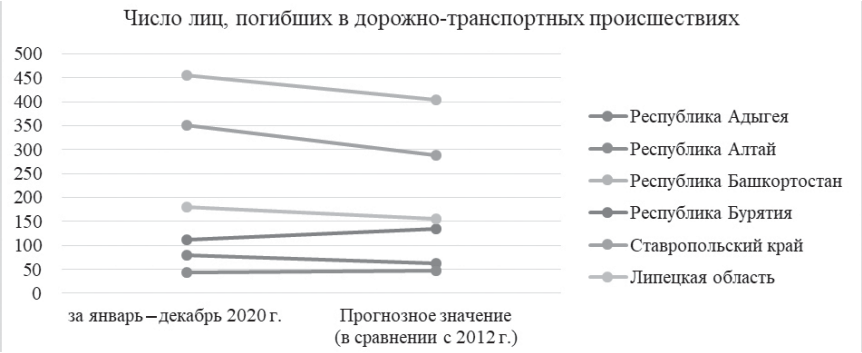


Рис. 1.25. Диаграмма наклона

Существует достаточно много специфических видов диаграмм, такие как биржевая, поверхностная, лепестковая (паутинообразная) и др. Один из вариантов – это график ***лепестковой диаграммы*** *(англ. radar chart)*, который по форме напоминает колесо, где каждый набор переменных отображается вдоль отдельной оси (рис. 1.26).

Рис. 1.26. Лепестковая диаграмма

Если диаграмма наклона может демонстрировать соотношения набора сущностей в два момента времени, то паутинообразная диаграмма намного больше. Диаграммы данного типа демонстрируют максимальную наглядность, а также хорошо подходят для иллюстрации изменчивости показателей по нескольким блокам. Кроме этого, данный тип подойдет при необходимости отобразить на одном графике отклонение переменных от некого набора стабильных значений.

Существуют некоторые рекомендации при выборе того или иного типа диаграммы. Например, процедуру выбора определенного способа визуализации иллюстрирует следующая схема (рис. 1.27)[1].

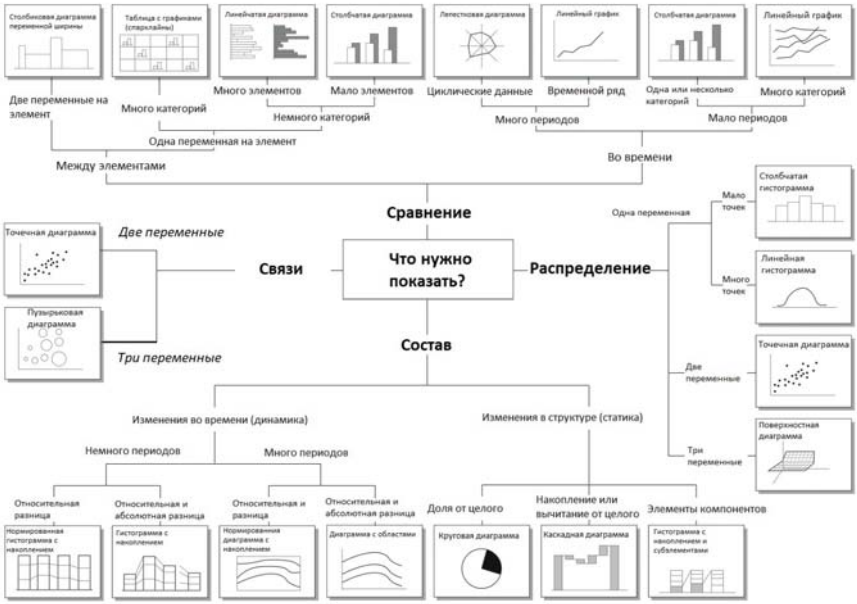


Рис. 1.27. Выбор типа диаграммы

Рассмотренные в данной главе методы визуализации данных являются общими для всех сфер деятельности[2]. Вместе с тем при решении задач научного исследования необходимо учитывать то,

[1] Опубликована на портале Extremepresentation.com (автор схемы A. Abela, перевод выполнен авторами учебного пособия).

[2] *Яу Н.* Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами / пер. с англ. Москва: Манн, Иванов и Фербер, 2013. 352 с.

что не существует типовых, шаблонных и стандартных задач в творческом процессе, к которым, безусловно, относится и научно-исследовательская деятельность.

1.3. Программное обеспечение

На современном этапе развития общества информационные технологии основаны на использовании разнообразных компьютерных программ анализа и визуализации данных. Как в научных исследованиях, так и в образовательном процессе предметом этой деятельности являются компьютерные файлы различных форматов – тексты, презентации, числовые таблицы, изображения и, на нынешнем этапе ускоренного развития образовательных технологий, файлы мультимедиа (аудио- и видеофайлы).

Наибольшее распространение получило прикладное (пользовательское) программное обеспечение (далее – ППО), отличительной особенностью которого является ориентация на выполнение определенных пользовательских специализированных задач. Программное обеспечение (далее – ПО) может быть классифицировано по типам, основным из которых является ПО офисного назначения: текстовые редакторы (MS Office Word, LibreOffice Writer и др.), электронные таблицы (MS Office Excel, LibreOffice Calc и др.), средства подготовки презентаций (MS Office PowerPoint, LibreOffice Impress и др.) и т. п. Кроме этого, к офисному ППО можно отнести некоторые графические редакторы (MS Paint, LibreOffice Draw и др.), а также пользовательские системы управления базами данных (ПСУБД), такие как MS Office Access, LibreOffice Base и др.

Отдельной категорией ППО являются браузеры, предназначенные для просмотра web-страниц, такие как Chrome, MS Explorer (Edge), Opera и др., которые относятся к проприетарному программному обеспечению. В классе браузеров к открытому (свободному) программному обеспечению относится Mozilla FireFox.

В отдельную группу можно вынести статистические пакеты, предназначенные для выявления явных и скрытых закономерностей в социально-экономических явлениях и процессах. Данный класс программ можно разделить на три группы: пользовательские (MS Excel, LibreOffice Calc. PSPP, Minitab), профессиональные (SPSS, Statistika, Gretl, IBM i2 Analyst's) и специализированные (Statistika Adv, SciDAVis, MATLAB, GNU Octave).

В последнее время наибольшее распространение получает ПО, основанное на облачных технологиях. Облачные вычисления (*англ.*

cloud computing) – модель обеспечения удобного сетевого доступа по требованию к некоторому общему фонду конфигурируемых вычислительных ресурсов (сетям передачи данных, серверам, средствам хранения данных, приложениям и сервисам – как вместе, так и по отдельности), которые могут быть оперативно предоставлены и освобождены с минимальными эксплуатационными затратами или обращениями к провайдеру. Наиболее крупными игроками в сфере публично-облачных вычислений являются такие гиганты как Amazon, Google, VMware, Microsoft. Наиболее часто встречающейся моделью облачных вычислений является технология «Программное обеспечение как услуга» (*англ. Software-as-a-Service, SaaS*), например, 1С: Предприятие, Microsoft Office 365 и др.

В зависимости от типа используемой лицензии программное обеспечение можно разделить на две категории (рис. 1.28): проприетарное (несвободное) и открытое (свободное)[1].

Рис 1.28. Лицензия на программное обеспечение

В первом случае в целях предотвращения использования, копирования или модификации программное обеспечение выпускается только в виде машинно-читаемых двоичных файлов, таким образом ограничивается доступ к исходному коду (закрытый исходный код). Вторая группа предполагает открытый доступ к исходным текстам программ. Данные типы программ предполагают использование GNU General Public License, то есть лицензии на свободное программное обеспечение, по которой автор передает разработанное им программное обеспечение в общественную собственность. Заметим,

[1] Данная классификация достаточно условна.

что использование открытого программного обеспечения позволяет повысить уровень информационной безопасности, так как в программном коде отсутствует бэкдор (*англ. back door* – тайный ход), то есть недекларированные (недокументированные) возможности (НДВ). Вместе с тем данный тип программного обеспечения также не застрахован от уязвимости «нулевого дня» (*англ. Zero-day exploit,* 0-*day attack*). Некоторые примеры разных видов программного обеспечения приведены в таблице 1.3.

Таблица 1.3. Программное обеспечение

Назначение	**Проприетарное**	**Открытое**
Графические и видеоредакторы	Microsoft Visio	Dia
	Adobe Illustrator, CorelDraw	Inkscape
	Movie Maker, Movavi	AviDeMux
Коммуникации	MS IE, Opera, Chrome	Mozilla Firefox
	MS Outlook, Becky, The Bat!	Mozilla Thunderbird
	ICQ, Trillian, MSN Messenger	Pidgin
Медиа	Windows Media Player	VLC media player
Антивирусы	Kaspersky Antivirus	ClamWin
Архиваторы	WinZip, WinRar	7-Zip
Файловый менеджер	Total Commander	MuCommander
Офисные приложения	Microsoft Office	OpenOffice, LibreOffice
Просмотр и печать документов	Adobe Acrobat	Sumatra

Безусловно, решать задачи визуализации статистических данных можно при помощи общедоступного программного обеспечения. Однако зачастую удобнее производить анализ данных и визуализацию при помощи специализированных инструментов, таких как статистические пакеты анализа. Условно данный вид ПО можно разделить на три большие группы: пользовательские, профессиональные и специализированные (таблица 1.4).

Таблица 1.4. Статистические пакеты анализа

Профессиональные	Пользовательские	Специализированные
SPSS	MS Excel	Statistika Adv
Statistika	LibreOffice Calc	SciDAVis
Gretl	PSPP	MATLAB
IBM i2 Analyst's	Minitab	GNU Octave

В свою очередь указанное ПО может быть также разделено на проприетарное и открытое. Данный перечень далеко не исчерпывающий и не включает многие хорошо известные пакеты анализа, такие как SOFA Statistics, STATA, EViews, Prognoz Platform, EViews, SalStat Statistics Package и др. Кроме этого, известно достаточно много надстроек и расширений MS Excel (StatTools, XLSTAT, StatPlus, XLR и др.), позволяющих реализовать множество статистических методов. Далее рассмотрим несколько статистических пакетов анализа, позволяющих в том числе и визуализировать данные.

GNU PSPP – программа для статистического анализа данных. Это бесплатный аналог проприетарной программы SPSS IBM и очень похожа на нее. PSPP[1] позволяет рассчитать описательную статистику, произвести различные статистические тесты, дисперсионный анализ, линейную и логистическую регрессию, кластерный анализ, анализ надежности и факторный анализ, непараметрические тесты и многое другое. PSPP предназначена для статистиков, социологов и обучающихся, которым требуется быстрый и удобный анализ статистических данных. К плюсам следует отнести ее бесплатность и свободное распространение, простоту в изучении и освоении. К недостаткам – ограниченный функционал методов и статистических тестов, отсутствие поддержки методов Data Mining[2] и AI (*англ. Artificial Intelligence* – искусственный интеллект).

GRETL (*англ. Gnu Regression, Econometrics and Time-series Library*) – это открытый, свободный и бесплатный кросс-платформенный программный пакет для эконометрического анализа. Отличительной особенностью GRETL является простой и интуитивно понятный мультиязычный интерфейс, множество методов оценивания (OLS, MLE, GMM)[3], обширный инструментарий для анализа временных рядов

[1] Данное название в виде аббревиатуры не имеет официальной расшифровки.

[2] Обобщенная группа методов «добычи» данных; глубинный анализ данных.

[3] Методы оценки неизвестных параметров моделей: OLS (Ordinary Least Squares) – метод наименьших квадратов; GMM (Generalized Method of Moments) – обобщенный метод моментов; MLE (Maximum Likelihood Estimation) – метод максимального правдоподобия.

(ARIMA, GARCH, ADL, VAR и др.)[1], поддержка моделей с ограниченной зависимой переменной (logit, probit, tobit и т. д.) и др. К недостаткам GRETL нужно отнести наличие только эконометрических методов, хотя и очень обширных, ориентацию на специалистов в области анализа данных (аналитиков) и некоторую сложность в освоении.

IBM i2 Analyst's Notebook предоставляет широкие возможности анализа данных, которые помогают преобразовать наборы разрозненных данных в высококачественную информацию. i2 Analyst's Notebook позволяет обрабатывать структурированные и неструктурированные данные, помогая аналитикам создавать единую аналитическую картину. Система обладает интуитивно понятным пользовательским интерфейсом, что сокращает время обучения пользователей, имеет широкий спектр инструментов визуального анализа, призванных помочь пользователям обнаружить ключевые связи, отношения, события, закономерности и тенденции, также включает в себя возможности анализа социальных сетей. Данные, полученные в результате анализа, можно использовать в виде интуитивно понятных визуальных диаграмм.

MS Excel – это программа, входящая в состав пакета Microsoft Office и предназначенная для работы с электронными таблицами. MS Excel – это универсальный инструмент, позволяющий обрабатывать разные форматы данных, а также хранить, организовывать и анализировать информацию. MS Excel позволяет работать с числовыми данными, текстом, создавать графики и различные варианты диаграмм. Помимо графиков и диаграмм MS Excel позволяет создавать различные типовые схемы за счет технологии SmartArt. Для решения нетривиальных (нетиповых) задач можно использовать инструменты для работы с фигурами, что значительно увеличивает возможности визуализации данных.

Начиная с версии MS Excel 2016, появилось несколько новых типов диаграмм: дерево (*англ. Treemap*), солнечные лучи (*англ. Sunburst*), ящик с усами (*англ. Box&Whisker*), каскадная (*англ. Waterfall*), обычная гистограмма (*англ. Histogramm*) и гистограмма типа Парето (*англ. Pareto*).

Практически аналогичным функционалом обладает программа ***Calc***, входящая в пакет LibreOffice. LibreOffice – это кроссплатформенный, свободно распространяемый офисный пакет с открытым исходным кодом, основанный на OpenOffice.org. Основные возможности Calc включают в себя выполнение достаточно

[1] ARIMA – autoregressive integrated moving average; GARCH – Generalized AutoRegressive Conditional Heteroscedasticity; ADL – autoregressive distributed lags; VAR – Vector AutoRegression.

сложных вычислений, позволяют организовывать, хранить и обрабатывать данные, создавать разнообразные диаграммы, а также открывать, редактировать и сохранять файлы в формате Microsoft Excel. Заметим, что при работе с электронными таблицами формата MS Excel наблюдаются некоторые проблемы с совместимостью.

Отдельно стоит обсудить ***Microsoft PowerPoint***, которая широко используется для подготовки и редактирования презентаций в различных областях человеческой деятельности, особенно в сфере науки и образования. Например, научные доклады во время защиты диссертации обычно сопровождаются презентацией. В образовательных организациях профессорско-преподавательский состав постоянно использует презентации во всех формах обучения.

Наиболее универсальным инструментом для подготовки числовых данных и их визуализации является таблица Microsoft Excel или ее аналоги с открытой лицензией LibreOffice Calc, OpenOffice Calc. Таблицы и диаграммы, созданные в этих программах, можно легко скопировать в текстовые документы в форматах doc, docx, rtf. Сами файлы, управляемые электронными таблицами, имеют множество различных расширений, но наиболее распространенными являются xls, xlsx и формат данных открытой электронной таблицы, не связанный с продуктами Microsoft – ods (*англ. Open Document Spreadsheet*).

Помимо графиков, диаграмм и типовых схем SmartArt, офисные пакеты позволяют создавать различные нетиповые схемы за счет технологии работы с иллюстрациями. Рассмотрим несколько нестандартных схем. На следующем рисунке может быть представлено схематическое описание концептуального правового или нормативного документа (рис. 1.29).

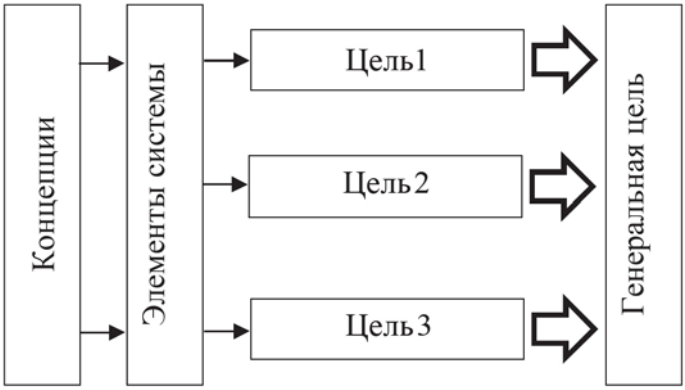


Рис. 1.29. Схематическое описание концептуального документа

Далее представлено схематическое отображение системы взаимосвязей элементов с целевыми индикаторами (рис. 1.30).

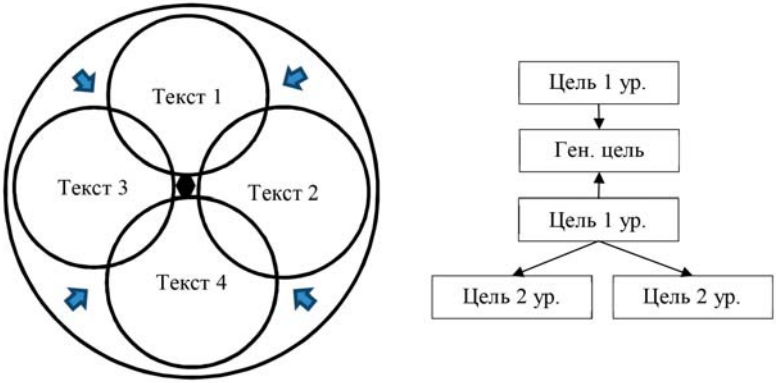


Рис. 1.30. Схематическое отображение системы взаимосвязей элементов с целевыми индикаторами

На следующем рисунке представлено схематическое описание объекта с системой целевых показателей (рис. 1.31).

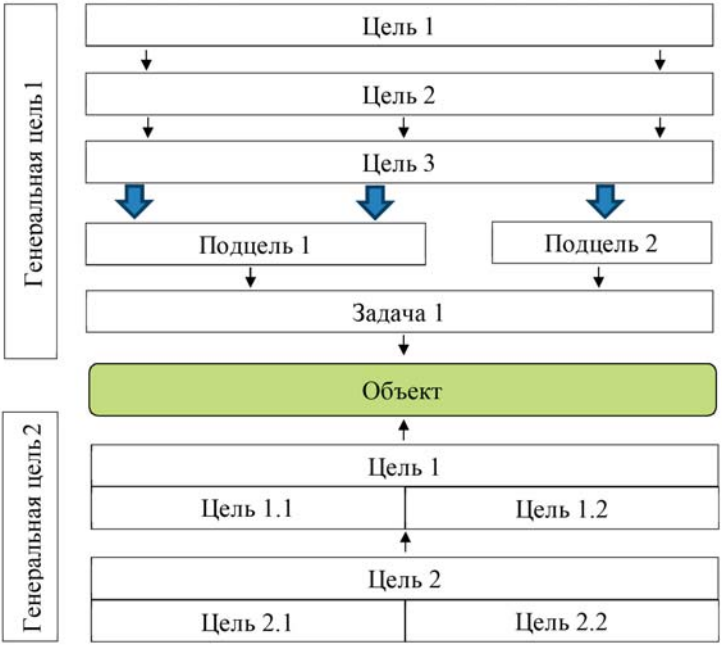


Рис. 1.31. Схематическое описание объекта с системой целевых показателей

Одной из наиболее часто используемых схем является описание иерархической системы. В большинстве случаев исследователь имеет дело с различными иерархическими системами, как типовыми, так и не типовыми. В качестве примера на рисунке представлена система управления Министерства внутренних дел Российской Федерации (рис. 1.32).

Рис. 1.32. Система управления Министерства внутренних дел Российской Федерации

В теории управления часто применяется следующая схема, иллюстрирующая систему управления (рис. 1.33).

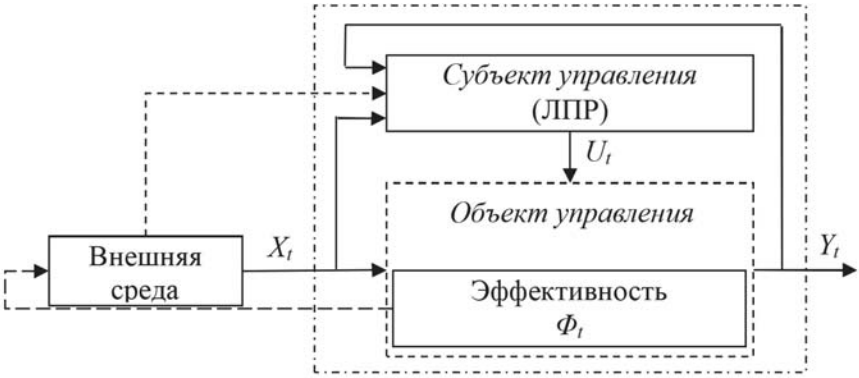


Рис. 1.33. Общая система управления

Эти схемы относятся к категории ***информационной визуализации*** и хорошо дополняют статистические данные, визуализируемые при помощи различных графиков и диаграмм (визуализация данных).

Глава 2. Основы математической статистики и анализ временных рядов

В главе рассматриваются основные понятия математической статистики, корреляционный анализ и дисперсия, а также компьютерные технологии анализа временных рядов[1].

2.1. Основные показатели математической статистики

Математическая статистика – это раздел прикладной математики, изучающий методы нахождения свойств случайных величин на основе результатов наблюдений и экспериментов. Математическая статистика основана на теории вероятностей и, в свою очередь, служит базой для разработки методов обработки и анализа статистических результатов в определенных областях человеческой деятельности.

Первая задача математической статистики – указать методы сбора и группировки статистической информации, полученной в результате наблюдений или в результате специальных экспериментов.

Вторая задача математической статистики – разработать методы анализа статистических данных в зависимости от цели исследования.

Современная математическая статистика разрабатывает способы определения количества тестов, необходимых до начала исследования (планирование тестирования) и во время исследования (последовательный анализ). Ее можно определить как науку о принятии решений в условиях неопределенности.

Таким образом, задача математической статистики заключается в создании методов сбора и обработки статистических данных для продвижения, подтверждения или опровержения обоснованных научных гипотез.

Генеральная совокупность – это совокупность всех мыслимых объектов определенного типа, над которыми производятся наблюдения с целью получения определенных значений определенной случайной величины.

Выборка (выборка совокупности) – это набор объектов, выбранных случайным образом из генеральной совокупности.

Последовательность вариаций в порядке возрастания называется ***серией вариаций***.

Дискретный статистический ряд – это ряд вариантов с соответствующими им частотами.

---

[1] Информационные технологии управления и организация защиты информации: учебник / В. В. Баранов и др. Москва: Академия управления МВД России, 2018. 456 с.

Характеристики дискретного статистического ряда: диапазон вариации, режим – вариант с наибольшей частотой, медиана – значение случайной величины в середине ряда.

Среднее арифметическое – это условное значение, которого на самом деле не существует. Есть действительно общая сумма. Следовательно, среднее арифметическое не является характеристикой наблюдения, а характеризует серию в целом.

Нормальное распределение, также называемое распределением ***Гаусса***, представляет собой распределение вероятностей, которое играет решающую роль во многих областях знаний. Социальная сфера не является исключением, где большинство ценностей обычно распределяется нормально.

Выявление закономерностей, которым подвержены случайные массовые явления, основано на исследовании статистических данных методами теории вероятностей и математической статистики[1].

Прикладное программное обеспечение MS Excel, входящее в пакет MS Office, позволяет осуществлять анализ данных при помощи двух основных подходов. Первый позволяет использовать специальный пакет анализа данных, а второй использует статистические функции, вводимые «вручную». Здесь и далее будем приводить как первый подход, так и второй.

Включение функции «Анализ данных» в MS Excel осуществляется в меню: Файл\Параметры\Надстройки, где следует нажать кнопку «Перейти» (рис. 2.1).

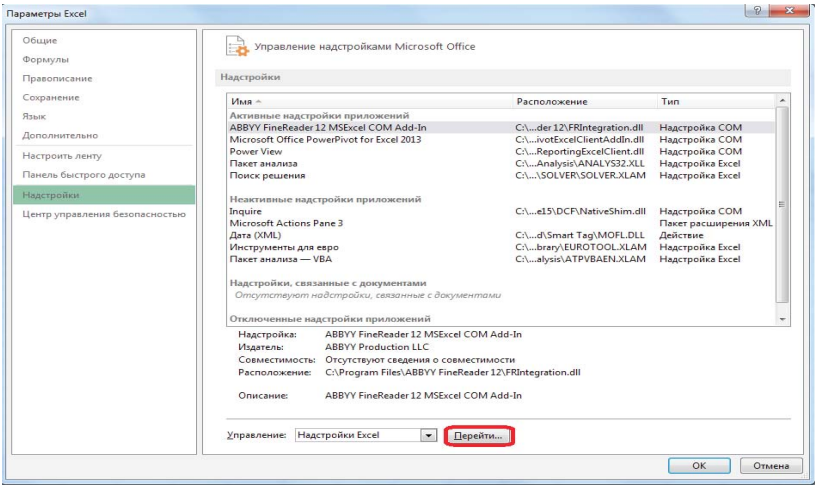


Рис. 2.1. Управление надстройками MS Excel

[1] *Носко В. П.* Эконометрика для начинающих. Москва, 2005. 379 с.

В появившемся окне установить галочку напротив пункта «Пакет анализа» и нажать кнопку «ОК» (рис. 2.2).

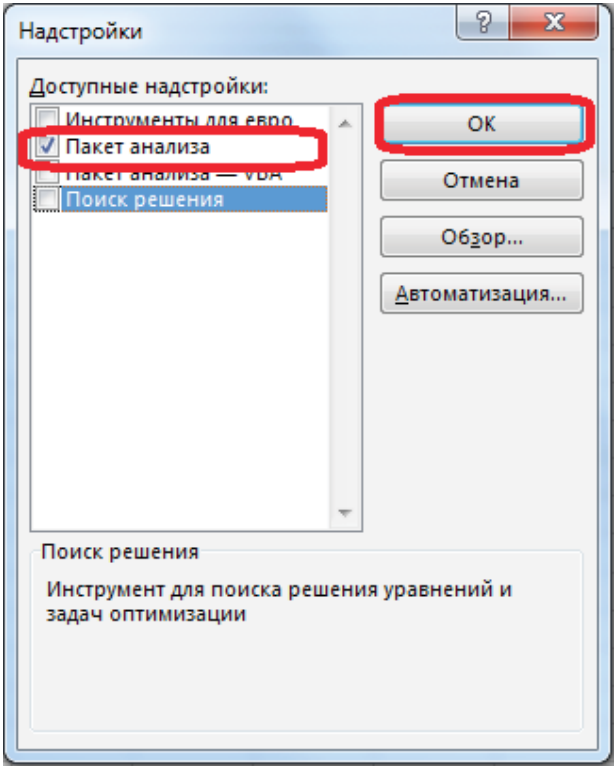


Рис. 2.2. Надстройки MS Excel

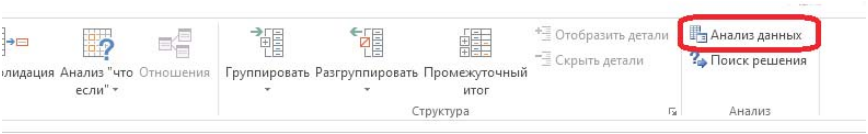


Рис. 2.3. Анализ данных MS Excel

Для расчета статистических показателей в LibreOffice Calc будут использоваться встроенные функции. Выбираем в главном меню пункт «Данные», далее «Статистика», после чего появится дополнительное меню с различными статистическими инструментами (рис. 2.4).

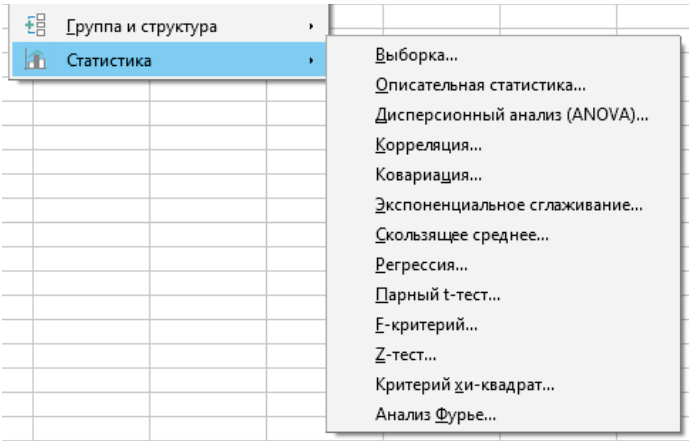


Рисунок 2.4. Анализ данных LibreOffice Calc

Заметим, что большинство функций LibreOffice и MS Excel совместимы, то есть наименование и параметры одинаковы. Это сделано для обеспечения совместимости двух популярных пакетов. Поэтому в большинстве случаев файлы формата xls и ods с автоматизированными расчетами будут без особых проблем работать в LibreOffice Calc и MS Excel.

Для расчета *описательной статистики* используем пакет анализа, встроенный в MS Excel. Выбираем в главном меню пункт «Данные», нажимаем кнопку «Анализ данных», в появившемся окне выбираем пункт «Описательная статистика» и нажимаем кнопку «ОК» (рис. 2.5).

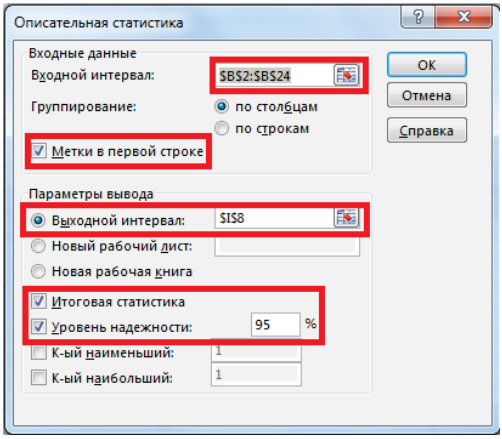


Рис. 2.5. Описательная статистика Excel

Задаем соответствующие параметры, и после нажатия кнопки «ОК» система выдаст результат следующего содержания (рис. 2.6):

Преступления	
Среднее	40188,04545
Стандартная ошибка	1223,756207
Медиана	40326,5
Мода	#Н/Д
Стандартное отклонение	5739,925398
Дисперсия выборки	32946743,57
Эксцесс	-1,496075474
Асимметричность	0,066451648
Интервал	16488
Минимум	31726
Максимум	48214
Сумма	884137
Счет	22
Уровень надежности(95,0%)	2544,94035

Рис. 2.6. Результаты описательной статистики Excel

Для расчета описательной статистики в LibreOffice Calc активируем главное меню и выберем пункт «Данные» – «Статистика» – «Описательная статистика» (рис. 2.7).

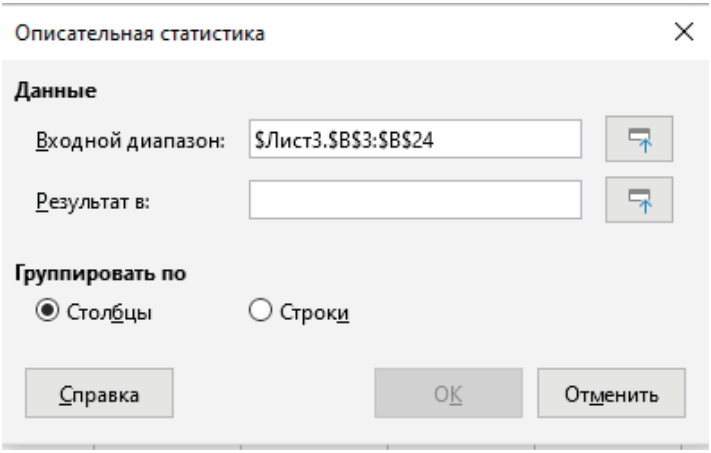


Рис. 2.7. Описательная статистика Calc

В появившемся окне задаем входной диапазон и результирующий диапазон, в нашем случае это B3:B24 и ячейка E2[1]. Задание диапазонов в Calc осуществляется нажатием кнопки  . После нажатия кнопки «ОК» получим следующий результат (рис. 2.8):

Среднее	40188,0454545455
Среднеквадратическое отклонение	1223,75620658216
Мода	#ЗНАЧ!
Медиана	40326,5
Первый квартиль	34925,5
Третий квартиль	45149,75
Дисперсия	32946743,5692641
Среднеквадратическое отклонение	5739,92539753472
Эксцесс	-1,49607547378158
Асимметрия	0,066451648385728
Диапазон	16488
Минимум	31726
Максимум	48214
Сумма	884137
Количество	22

Рис. 2.8. Результаты описательной статистики Calc

В отличие от MS Excel большинство инструментов анализа данных в LibreOffice Calc не позволяют задать подписи данных (метки).

Рассмотрим один из наиболее распространенных инструментов статистического анализа – корреляционный. Для этого обратимся к процедуре расчета коэффициента корреляции, который показывает меру тесноты взаимосвязи между двумя и более случайными величинами и рассчитывается по формуле:

$$r_{xy} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(y_i - \bar{y})^2}}; \; -1 < r_{xy} < 1.$$

Рассчитаем коэффициент корреляции «вручную» (рис. 2.9).

	A	B	C	D	E	F	G	H	I	J
1		y	x	$x_i - x_{ср}$	$y_i - y_{ср}$	$(x_i - x_{ср})*(y_i - y_{ср})$	$(x_i - x_{ср})^2$	$(y_i - y_{ср})^2$		
2	1	16 842	40,6	7,3	1648,0	11983,3	52,9	2715904,0		
3	2	15 895	33,2	-0,1	701,0	-90,1	0,0	491401,0		
4	3	14 888	29,5	-3,8	-306,0	1171,5	14,7	93636,0		
5	4	14 930	32,2	-1,1	-264,0	297,9	1,3	69696,0		
6	5	14 296	32,5	-0,8	-898,0	744,1	0,7	806404,0		
7	6	15 792	32,8	-0,5	598,0	-316,1	0,3	357604,0		
8	7	13 715	32,5	-0,8	-1479,0	1225,5	0,7	2187441,0		
9	СР	15194,0	33,3							
10	Σ	106358	233,3			15016,1	70,47428571	6722086	21765,44	0,689906

Рис. 2.9. Расчет корреляции

[1] Система автоматически расширит указанный диапазон до нужного размера.

Для автоматизированного расчета коэффициента корреляции используем пакет анализа, встроенный в MS Excel. Выбираем в главном меню пункт «Данные» и нажимаем кнопку «Анализ данных», появится окно с инструментами анализа, в котором выбираем пункт «Корреляция», нажимаем кнопку «ОК» и в появившемся окне (рис. 2.10) задаем следующие параметры:

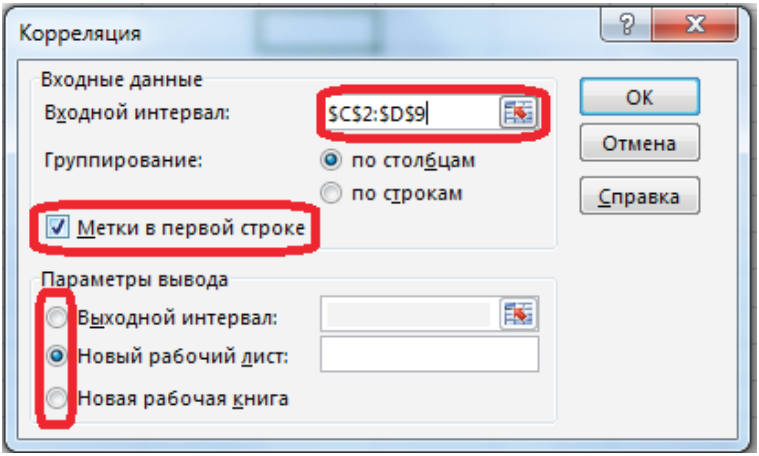


Рис. 2.10. Параметры расчета корреляции

Выбираем диапазон данных нажатием кнопки  MS Excel[1]. Указываем исходные данные, представленные ниже, и наличие меток[2].

Преступления	Безработица
16 842	40,6
15 895	33,2
14 888	29,5
14 930	32,2
14 296	32,5
15 792	32,8
13 715	32,5

Кроме этого, можно задать и параметры вывода (по умолчанию установлен вывод «Новый рабочий лист», и система направит

[1] Здесь и далее нажатие данной кнопки позволяет задать диапазон данных.
[2] Метки – это наименования полей данных.

результаты во вновь создаваемый следующий лист). После нажатия кнопки «ОК» система выдаст результат следующего содержания:

	Преступления	**Безработица**
Преступления	1	
Безработица	0,689905772	1

Полученный коэффициент корреляции равен ≈ 0,7 (0,689905772), который можно интерпретировать по шкале Чеддока.

Количественная мера тесноты связи	**Качественная мера силы связи**
0,1–0,3	Слабая
0,3–0,5	Умеренная
0,5–0,7	Заметная
0,7–0,9	Высокая
0,9–0,99	Весьма высокая

Таким образом, можно говорить о наличии положительной высокой связи переменных «Преступления» и «Безработица». При расчете коэффициента корреляции большое значение имеет знак. Так, при положительном знаке говорят о положительной (прямой) связи, а при отрицательном – об отрицательной (обратной) связи (рис. 2.11).

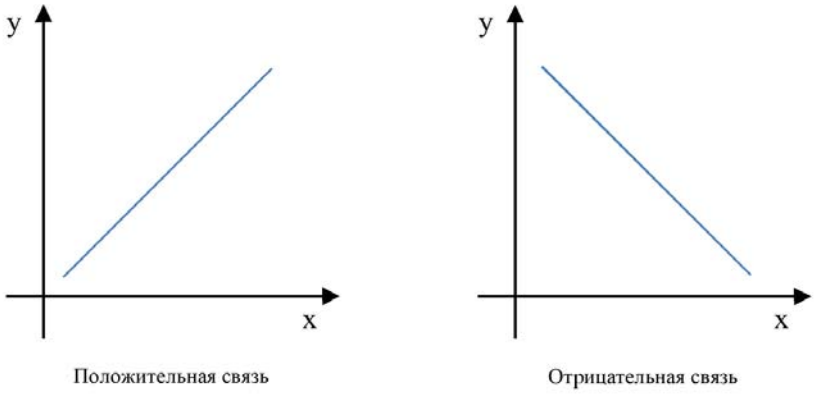


Рис. 2.11. Корреляционная связь

Аналогичные результаты получим, применяя «ручной» расчет коэффициента корреляции. Для этого активируем ячейку, в которую необходимо поместить результат, и нажимаем кнопку «Вста-

вить функцию». В появившемся окне выберем из выпадающего списка «Категория» группу функций «Статистические» (рис. 2.12)[1].

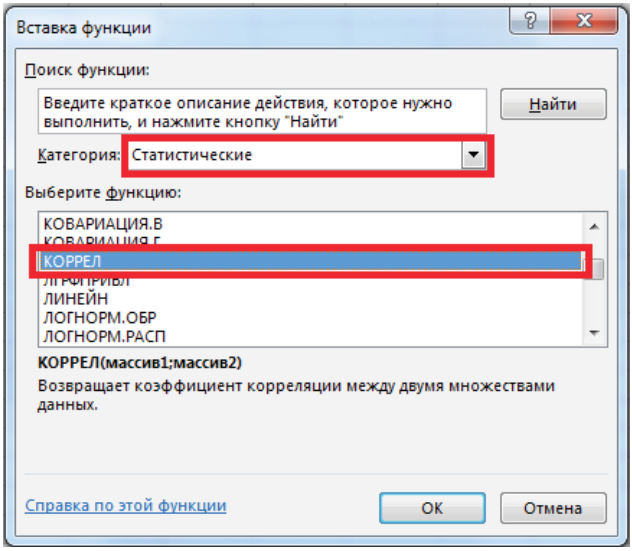


Рис. 2.12. Вставка функции «коррел»

В списке функций найдем функцию «КОРРЕЛ», и после нажатия кнопки «ОК» появится окно ввода аргументов функции следующего вида (рис. 2.13), в котором необходимо задать «Массив 1» и «Массив 2».

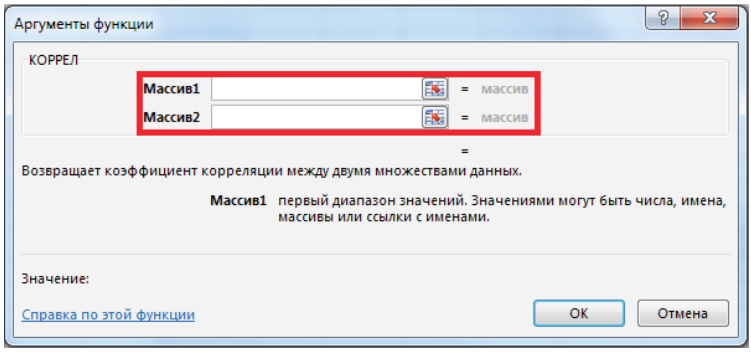


Рис. 2.13. Выбор аргументов функции

[1] По умолчанию установлена категория «10 недавно использовавшихся», и если в списке требуемой функции нет, то нужно выбрать соответствующую категорию или «полный алфавитный перечень».

После нажатия кнопки «ОК» в целевой ячейке появится результат следующего вида (рис. 2.14):

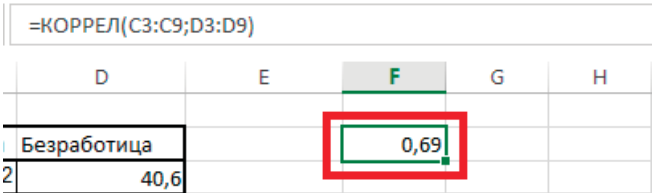


Рис. 2.14. Результат функции

Теперь рассмотрим процедуру расчета коэффициента корреляции в LibreOffice Calc. В ней, как и в MS Office, рассчитать коэффициент корреляции можно как в «ручном» режиме, так и через интерфейс, то есть диалоговое окно.

Инструменты анализа данных в LibreOffice Calc находятся также в главном меню, в котором выбираем пункт «Данные» – «Статистика» – «Корреляция» (рис. 2.15).

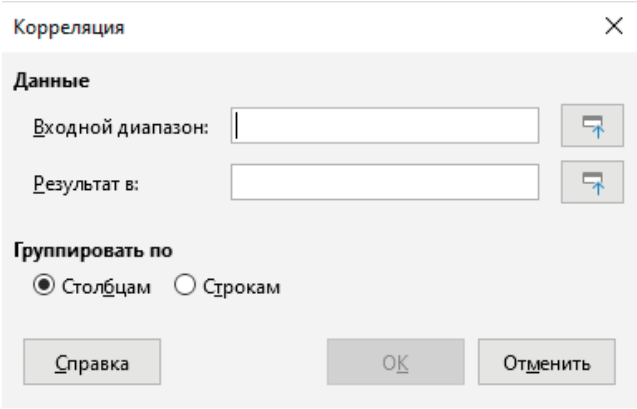


Рис. 2.15. Форма «Корреляция»

Запустится окно, в котором задаем входной диапазон и результирующий диапазон, в нашем случае это C3:D9 и ячейка F2. После нажатия кнопки «ОК» получим результат следующего вида (рис. 2.16):

Корреляции	Столбец 1	Столбец 2
Столбец 1	1	
Столбец 2	0,68990577	1

Рис 2.16. Коэффициент корреляции

В «ручном» режиме можно рассчитать коэффициент корреляции, воспользовавшись функцией «КОРРЕЛ», которая есть как в MS Excel, так и в LibreOffice Calc. Активируем нужную ячейку, куда необходимо поместить результат, и нажмем кнопку «Мастер функций». В появившемся окне находим функцию «КОРРЕЛ», выбираем ее и нажимаем кнопку «Далее». В следующем окне введем следующие параметры: «Данные 1» – это имеющиеся показатели *y*, то есть диапазон данных C3:C9; «Данные 2» – это имеющиеся показатели *x*, то есть диапазон номеров ряда D3:D9. После нажатия кнопки «ОК» в целевую ячейку будет помещен результат – коэффициент корреляции (≈ 0,69).

2.2. Дисперсионный анализ и временные ряды

Под дисперсионным анализом понимается статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния[1]. Дисперсионный анализ (*англ. Analysis of variance* – ANOVA) характеризует «разброс» значений переменных[2].

Одним из ключевых элементов дисперсионного анализа, да и всей математической статистки, является понятие статистической гипотезы. Статистическая гипотеза – это некоторое предположение относительно исследуемой генеральной совокупности[3].

Основная (нулевая) гипотеза H_0 – это гипотеза, которой мы придерживаемся, пока наблюдения не заставят признать обратное. Ей всегда сопутствует **альтернативная** (конкурирующая) гипотеза H_1.

Статистические методы не позволяют доказать гипотезу. На основе наблюдений мы можем опровергнуть гипотезу.

---

[1] *Шеффе Г.* Дисперсионный анализ. Москва: Наука: Главная редакция физико-математической литературы, 1980. 512 с.

[2] *Новиков Д. А.* Статистические методы в педагогических исследованиях (типовые случаи). Москва: МЗ-Пресс, 2004. 67 с.

[3] *Фадеева Л. Н.* Теория вероятностей и математическая статистика: учебное пособие / Л. Н. Фадеева, А. В. Лебедев. 2-е изд., перераб. и доп. Москва: Эксмо, 2010. 496 с.

Процедура проверки гипотезы состоит из нескольких этапов (рис. 2.17):

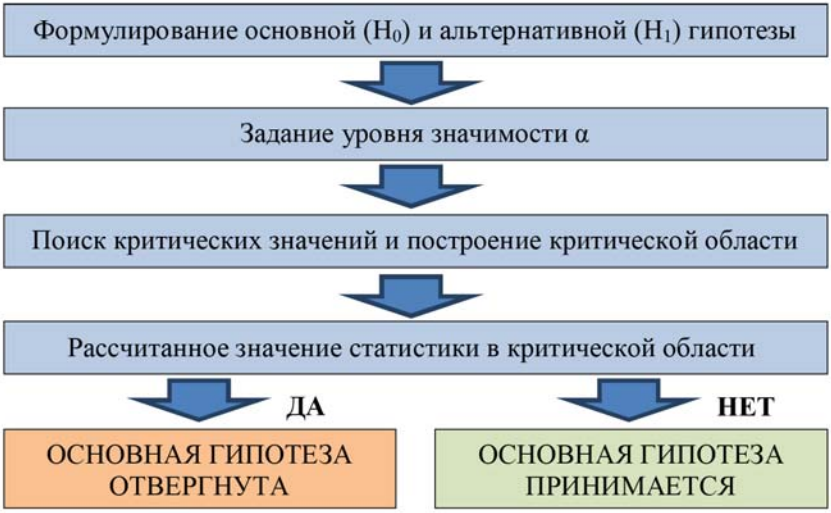


Рис. 2.17. Процедура проверки гипотез

Для определения ситуации, при которой необходимо опровергнуть гипотезу, используются понятия ошибки первого и второго родов. **Ошибка первого рода** – это ситуация, когда H_0 отвергается, но на самом деле она верна. **Ошибка второго рода** – это ситуация, когда H_0 принимается, хотя она неверна.

Чаще всего **уровень значимости или вероятность** ошибки первого рода обозначается греческой буквой α, а вероятность ошибки второго рода обычно обозначается буквой β. Уменьшение ошибки первого рода приводит к увеличению ошибки второго рода и наоборот. В научных исследованиях α чаще всего берут равным 0,1 (10 %), 0,05 (5 %) или 0,01 (1 %). В зависимости от того, какая гипотеза является основной, а какая альтернативной, ошибки первого и второго родов меняются местами.

Случайная величина, построенная по наблюдениям для проверки нулевой гипотезы, называется статистикой критерия и чаще всего обозначается Z. Для решения задачи выбирается уровень значимости и статистика критерия, на основе которой делается вывод о справедливости гипотезы. При справедливости основной гипотезы известно, с какой вероятностью какое значение принимает статистика критерия. Если эта вероятность очень маленькая, то гипотезу придется отвергнуть.

Мощностью критерия называется вероятность не совершить ошибку второго рода, то есть $1 - \beta$. А наиболее мощным критерием из всех

критериев с уровнем значимости α называется тот, который обладает наибольшей мощностью.

Критической областью называется область значений статистики критерия, при которых отвергается H_0. Критические значения – это граница критической области. Существует три вида критических областей: левосторонняя (рис. 2.18), правосторонняя (рис. 2.19) и двусторонняя (рис. 2.20).

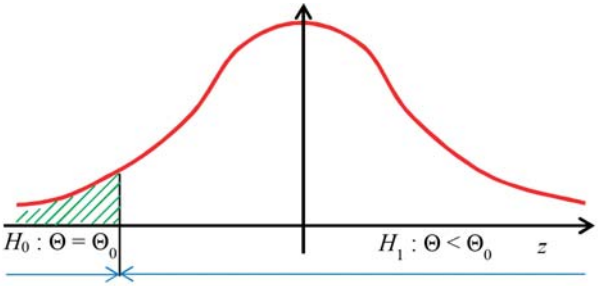


Рис. 2.18. Левосторонняя критическая область

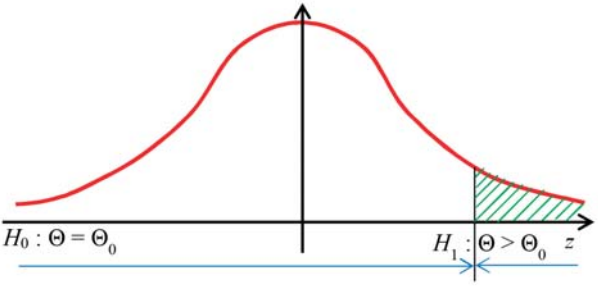


Рис. 2.19. Правосторонняя критическая область

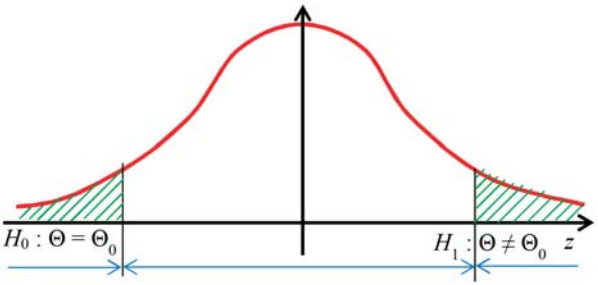


Рис. 2.20. Двусторонняя критическая область

Минимальный уровень значимости[1] (P–value) – это минимальное значение α, при котором основная гипотеза еще отвергается, обычно принимается на уровне значимости 1 %, 5 % или 10 %[2].

Далее рассмотрим элементы **анализа временных рядов**. В теории выделяют следующие элементы временного ряда[3]: регулярную компоненту, включающую в себя тренд (T), сезонность (S) и цикличность (C), а также случайную компоненту (ε). В основном применяется либо аддитивная форма[4] представления динамического ряда (T+S+C+ε), либо мультипликативная (T·S·C·ε). В сокращенном виде аддитивную модель можно записать как $y_{add} = \sum_{i=1}^{n} \beta_n x_n$, а мультипликативную как $y_{mlt} = \prod_{i=1}^{n} \beta_n x_n$.

В информационно-аналитической деятельности[5] принято выделять прогнозы со следующей периодичностью: за текущий период (ежедневно, ежемесячно), за отчетный год и прогнозы за длительный период (год и более). Ясно, что сезонность может быть только в пределах года, а циклическая компонента может присутствовать в достаточно длинных рядах. Выделить цикличность в исследовании преступности достаточно сложно, поэтому в практической деятельности данную компоненту временного ряда обычно не учитывают[6]. Таким образом, аддитивная модель временного ряда без учета цикличности будет представлена в виде (T+S+ε).

Метод экстраполяции основан на распространении тенденций развития объекта, процесса или явления в ретроспективе на будущее состояние объекта прогнозирования. Метод **статистической экстраполяции временного ряда** является одним из основных методов прогнозирования. Вместе с тем к недо-

---

[1] В прикладных эконометрических исследованиях уровень значимости обычно указывается в виде звездочки «*» рядом с коэффициентом: * - 10%-ый уровень значимости; ** - 5%-ый уровень значимости; *** - 1%-ый уровень значимости. По умолчанию применяется 5%-ый уровень значимости.

[2] *Новиков Д. А., Новочадов В. В.* Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: ВолГМУ, 2005. 84 с.

[3] Здесь и далее временной ряд и динамический рассматриваются как синонимы.

[4] Здесь и далее рассматривается только аддитивная форма представления временного ряда.

[5] Вопросы организации информационно-аналитической работы в управленческой деятельности органов внутренних дел: приказ МВД России от 26 сентября 2018 г. № 623.

[6] Заметим, что в отличие от экономических циклов (циклы Дж. Китчина, К. Жюгляра, С. Кузнеца и волны Н. Кондратьева) в исследовании социальных систем (преступности) нет единого мнения о наличии цикличности.

статкам данного метода можно отнести невозможность учесть сезонность и цикличность.

Рассмотрим методику аналитического выравнивания временного ряда (подбора аналитической функции). Основная идея метода аналитического выравнивания временного ряда заключается в применении рассмотренного ранее метода наименьших квадратов (далее – МНК) для подбора функции, наилучшим образом описывающей объект прогнозирования.

Наиболее простым и распространенным видом функции является линейная, в которой есть два коэффициента. Коэффициент сдвига, или свободный член, отражает точку пересечения с осью ординат, коэффициент при параметре x показывает наклон графика. В общем виде линейный тренд можно представить как: $y_t=\alpha+\beta x_t$ где y – значение, t – номер ряда, α и β –неизвестные параметры, оцениваемые МНК.

Для подбора аналитической функции с помощью MS Excel на первом этапе нужно построить график, для чего выделяем набор данных (рис. 2.21).

		Данные	Прогноз
1	2009	110565	
2	2010	130848	
3	2011	256007	
4	2012	271631	
5	2013	322181	
6	2014	353879	
7	2015	385152	
8	2016	468235	
9	2017	517694	
10	2018	557672	

Рис. 2.21. Выбор данных

В меню выбираем пункт «Вставка» – «Вставить график» – «График», система вставит график, и после нажатия правой клавишей мыши на графике появится контекстное меню следующего вида (рис. 2.22):

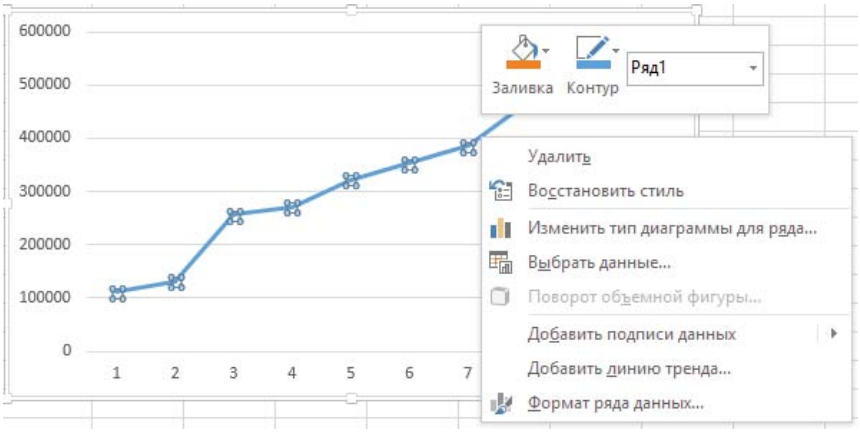


Рис. 2.22. Добавление линии тренда

Далее выбираем пункт «Добавить линию тренда», после чего появится окно справа следующего вида (рис. 2.23):

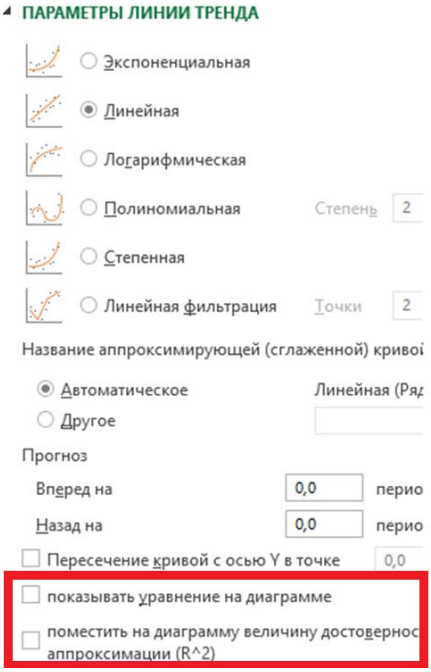


Рис. 2.23. Параметры линии тренда

В появившемся окне выберем пункт «Линейная» (она будет выбрана по умолчанию) и поставим галочки напротив пунктов «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2)»[1]. На графике (рис. 2.24) появятся следующие элементы (линия тренда, уравнение и R^2):

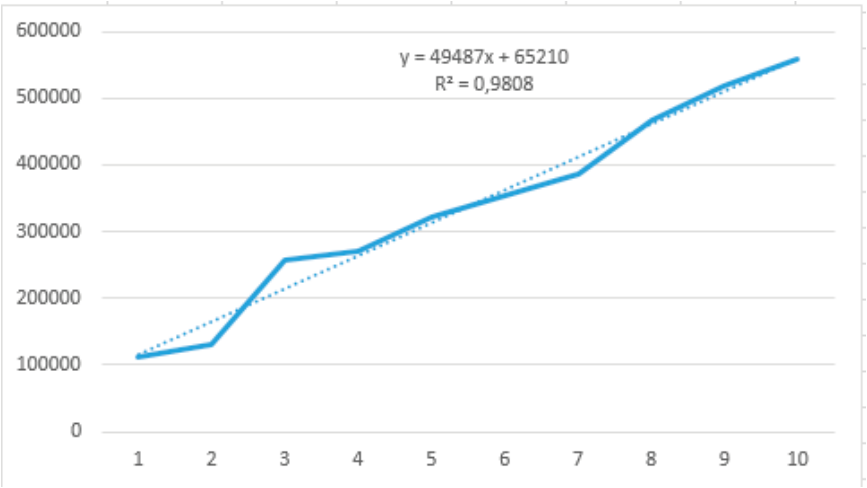


Рис. 2.24. Линия тренда

Рассмотрим вкратце полученные результаты и их значение:

1. Линия тренда – это теоретическая линия, подобранная к эмпирическим данным на основе МНК;

2. Уравнение – линейное уравнение вида $y=bx+a$, где a и b – неизвестные параметры[2], a – свободный член, который показывает точку пересечения с осью ординаты, а коэффициент b (коэффициент при x) показывает наклон графика. R^2 – коэффициент аппроксимации[3], принимает значения от 0 до 1 и показывает степень приближения теоретических (рассчитанных значений) значений $\hat{y}$ к эмпирическим значениям y. В нашим примере уравнение y=49487·x+65210, а $R^2 \approx 0{,}98$, что является довольно высоким значением[4]. X – это номер

[1] То есть коэффициент корреляции, возведенный в квадрат.

[2] Очевидно, что свободный член может располагаться как в конце, так и в начале уравнения. Заметим, что в литературе чаще всего встречается классическое расположение свободного члена в начале уравнения y=a+b*x.

[3] Иногда его также называют коэффициентом детерминации R^2.

[4] В практической деятельности такое высокое значение коэффициента встречается достаточно редко, обычно он намного ниже.

ряда, на получившемся графике он указан по оси абсцисс. Введем полученные данные в нашу таблицу (рис. 2.25).

	A	B	C	D	E
1					
2			Данные	Прогноз	
3	1	2009	110565		
4	2	2010	130848		
5	3	2011	256007		
6	4	2012	271631		
7	5	2013	322181		
8	6	2014	353879		
9	7	2015	385152		
10	8	2016	468235		
11	9	2017	517694		
12	10	2018	557672		
13	11	2019		=49487*A13+65210	
14	12	2020			

Рис. 2.25. Ввод уравнения

Зафиксируем введенное уравнение нажатием кнопки «Enter» и распространим на следующую ячейку. Таким образом, неизвестный параметр x – это номер временного ряда, для 2019 г. x=11, для 2020 г. x=12, для 2021 г. x=13 и т. д. Для прогноза на 2019 г. уравнение будет y=49487*11+65210. Построенный прогноз будет представлен на следующем графике (рис. 2.26):

		Данные	Прогноз
1	2009	110565	
2	2010	130848	
3	2011	256007	
4	2012	271631	
5	2013	322181	
6	2014	353879	
7	2015	385152	
8	2016	468235	
9	2017	517694	
10	2018	557672	560080
11	2019		609567
12	2020		659054

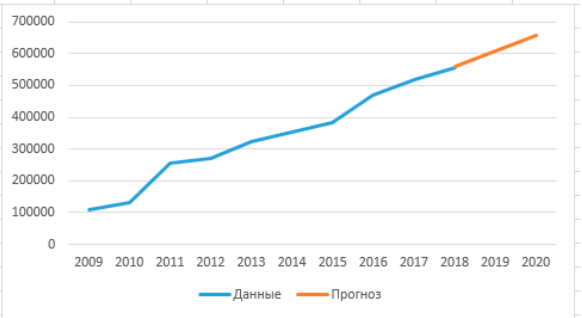


Рис. 2.26. Прогноз

Неизвестные параметры можно рассчитать и «вручную», используя инструмент «Регрессия» в пакете анализа. В появившемся окне в качестве входного интервала *Y* задаем известные значения временного ряда, а в качестве входного интервала X – номера временного ряда (рис. 2.27).

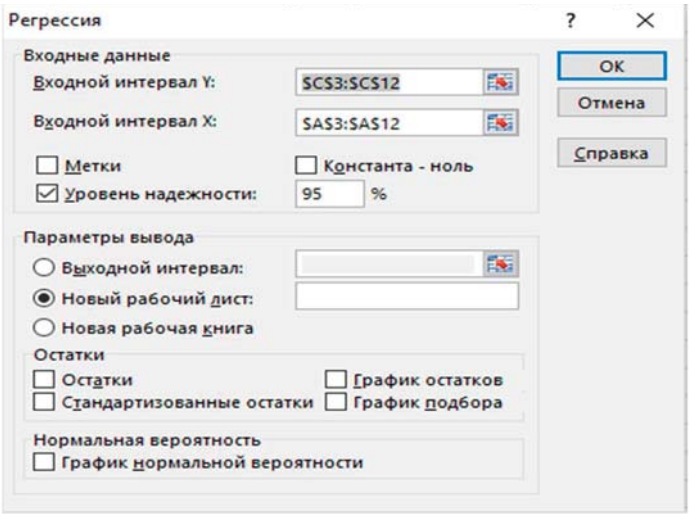


Рис. 2.27. Регрессия

После нажатия кнопки «OK» на отдельном листе появятся следующие результаты (рис. 2.28):

1	ВЫВОД ИТОГОВ								
2									
3	*ессионная статистика*								
4	Множеств	0,990344							
5	R-квадрат	0,980781							
6	Нормиров	0,978073							
7	Стандарти	22246,03							
8	Наблюдени	10							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*ачимость F*			
12	Регрессия	1	2,02E+11	2,02E+11	408,2479	3,76E-08			
13	Остаток	8	3,96E+09	4,95E+08					
14	Итого	9	2,06E+11						
15									
16	*Ко*	*ффициен*	*ртная о*	*татисти*	*²-Значени*	*ижние 95%*	*рхние 95%*	*жние 95,0*	*рхние 95,0%*
17	Y-пересе	65210,2	15196,93	4,291012	0,002648	30166,02	100254,4	30166,02	100254,4
18	Перемен	49486,58	2449,207	20,20515	3,76E-08	43838,7	55134,46	43838,7	55134,46
19									

Рис. 2.28. Регрессионный анализ

Обратите внимание, что рассчитанные значения коэффициентов аналогичны предыдущим результатам.

Технология аналитического выравнивания в LibreOffice Calc практически такая же, как и в MS Excel. На первом этапе осуществляется вставка графика, для чего выделяем набор данных и выбираем в главном меню пункт «Вставка» и элемент «Диаграмма». Выбираем тип диаграмм «Линии» и вид диаграммы «Только линии» и нажимаем кнопку «Готово». На получившейся диаграмме двойным щелчком мыши переходим в режим редактирования, выделяем ряд данных и нажимаем на нее правой кнопкой мыши. В появившемся контекстном меню выделяем пункт «Вставить линию тренда». В появившемся окне (рис. 2.29) переключаемся на закладку «Тип» и в следующем окне ставим галочки напротив пунктов «Показать уравнение» и «Показать коэффициент детерминации (R^2)».

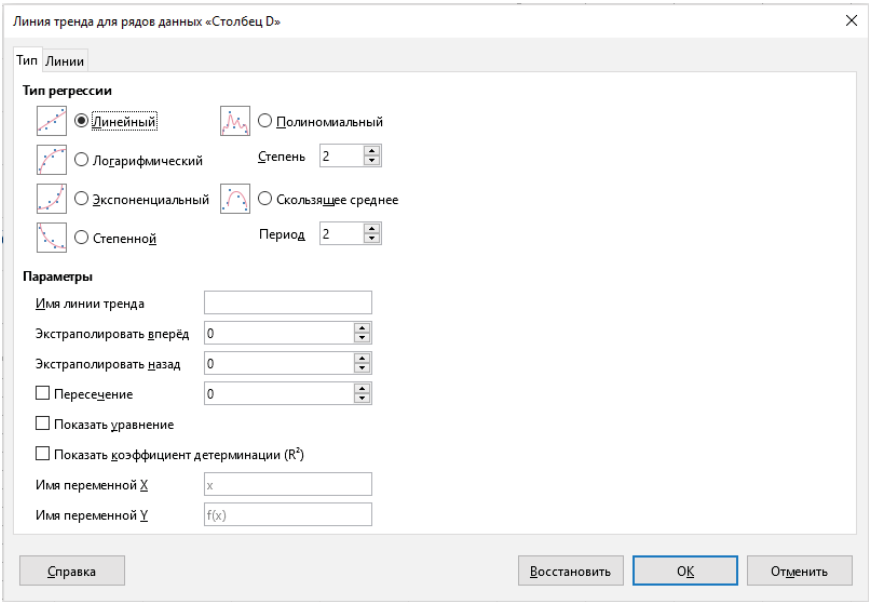


Рис. 2.29. Линия тренда для рядов данных

Как и при применении MS Excel, на диаграмме появятся линия тренда, уравнение и коэффициент детерминации (рис. 2.30).

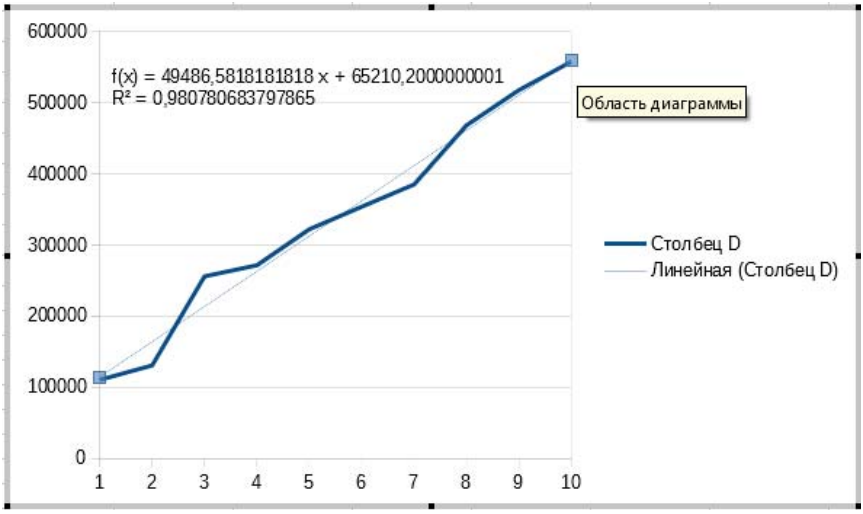


Рис. 2.30. Уравнение тренда

Полученные значения следующие: уравнение тренда y=49486,6*x+65210,2; R^2=0,98. Обратим внимание на то, что полученные результаты аналогичны тем, что получили при применении MS Excel.

2.3. Нелинейные модели и индексы сезонности

Наряду с линейным трендом можно построить и некоторые другие виды аппроксимирующих кривых[1], иногда их называют нелинейными[2]. Наиболее распространенными видами являются: экспоненциальная, логарифмическая, степенная и полиномиальная[3]. Например, при подборе линейного тренда для следующего графика (рис. 2.31) коэффициент аппроксимации R^2 составляет всего 0,59, что явно недостаточно для построения качественной модели.

[1] Иногда их еще называют сглаживающими кривыми.

[2] Здесь и далее нелинейные и криволинейные функции будут восприниматься как синонимы.

[3] Строго говоря, полином является степенной функцией, и степень полинома определяет количество пиков временного ряда.

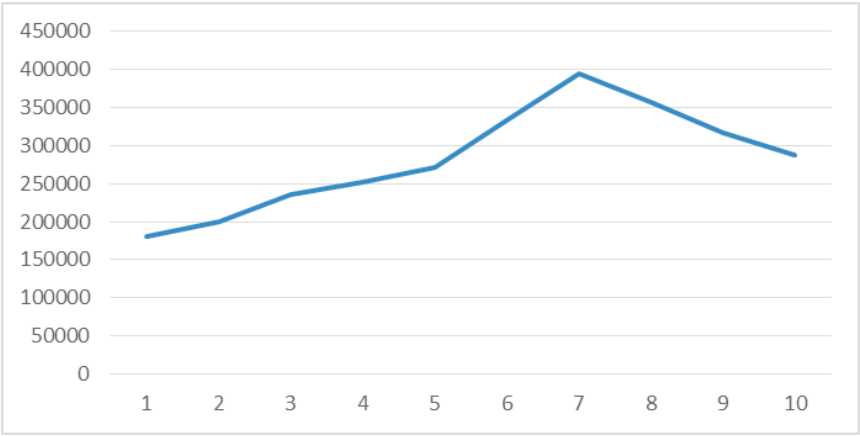


Рис. 2.31. График временного ряда

Кроме этого, при построении прогноза получается возрастающий тренд, хотя последние четыре отчетных периода наблюдалось снижение. В этой связи целесообразно попытаться подобрать иные формы зависимостей. В следующей таблице будут представлены различные виды уравнений и соответствующие им коэффициенты детерминации.

Таблица 2.3. Формы уравнений

Название	**Уравнение**	**R^2**
Линейная	$y = 17487x + 187210$	0,59
Экспоненциальная	$190738e^{0,0669x}$	0,65
Логарифмическая	$y = 78068\ln(x) + 165469$	0,69
Степенная	$y = 174546x^{0,3025}$	0,78
Полином 2-й степени	$y = -4213,4x^2 + 63833x + 94516$	0,81
Полином 3-й степени	$y = -1283x^3 + 16956x^2 - 33802x + 204597$	0,93

Очевидно, что для прогноза наилучшим образом подойдут полиномиальные формы уравнений. Однако высокие степени для практических расчетов применять сложно, поэтому мы ограничимся только полиномом 2-й степени. Кроме этого, используя полином, можно подобрать такую теоретическую функцию, которая будет иметь достаточно высокие значения коэффициента аппроксимации, однако интерпретировать полученные результаты будет зачастую невозможно.

Добавляем линию тренда, в параметрах выбираем полиномиальную и указываем степень 2, которая стоит по умолчанию (рис. 2.32).

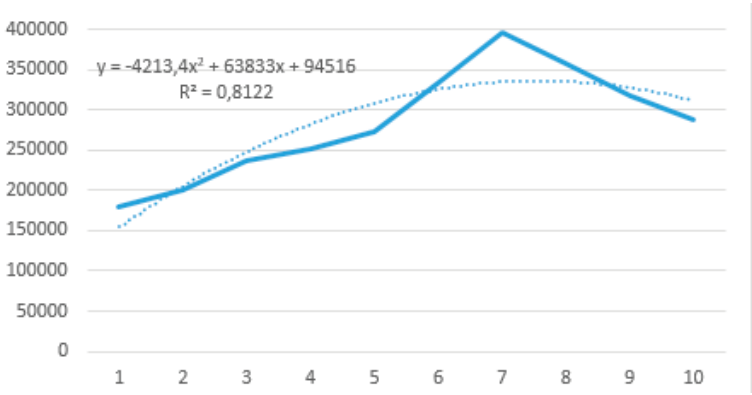


Рис. 2.32. График полинома 2-й степени

Вводим полученное уравнение в ячейку и фиксируем изменения нажатием клавиши «Enter»[1] (рис. 2.33).

СУММ | =-4213,4*B12^2+63833*B12+94516

	A	B	C	D	E	F	G	H
1								
2				Данные	Прогноз			
3		1	2009	180565				
4		2	2010	200848				
5		3	2011	236007				
6		4	2012	251631				
7		5	2013	272181				
8		6	2014	333879				
9		7	2015	395152				
10		8	2016	358235				
11		9	2017	317694				
12		10	2018	287672	=-4213,4*B12^2+63833*B12+94516			
13		11	2019					
14		12	2020					

Рис. 2.33. Ввод уравнения в таблицу

Распространим введенную формулу на следующие отчетные периоды и построим график (рис. 2.34). Аналогичным образом рассчитываются и остальные формы уравнений.

[1] Для возведения в квадрат можно просто умножить x на x либо использовать специальный знак «^». Таким образом, возведение в степень числа x будет обозначаться «x^2», где 2 – это показатель степени.

Рис. 2.34. График функции

Рассмотрим теперь методику построения модели при помощи Libre Office Calc. В целом алгоритм построения схож с MS Excel, но технология имеет некоторые отличительные особенности. Построим не полиномиальное уравнение, а логарифмическое. На первом этапе также строится диаграмма, для чего выделяется набор данных и активируется меню «Ставка» – «Диаграмма». В появившемся окне выбираем пункт «Линии» и 3-й значок «Только линии», и после нажатия кнопки «Готово» появится диаграмма. Переходим в режим редактирования двойным щелчком мыши, после нажатия правой кнопкой мыши на диаграмме активируется меню, в котором выбираем пункт «Вставить линию тренда». Далее переключаемся на закладку «Тип» и выбираем логарифмический тип регрессии (рис. 2.35).

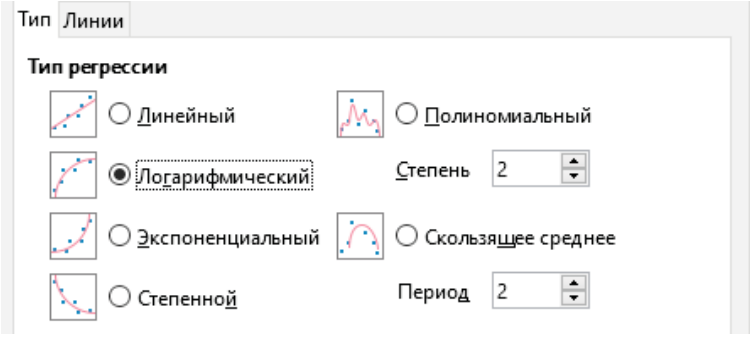


Рис. 2.35. Тип регрессии

Ставим галочки напротив пунктов «Показать уравнение» и «Показывать коэффициент детерминации (R^2)», и после нажатия кнопки «ОК» появится диаграмма следующего вида (рис. 2.36):

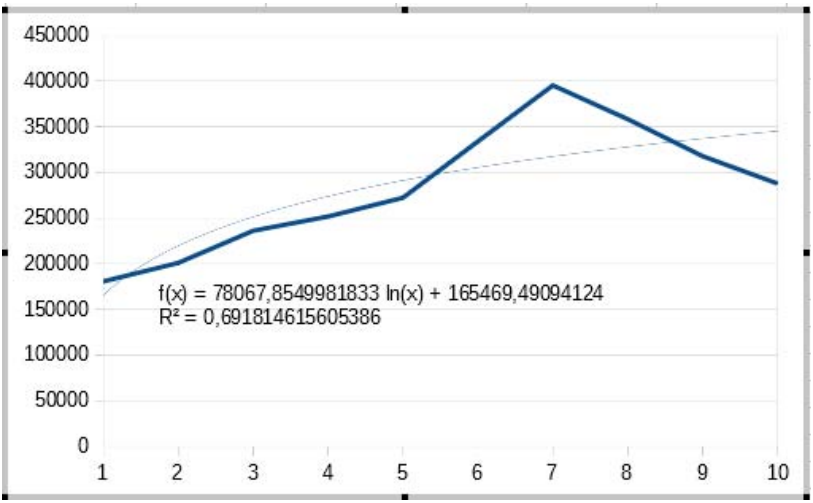
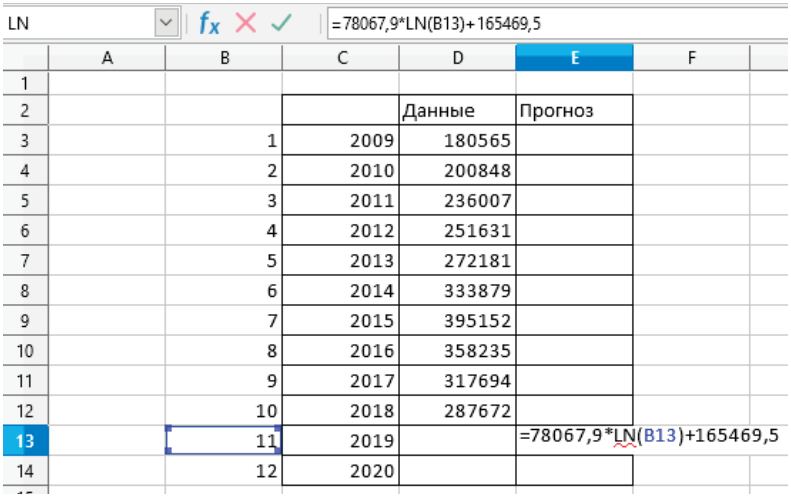


Рис. 2.36. Диаграмма с логарифмами

Введем полученное уравнение в ячейку и зафиксируем (рис. 2.37).



	A	B	C	D	E	F
1						
2				Данные	Прогноз	
3		1	2009	180565		
4		2	2010	200848		
5		3	2011	236007		
6		4	2012	251631		
7		5	2013	272181		
8		6	2014	333879		
9		7	2015	395152		
10		8	2016	358235		
11		9	2017	317694		
12		10	2018	287672		
13		11	2019		=78067,9*LN(B13)+165469,5	
14		12	2020			

Рис. 2.37. Ввод уравнения в ячейку

Распространим введенную формулу на все ячейки и построим график фактических и прогнозных значений (рис. 2.38).

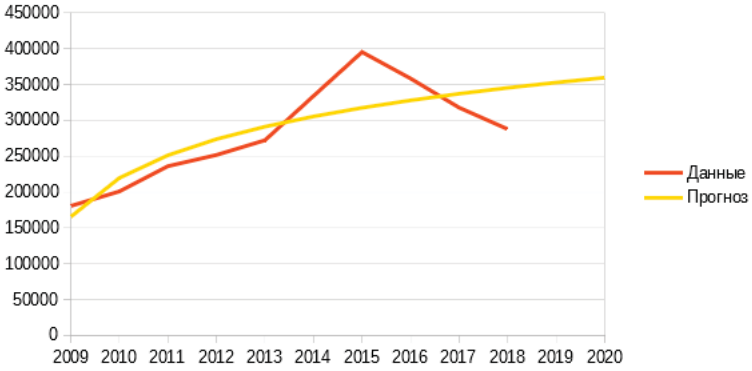


Рис. 2.38. График фактических и прогнозных значений

Для расчета неизвестных параметров уравнения также можно воспользоваться функцией «ПРЕДСКАЗ.ЛИНЕЙН», которая есть как в MS Excel, так и в LibreOffice Calc. Активируем нужную ячейку, куда нужно поместить прогноз, через «Мастер функций» находим функцию «ПРЕДСКАЗ.ЛИНЕЙН», выберем ее. В появившемся окне введем следующие параметры: «Значение» – это показатель x, то есть ячейка B13; «Данные Y» – это имеющиеся показатели y, то есть диапазон данных D3:D12; «Данные X» – это имеющиеся номера временного ряда, то есть диапазон номеров ряда B3:D12. Все должно получиться так, как указано на следующем рисунке (рис. 2.39).

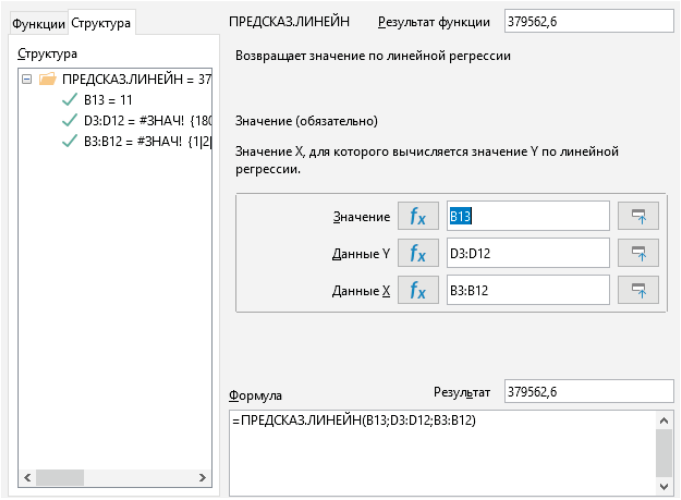


Рис. 2.39. Функция «ПРЕДСКАЗ.ЛИНЕЙН»

В результате в целевой ячейке мы получим прогнозное значение. Данный подход позволяет рассчитать только линейную форму уравнения, что не всегда целесообразно.

Индексы сезонности. Рассмотренный метод подбора аналитической функции предназначен для анализа годовых данных и не учитывает фактор сезонности. В случае если мы располагаем еженедельным или ежемесячным набором данных, то для прогнозирования необходимо учитывать фактор сезонности. Например, имеются ежемесячные данные о зарегистрированных преступлениях с января 2017 по декабрь 2019 г. (всего 36 точек). Необходимо построить прогноз на 2020 г. На первом этапе вставляем график и добавляем линию тренда (рис. 2.40). Нас интересует полученное уравнение $y = 11{,}947x + 14946$.

Рис. 2.40. Прогноз

На следующем этапе включим сезонные изменения в модель, для чего рассчитаем сезонные индексы по формуле:

$$\hat{y}_t = \frac{\sum y_t}{\sum y} \cdot 12,$$

где $\hat{y}_t$ – сезонный индекс t периода, y – набор всех данных, t – номер периода. Альтернативная формула для расчета индексов сезонности может быть представлена как отношение среднего арифметического за определенный период (месяц, квартал) к среднему арифметическому по всей совокупности данных:

$$\hat{y}_t = \frac{\overline{y_t}}{\bar{y}},$$

где $\overline{y_t}$ – среднее значение t периода, $\bar{y}$ – среднее значение всех данных. Рассчитаем сезонные индексы для каждого месяца (рис. 2.41).

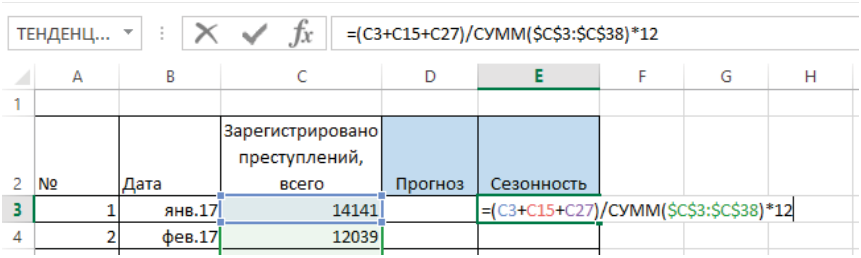


Рис. 2.41. Формула индексов сезонности

На следующем рисунке представлены рассчитанные индексы (рис. 2.42).

Сезонность
0,91
0,84
1,00
0,98
1,03
1,02
1,05
1,07
1,00
1,19
0,99
0,93

Рис. 2.42. Таблица рассчитанных индексов сезонности

На следующем этапе умножаем прогноз на соответствующий индекс (рис. 2.43).

38	36	дек.19	14379	
39	37	янв.20	15388	=C39*E3
40	38	фев.20	15400	

Рис. 2.43. Прогноз

И распространяем введенную формулу на все месяцы. Получившиеся значения представлены ниже (рис. 2.44).

39	37	янв.20	15388	14027
40	38	фев.20	15400	12922
41	39	мар.20	15412	15411
42	40	апр.20	15424	15059
43	41	май.20	15436	15887
44	42	июн.20	15448	15728
45	43	июл.20	15460	16254
46	44	авг.20	15472	16524
47	45	сен.20	15484	15469
48	46	окт.20	15496	18492
49	47	ноя.20	15508	15303
50	48	дек.20	15519	14389

Рис. 2.44. Прогнозная таблица

Приведем построенный график (рис. 2.45).

Рис. 2.45. Прогноз

Рассмотрим технологию прогнозирования с использованием LibreOffice Calc, но в отличие от предшествующего метода устраним влияние тренда. На первом этапе рассчитаем тренд при помощи функции «ПРЕДСКАЗ», как указано на следующем рисунке (рис. 2.46), и распространим формулу ниже.

	A	B	C	D	E	F
1		Месяц	Данные	Тренд		
2	1	Январь	14141	=ПРЕДСКАЗ(A2;$C$2:$C$37;$A$2:$A$37)		
3	2	Февраль	12039			
4	3	Март	16263			
5	4	Апрель	15125			
6	5	Май	15678			
7	6	Июнь	15693			

Рис. 2.46. Функция «Предсказ»

На следующем этапе рассчитаем индексы сезонности и также распространим введенную формулу на все ячейки (рис. 2.47).

	A	B	C	D	E	F	G	H
1		Месяц	Данные	Тренд	Индекс сезонности			
2	1	Январь	14141	14958	=СРЗНАЧ(C2;C14;C26)/СРЗНАЧ($C$2:$C$37)			
3	2	Февраль	12039	14970				

Рис. 2.47. Индекс сезонности

На следующем этапе умножаем полученные значения тренда на индекс сезонности и также распространяем на требуемый диапазон[1] (рис. 2.48).

38	37	Январь		15388	=D38*E2
39	38	Февраль		15400	12922
40	39	Март		15412	15411
41	40	Апрель		15424	15059

Рис. 2.48. Прогноз с сезонностью

Полученные прогнозные значения лучше всего отобразить на графике (рис. 2.49).

[1] В нашем примере на год.

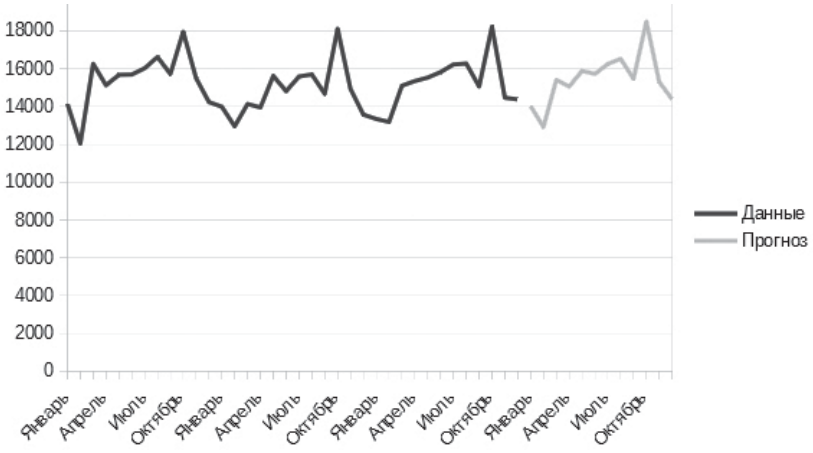


Рис. 2.49. График прогноза с сезонностью

Вторая группа методов анализа временных рядов основана на предположении, что каждый последующий показатель зависит от предыдущего. Эти методы основаны на сглаживании временного ряда, а также на экстраполяции.

Для расчета прогноза методом **скользящего среднего** необходимо определиться с периодом прогнозирования[1]. В нашем примере период прогнозирования представлен кварталом, то есть n=3.

Скользящее среднее рассчитывают по формуле:

$$S_t = \frac{y_1 + y_2 + \cdots + y_n}{n},$$

где S – прогнозные значения, n – период.

Например, для n=3

$$S_{\text{фев}} = \frac{14141 + 12039 + 16263}{3} \approx 14148,$$

$$S_{\text{мар}} = \frac{12039 + 16263 + 15125}{3} = 14476,$$

$$S_{\text{апр}} = \frac{16263 + 15125 + 15678}{3} = 15689$$

и т. д.

[1] Строгих рекомендаций по выбору периода расчета скользящего среднего не существует.

Введем формулу во вторую ячейку листа MS Excel следующим образом и распространим введенную формулу до предпоследней ячейки (рис. 2.50).

	A	B	C	D
1		Месяц	Данные	Скользящее среднее, n=3
2	1	Январь	14141	
3	2	Февраль	12039	=СРЗНАЧ(C2:C4)
4	3	Март	16263	14476
5	4	Апрель	15125	15689
6	5	Май	15678	15499

Рис. 2.50. Скользящее среднее

После того, как мы рассчитали скользящее среднее для всех периодов, нужно построить прогноз на январь 2020 г. по следующей формуле:

$$y_{t+1} = S_{t-1} + {}^{1}/_{k}\,(y_t - y_{t-1}),$$

где t – текущий отчетный период, y_{t+1} – прогнозируемый показатель, S_{t-1} – скользящее среднее за предыдущий период, t+1 – прогнозный период, k – интервал сглаживания, y_t – текущее значение показателя, y_{t-1} – предыдущее значение показателя. Подставляя значения в формулу, получим:

$$y_{\text{янв}} = 15689 + {}^{1}/_{3}\,(14379 - 14464) = 15661,$$

затем определим скользящую среднюю за декабрь:

$$S_{\text{дек}} = \frac{14141 + 12039 + 16263}{3} \approx 14148.$$

Повторяем процедуру. Строим прогноз на следующий месяц. Затем рассчитываем скользящее среднее и т. д.

MS Excel позволяет рассчитать скользящее среднее при помощи пакета анализа (рис. 2.51).

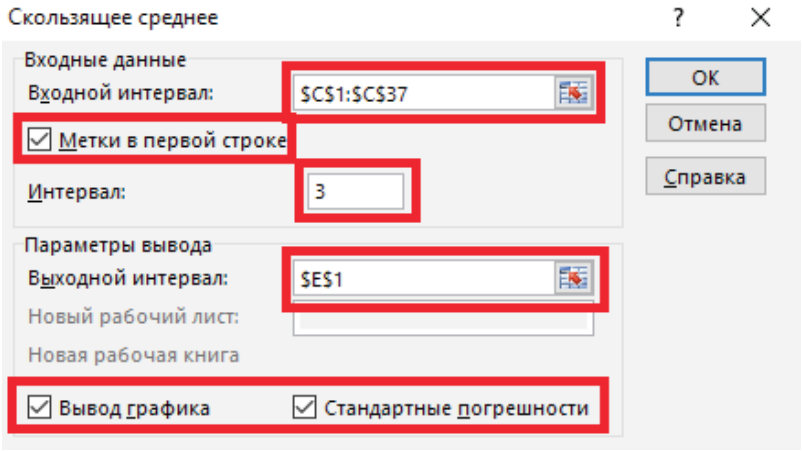


Рис. 2.51. Параметры скользящего среднего

В появившемся окне задаем параметры: «Входной интервал» – в нашем случае это имеющиеся данные в диапазоне C1:C37; «Метки в первой строке» – если выделяем заголовок данных; «Интервал» – в нашем примере это 3. Можно также отметить пункт «Вывод графика». Пункт «Стандартные погрешности» отмечать нет необходимости. После того, как задаем диапазон «Выходного интервала» (достаточно отметить одну ячейку, система автоматически расширит колонку в указанной ячейке до нужного размера) и нажимаем кнопку «ОК», мы получим следующие результаты[1] (рис. 2.52):

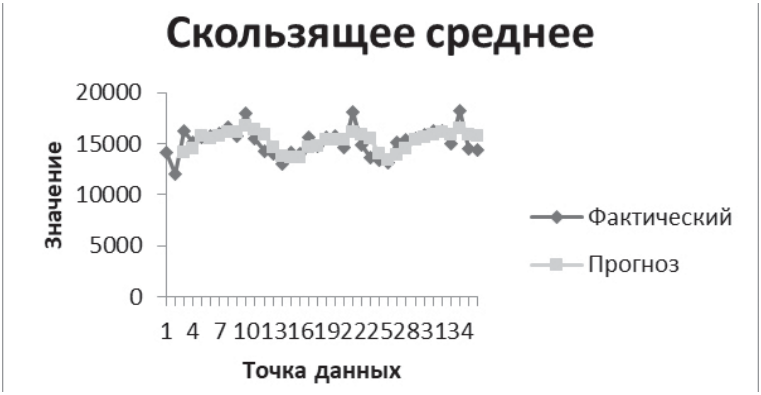


Рис. 2.52. Скользящее среднее

[1] Для удобства отображения данных нужно удалить значения «#Н/Д» в ячейках.

Второй метод основан на экспоненциальном сглаживании, отличительной особенностью которого является то, что прогнозировать с его помощью можно только на один период вперед. Кроме этого, простое экспоненциальное сглаживание не учитывает сезонность. Расчет осуществляется по формуле:

$$S_{t+1} = \alpha \cdot y_t + (1 - \alpha) \cdot S_t,$$

где t – текущий отчетный период, S_{t+1} – прогнозируемый показатель, S_t – сглаженный показатель за текущий период, y_t – текущее значение показателя, α – параметр сглаживания (фактор затухания) принимает значения от 0,1 до 0,9. Рассмотрим пример, когда параметр сглаживания (фактор затухания) α=0,5. Подставляя значения в формулу, получим:

$$S_{t+1} = 0{,}5 \cdot 14379 + 0{,}5 \cdot 15679{,}2 \approx 15029.$$

На следующем рисунке представлены параметры соответствующего инструмента в MS Excel (рис. 2.53).

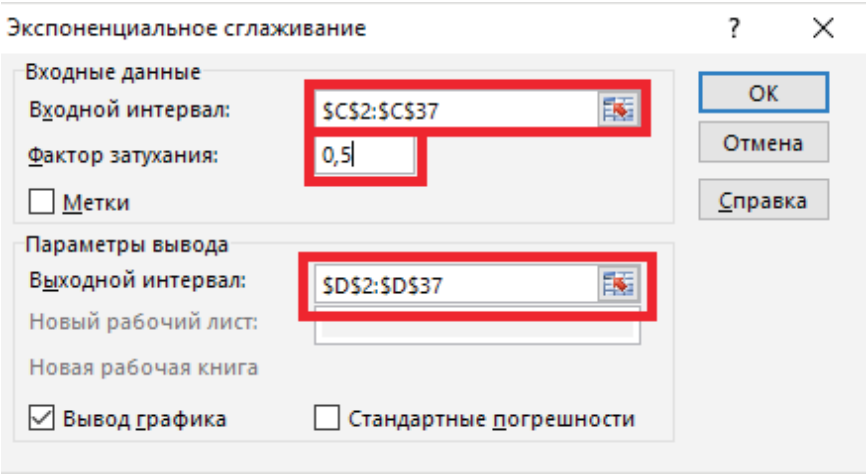


Рис. 2.53. Экспоненциальное сглаживание

На следующих двух рисунках (рис. 2.54 и рис. 2.55) представлены две диаграммы, в которых указан разный интервал сглаживания (0,5 и 0,05).

Рис. 2.54. Экспоненциальное сглаживание (α=0,5)

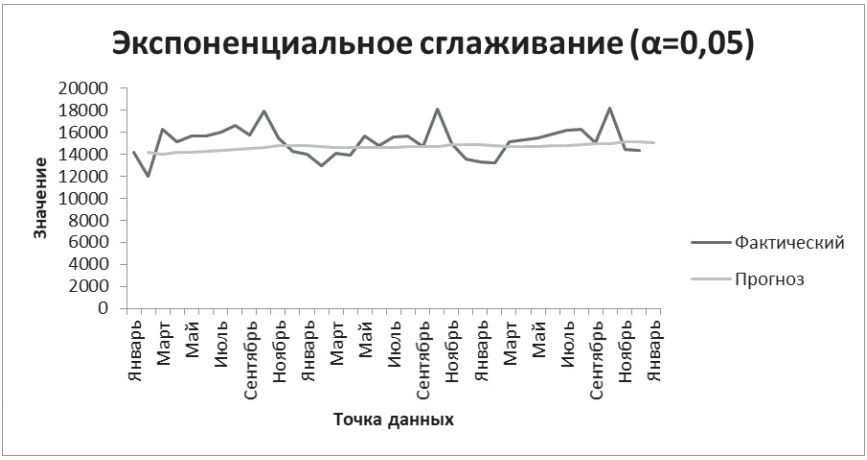


Рис. 2.54. Экспоненциальное сглаживание (α=0,05)

Отчетливо видна разница между двумя моделями с разными интервалами затухания.

На практике основной проблемой является выбор значения фактора затухания α. Один из подходов заключается в выборе параметра сглаживания по формуле: $\alpha = 2/(n+1)$, то есть чем длиннее временной ряд, тем ниже фактор затухания α.

Данные методы в качестве фактора, оказывающего влияние на исследуемый временной ряд, оперируют только одной переменной, это временная переменная *t*. В следующей главе рассмотрим

ситуацию, когда на исследуемый процесс (явление) оказывает влияние иной фактор (факторы).

С версии MS Excel 2016 появилось несколько новых функций, позволяющих осуществлять прогнозирование с использованием аддитивного (ПРЕДСКАЗ.ETS.ADD) или мультипликативного (ПРЕДСКАЗ.ETS.MULT) алгоритмов экспоненциального сглаживания (ETS). Функции «ПРЕДСКАЗ.ETS» позволяют рассчитать будущее значение на основе ретроспективных данных. Функции «ПРЕДСКАЗ.ETS.PI.ADD» и «ПРЕДСКАЗ.ETS.PI.MULT» позволяют рассчитать прогнозный интервал, а «ПРЕДСКАЗ.ETS.STAT.ADD» и «ПРЕДСКАЗ.ETS.STAT.MULT» позволяют рассчитать статистку, соответственно, для аддитивного и мультипликативного алгоритмов экспоненциального сглаживания. Кроме этого, функция «ПРЕДСКАЗ.ETS.СЕЗОННОСТЬ» позволяет определить количество элементов в периоде. Заметим, что данные функции поддерживаются Libre Offce Calc без каких-либо ограничений. Применение данных функций при решении задач анализа и прогнозирования существенно упрощает объемы «ручной» работы.

Следует иметь в виду, что набор функций чрезвычайно широк, и решить те или иные задачи возможно за счет различных способов. Например, для ручного расчета уравнения тренда можно использовать две функции «ОТРЕЗОК» и «НАКЛОН», которые возвращают значения коэффициентов *a* (отрезок, отсекаемый на оси линией регрессии) и *b* (наклон линии регрессии) линейной модели. Также наряду с функцией «ПРЕДСКАЗ» возможно использование функций «ТЕНДЕНЦИЯ» и «РОСТ».

Рассмотренные здесь методы анализа временных рядов не претендуют на полное их описание. Существует достаточно большой класс авторегрессионных моделей (*англ. AutoRegressive*, AR-модели), а также их различные комбинации. Например, сочетание скользящего среднего (*англ. moving average,* MA-модели) и авторегрессионной AR-модели называется ARMA (*англ. AutoRegressive Moving Average*). Расширение ARMA для нестационарных временных рядов называется ARIMA (*англ. AutoRegressive Integrated Moving Average*), эта же модель с учетом сезонности – SARIMA (*англ. Season AutoRegressive Integrated Moving Average*).

Глава 3. Регрессионный анализ

В данной главе рассматриваются линейные и нелинейные модели анализа данных, модели с конкретными переменными, методы оценки коэффициентов модели и компьютерные технологии для построения парных и множественных регрессионных моделей.

3.1. Модели линейной регрессии

Линейная регрессия – это статистически используемая регрессионная модель зависимости переменной (объявленной, зависимой) *y* от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) *x* с функцией линейной зависимости.

Основным методом оценки неизвестных параметров регрессионной модели является метод наименьших квадратов (*англ. ordinary least squares*, OLS). Модель линейной регрессии является наиболее широко используемой и изучаемой в эконометрике.

Основная цель регрессионного анализа – оценить функциональную связь между независимой переменной *X* и условным ожиданием зависимой переменной *Y*. Парная регрессия – это модель, в которой (среднее) теоретическое значение зависимой переменной *Y* принимается во внимание как функция независимой переменной *X*. Множественная регрессия – это модель, в которой рассматривается теоретическое (среднее) значение зависимой переменной *Y* как функции нескольких независимых переменных *X*1, *X*2, ... *X*m.

Спецификация модели – формулировка типа модели на основе соответствующей теории взаимосвязи между переменными. Определяется состав переменных и математическая функция, отражающая взаимосвязь между ними.

Мультиколлинеарность – это линейная связь между двумя или более независимыми переменными[1].

Криминологические модели являются характерными представителями факторных социальных моделей, с помощью которых они пытаются формализовать сложные социальные процессы, связанные с противоправным поведением, возникновением преступности и формированием механизмов правового регулирования.

Построение модели – это итеративный процесс поиска эффективных независимых переменных. Основная цель – попытаться объ-

[1] *Торопов Б. А., Гонов Ш. Х.* Статистические методы принятия управленческих решений: сборник задач (задачник). Москва: Академия управления МВД России, 2019. 76 с.

яснить зависимые переменные, которые необходимо моделировать и пересмотреть с помощью инструмента регрессии, и определить, какие величины являются эффективными предикторами. Затем удаляем и (или) добавляем независимые переменные, пока не найдется наиболее подходящая модель регрессии.

Наиболее распространенными моделями регрессии являются линейные модели, однако функциональная форма зависимости может быть различной. На следующем рисунке представлены наиболее распространенные функциональные формы (рис. 3.1).

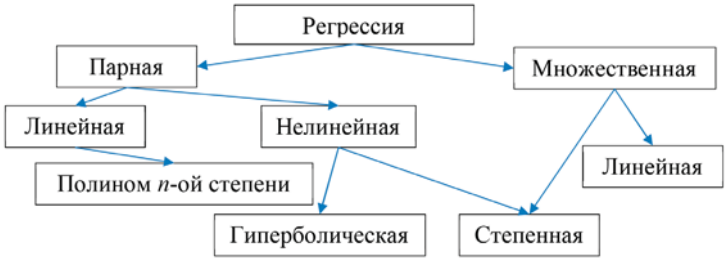


Рис. 3.1. Функциональные формы моделей

В графическом виде некоторые функции представлены ниже (рис. 3.2).

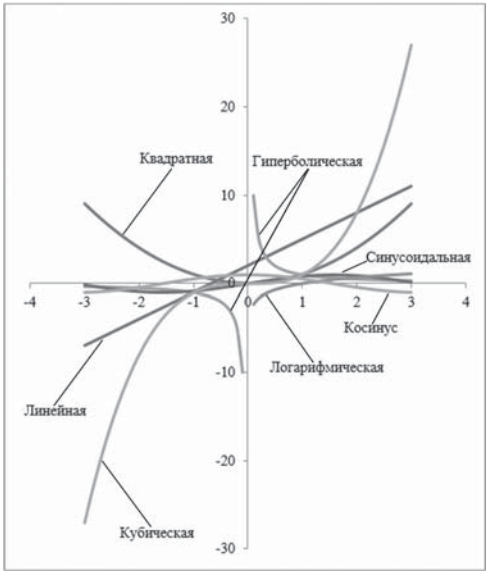


Рис. 3.2. Функции

Для оценки неизвестных параметров уравнения могут применяться различные методы, наиболее распространенным является МНК. В таблице представлены некоторые виды моделей и соответствующие им математические выражения (таблица 3.1).

Таблица 3.1. Некоторые виды моделей

Наименование	**Уравнение**
Линейная	$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon$
Полином k-ой степени	$y = \alpha + \beta_{11} x_1 + \beta_{12} x_1^2 + \cdots + \beta_{1k} x_1^k + \cdots$ $\ldots + \beta_{p1} x_p + \beta_{p2} x_{p2}^2 + \cdots + \beta_{pk} x_p^k + \varepsilon$
Обратная	$y = \dfrac{1}{\alpha + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon}$
Степенная	$y = \alpha \cdot x_1^{\beta_1} \cdot \ldots \cdot x_p^{\beta_p} \cdot \varepsilon$
Показательная	$y = \alpha \cdot \beta_1^{x_1} \cdot \ldots \cdot \beta_p^{x_p} \cdot \varepsilon$
Полулогарифмическая	$y = \alpha + \beta_1 \ln x_1 + \cdots + \beta_k \ln x_p + \varepsilon$

В общем виде модель парной регрессии может быть представлена как $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$, $i = 1, \ldots n$, где α и β – неизвестные параметры модели, y_i – зависимая переменная, x_i – зависимая переменная (регрессор) и ε_i – случайная компонента (случайная ошибка) модели, n – количество наблюдений, i – номер наблюдения.

В «Анализе данных» выбираем пункт «Регрессия» и в появившемся окне (рис. 3.3) в качестве входного интервала *Y* задаем значения зависимой переменной, а в качестве входного интервала *X* – значения независимой переменной (независимых переменных).

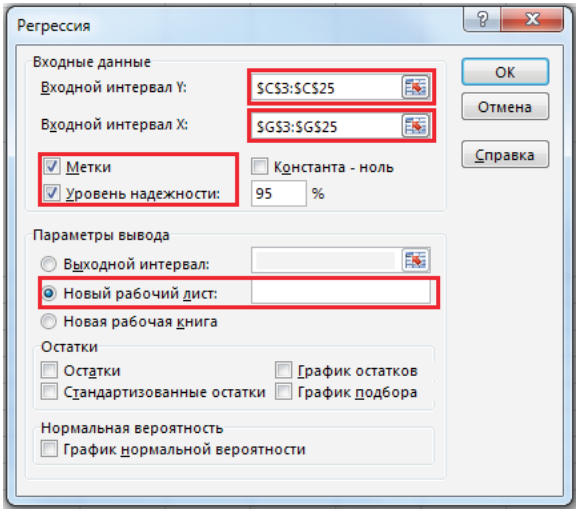


Рис. 3.3. Параметры расчета регрессии

Например, в качестве параметра *Y* будут выступать преступления, а в качестве *X* – численность сотрудников патрульно-постовой службы полиции (далее – ППСП). После нажатия кнопки «ОК» на новом листе появятся следующие результаты:

Регрессионная статистика	
Множественный R	0,85
R-квадрат	0,73
Нормированный R-квадрат	0,71
Стандартная ошибка	3064,8
Наблюдения	22

Дисперсионный анализ

	df	**SS**	**MS**	**F**	**Значимость F**
Регрессия	1	504026628	504026628	53,6612454	4,42E-07
Остаток	20	187854987	9392749,346		
Итого	21	691881615			

	Коэффициенты	**Стандартная ошибка**	**t-статистика**	**P-Значение**
Y-пересечение	10159,20099	4151,03493	2,44738991	0,02374833
ППСП	15,81146374	2,15844856	7,325383638	4,4218E-07

Проинтерпретируем некоторые полученные результаты.

R-квадрат. Полученная модель описывает 73 % вариации, оставшиеся 27 % – это другие неучтенные в модели факторы[1].

P-Значение. На 5 % уровне вероятности будем принимать данное значение не больше 0,05.

F-статистика *Фишера*. Должна быть больше, *чем значимость F.*

Полученная модель представлена в виде выражения, на основе которой рассчитаем модельные значения для следующей таблицы:

№	Преступления	ППСП	Модель	№	Преступления	ППСП	Модель
1	39662	1972	41317	12.	41735	2080	43023
2	44492	2013	41965	13.	41809	2252	45741
3	48214	1856	39484	14.	36126	1731	37509
4	48135	2199	44903	15.	34706	1483	33591
5	48023	2251	45725	16.	35584	1562	34839
6	38913	2062	42739	17.	33839	1638	36040
7	40991	2032	42265	18.	33476	1536	34428
8	44383	2194	44824	19.	36259	1409	32421
9	48088	2211	45093	20.	34336	1575	35044
10	45369	2327	46926	21.	31726	1569	34949
11	45968	2264	45930	22.	32303	1566	34902

Построим линейный график с фактическими и модельными данными (рис. 3.4).

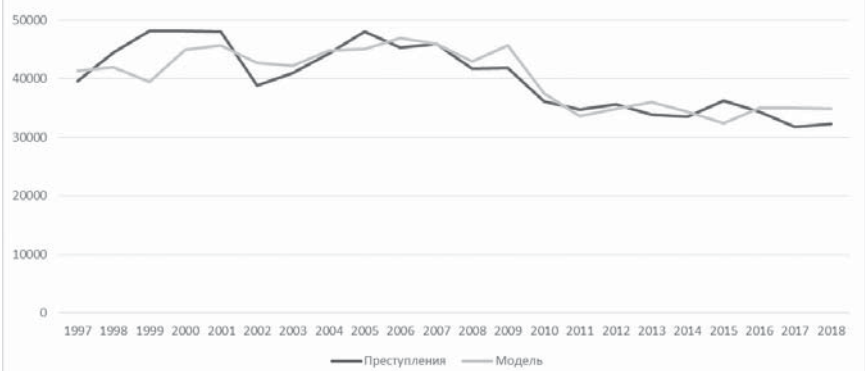


Рис. 3.4. График фактических и модельных значений

[1] Для практических целей моделирования значение R-квадрат принимается не менее 0,7.

На следующем этапе рассчитаем прогнозные значения уровня преступности на 2019 и 2020 г. при условии увеличения сотрудников ППСП ежегодно на 100 чел. В этом случае прогноз составит:

	ППСП	**Преступлений**
2019 г.	1666	36482
2020 г.	1766	38062

При исследовании организационных (социальных и экономических) систем парная регрессия не в полной мере может отразить степень влияния различных факторов, поэтому в анализе данных большее распространение получили модели множественной регрессии. В общем виде модель множественной регрессии может быть представлена как:

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \beta_k \cdot x_k + \varepsilon,$$

где a и β_k – неизвестные параметры модели, y – зависимая переменная, x_k – зависимая переменная (регрессор), ε – случайная компонента (случайная ошибка) модели и k – количество зависимых переменных (регрессоров).

В компактном виде данная модель записывается как:

$$y = a + \sum_{i=1}^{k} \beta_k x_k + \varepsilon.$$

На предыдущем этапе мы рассмотрели ситуацию, когда в качестве независимой (влияющей) переменной выступал только один фактор (численность сотрудников ППСП). Теперь рассмотрим ситуацию, когда в качестве независимых переменных выступают несколько факторов (множественная регрессия).

В инструменте «Анализ данных» выбираем пункт «Регрессия» и в появившемся окне задаем входной параметр *Y* и входной параметр *X*. В качестве *Y* будут выступать зарегистрированные преступления, а в качестве *X* факторы – безработица, численность сотрудников ППСП и численность сотрудников ДПС. После нажатия кнопки «ОК» на новом листе появятся следующие результаты:

Регрессионная статистика	
Множественный R	0,92
R-квадрат	0,84
Нормированный R-квадрат	0,81

Стандартная ошибка	2475,8
Наблюдения	22

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	3	5,82E+08	1,94E+08	31,62653	2,17E-07
Остаток	18	1,1E+08	6129378		
Итого	21	6,92E+08			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересечение	-3795,71	6326,487	-0,59997	0,556004
ППСП	10,77304	2,248609	4,790978	0,000146
ДПС	8,226943	3,725148	2,208488	0,040419
Численность безработных	299,2542	95,89231	3,120732	0,005905

Проинтерпретируем полученные результаты.

1. *R*-квадрат. Полученная модель описывает 84 % вариации, оставшиеся 16 % – это другие неучтенные в модели факторы.

2. *P*-Значение. На 5 % уровне вероятности будем принимать данное значение не больше 0,05.

3. *F*-статистика. Должна быть больше, чем значимость F.

Полученная модель выглядит как $y = -3795{,}7 + 10{,}8 \cdot x_1 + 8{,}2 \cdot x_2 + 299{,}2 \cdot x_3$, где x_1 – численность сотрудников ППСП, x_2 – численность сотрудников ДПС, x_3 – численность безработных. На основе полученной модели рассчитаем модельные значения для следующей таблицы:

Зарегистрировано преступлений	Численность сотрудников ППСП	Численность сотрудников ДПС	Численность безработных, тыс. чел.	Модель
39662	1972	1666	40,1	
44492	2013	1781	42,2	
48214	1856	1742	40,2	
48135	2199	1638	40,6	
48023	2251	1867	35,7	
38913	2062	1602	23,7	

40991	2032	1933	34,7	
44383	2194	1943	32,9	
48088	2211	1995	29,3	
45369	2327	1607	40,8	
45968	2264	1577	41,4	
41735	2080	1649	34,4	
41809	2252	1653	28,5	
36126	1731	1491	23,9	
34706	1483	1658	22,9	
35584	1562	1824	21,1	
33839	1638	1821	22,2	
33476	1536	1765	29,6	
36259	1409	1558	32,8	
34336	1575	1572	30,9	
31726	1569	1561	28,0	
32303	1566	1395	28,1	

В соответствующую ячейку введем следующее выражение (рис. 3.5):

Q	R	S	T	U	V
Преступления	ППС	Модель (I)	Модель (II)		
39662	1972	41316,8	=-3795,7+10,8*G4+8,2*H4+299,2*I4		
44492	2013	41964,6	45175,14		
48214	1856	39484	42561,34		
48135	2199	44903,4	45532,62		
48023	2251	45725	46505,94		

Рис. 3.5. Уравнение в ячейке по модели

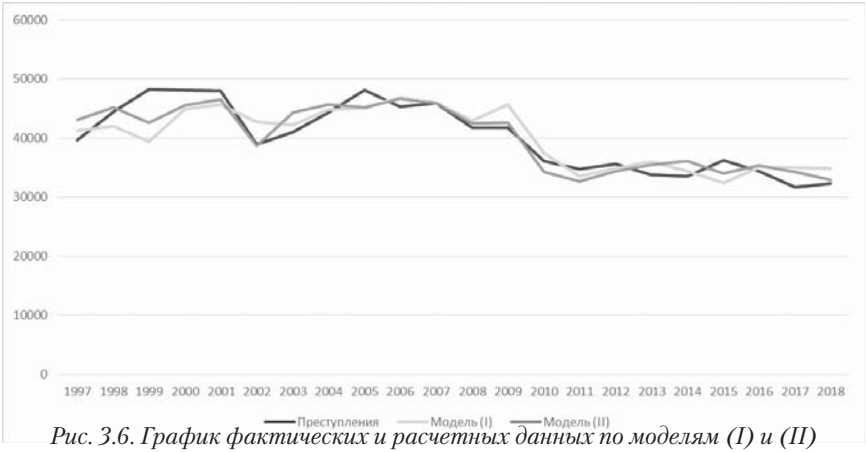


Рис. 3.6. График фактических и расчетных данных по моделям (I) и (II)

На следующем этапе рассчитаем прогнозные значения уровня преступности на 2019 и 2020 г. при условии увеличения сотрудников ППСП и ДПС ежегодно на 50 чел. и сокращения безработных на 1 тыс. чел. по сравнению с 2018 г. В этом случае прогноз составит:

	ППСП	**ДПС**	**Безработных**	**Преступлений**
2019 г.	1616	1445	27,1	33614,42
2020 г.	1666	1495	26,1	34265,22

Ранее мы рассмотрели процедуру построения линейной модели парной регрессии. Теперь рассмотрим процедуру построения регрессионной модели в виде полинома n-ой степени:

$$y = a + b_1 x_1 + b_2 {x_2}^2 + \cdots + b_n {x_n}^n.$$

На практике достаточным считается показатель 2-й и 3-й степени, применение более старших степеней связано с вычислительными сложностями, и рассматривать мы их не будем. В общем виде модель парной регрессии в виде полинома 2-й степени представлена ниже:

$$y = a + b_1 x_1 + b_2 {x_2}^2.$$

Вернемся к задаче, сформулированной ранее. На первом этапе произведем замену переменных. Так, X_1 – это численность ППСП, X_2 – это показатель X_1, возведенный в квадрат, Y – количество зарегистрированных преступлений. В ячейку D4 введем следующее значение (рис. 3.7):

	A	B	C	D
		Преступления	ППС	ППС^2
		Y	X_1	X_2
1997		39662	1972	=C4^2
1998		44492	2013	
1999		48214	1856	

Рис. 3.7. Возведение в квадрат показателя X_1

Таким образом, в ячейке появился показатель X, возведенный в квадрат. Распространим введенную формулу на все ячейки, для этого находим в правом нижнем углу ячейки значок + и, удерживая, протягиваем до конца. Формула автоматически распространится на все ячейки[1].

Далее рассчитаем неизвестные коэффициенты (a, b_1, b_2), для чего воспользуемся встроенной в MS Excel функцией «ЛИНЕЙН». Активируем свободную ячейку, в нашем примере это J4, и вставляем функцию «ЛИНЕЙН». В появившемся окне заполняем указанные поля (рис. 3.8).

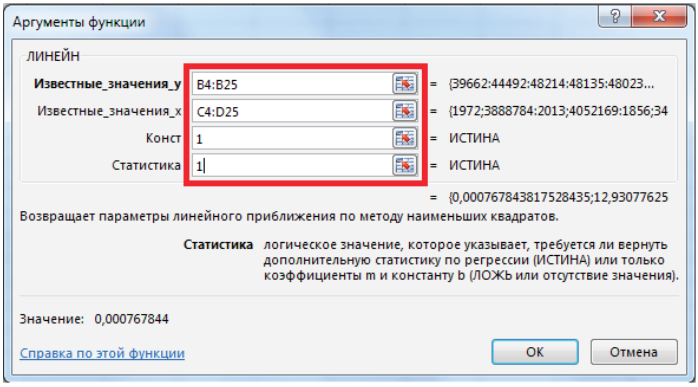


Рис. 3.8. Аргументы функции «ЛИНЕЙН»

В целевой ячейке появится некоторое значение. Проблема в том, что данная функция возвращает только одно значение, и это значение последнего коэффициента. Для того чтобы получить значения остальных коэффициентов, необходимо выполнить следующие действия: выделяем ячейки по следующей схеме: пять строк, начиная от уже рассчитанного коэффициента, и три столбца[2] (рис. 3.9).

[1] Можно использовать функции «Копировать» и «Вставить», результат будет тот же.
[2] Три (3) – это количество неизвестных коэффициентов.

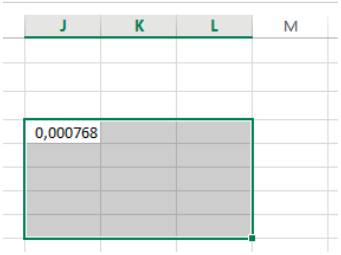


Рис. 3.9. Выделение целевого массива

Далее нажимаем клавишу «F2» и сочетание клавиш «Ctrl + Shift + Enter». Система заполнит массив данными следующего вида (рис. 3.10):

0,000767844	12,93077626	12790,26
0,011184943	42,02046109	38561,6
0,728554136	3143,986763	#Н/Д
25,49777034	19	#Н/Д
504073212,4	187808402,6	#Н/Д

Рис. 3.10. Массив данных

Неизвестные коэффициенты находятся в первой строке и располагаются справа налево. Таким образом, в нашем примере коэффициенты[1] равны:

$a \approx 12790{,}26;\ b_1 \approx 12{,}93;\ b_2 \approx 0{,}00077;$

$$y = 12790{,}26 + 12{,}93 \cdot x_1 + 0{,}00077 \cdot x_2{}^2.$$

Введем полученные значения в таблицу и распространим введенную формулу по всем ячейкам (рис. 3.11).

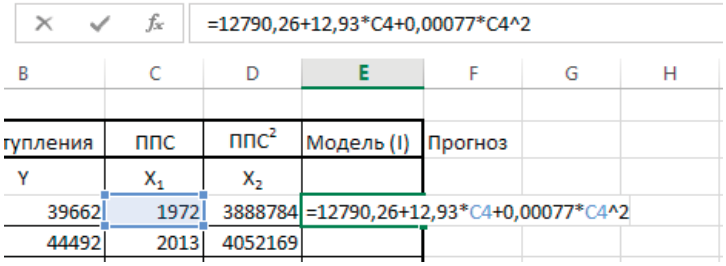


Рис. 3.11. Ввод формулы

[1] Коэффициенты даны с округлением по общим правилам.

На следующем этапе построим график фактических и модельных значений (рис. 3.12).

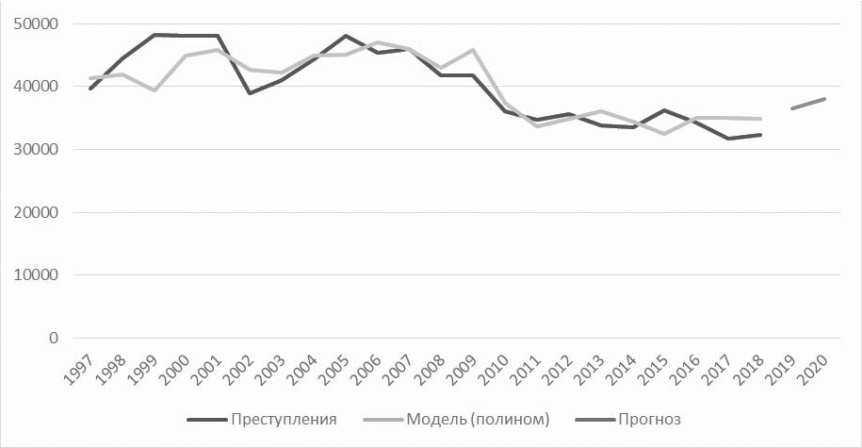


Рис. 3.12. График фактических и модельных значений. Полином (I)

На следующем этапе рассчитаем прогнозные значения уровня преступности на 2019 и 2020 г. при условии увеличения сотрудников ППСП ежегодно на 100 чел. Так, прогноз составит:

	ППСП	**Преступлений**
2019 г.	1666	36469
2020 г.	1766	38026

Чаще всего в качестве инструмента моделирования применяются линейные модели, но ограничиваться только их применением при исследовании не вполне корректно. Зачастую бывает так, что из-за сложности социально-правовых явлений и процессов их моделирование возможно только на основе нелинейных моделей.

3.2. Модели нелинейной регрессии

На предыдущем этапе мы рассмотрели ситуацию, когда в качестве зависимой (влияющей) переменной выступают несколько факторов. Зачастую бывает так, что линейные модели не в полной мере описывают социальное явление (процесс) либо полученные результаты свидетельствуют о низком качестве построенной модели.

Существуют несколько методов улучшения построенных моделей, в том числе когда линейные модели дают неудовлетворитель-

ные результаты. В этом случае можно перейти от линейной формы модели к нелинейной. Существует два класса таких моделей:

1. Нелинейные относительно независимых переменных, но линейные по параметрам (например, полиномиальная модель регрессии);

2. Нелинейные по независимым переменным и по оцениваемым параметрам (например, степенная модель регрессии).

Так, степенные функции применяются для исследования зависимости между объемом произведенной продукции и двух факторов производства – труда и капитала. Наряду с линейной одной из наиболее распространенных производственных функций является модель типа Кобба – Дугласа, имеющая форму степенной функции. В общем виде функцию этого типа можно представить так:

$$Y_i = AK_i^{\alpha_i} L_i^{\beta_i}, \alpha_i\, \beta_i \in [0,1], i = 1, \dots, N,$$

где α_i и β_i – оцениваемые параметры модели, Y_i – результаты, K_i – капитал и L_i – труд. Данную функцию удобно представить в логарифмическом виде[1]:

$$\ln Y_i = \ln A + \alpha \ln K_i + \beta \ln L_i + \ln \varepsilon_i.$$

Рассмотрим процедуру построения модели на основе исходных данных, полученных из Единой межведомственной информационно-статистической системы (ЕМИСС). На первом этапе из ЕМИСС (https://www.fedstat.ru) получим данные о валовом внутреннем продукте, основных средствах и численности занятых и составим следующую таблицу (таблица 3.1):

Таблица 3.1. Исходные данные для построения производственной функции Кобба – Дугласа

Год	Валовой внутренний продукт (в ценах 2016 г., млрд руб.)	Основные средства (млрд руб.)	Численность занятых, (млн чел.)
	Y	K	L
2011	81 750,6	238 100,9	70 856,6
2012	85 040,3	263 599,4	71 545,4

[1] Данная операция называется методом замены переменных.

2013	86 533,1	276 494,9	71 391,5
2014	87 170,2	286 081,4	71 539,0
2015	85 450,6	306 326,1	72 323,6
2016	85 616,1	340 098,5	72 392,6
2017	87 179,3	325 857,1	72 142,0
2018	89 626,6	348 492,7	72 354,4
2019	91 596,7	499 311,6	71 764,5
2020	89 138,9	507 269,2	70 460,8

На листе MS Excel эти данные будут располагаться в диапазоне данных [C3:E12]. Произведем логарифмирование переменных, для этого в соответствующие ячейки [F3:H12] введем значения, взятые с натуральными логарифмами. Для этого в ячейку [F3] введем следующее выражение: «=LN(C3)» (рис. 3.13).

Рис. 3.13. Ввод выражения

Фиксируем изменения нажатием клавиши «Enter» и, используя инструмент автозаполнения, распространяем введенное выражение по строкам и столбцам. Должны получиться следующие результаты (рис. 3.14):

ln(Y)	ln(K)	ln(L)
11,3114	12,3804	11,1684
11,3509	12,4822	11,1781
11,3683	12,5299	11,1759
11,3756	12,5640	11,1780
11,3557	12,6324	11,1889
11,3576	12,7370	11,1899
11,3757	12,6942	11,1864
11,4034	12,7614	11,1893
11,4252	13,1210	11,1811
11,3980	13,1368	11,1628

Рис. 3.14. Расчет значений

Далее рассчитаем неизвестные коэффициенты A, α, β методом наименьших квадратов с использованием функции MS Excel «ЛИНЕЙН». Данная функция возвращает параметры линейного приближения по МНК и имеет несколько аргументов:

1. Известные значения Y – зависимая переменная (валовой внутренний продукт);

2. Известные значения X – независимые переменные, или регрессоры (x_1 – основные средства и x_2 – численность занятых);

3. Логическое выражение, определяющее равенство нулю свободного члена (=1);

4. Логическое выражение, определяющее необходимость возврата дополнительной статистики (=1)[1].

Выделяем 5 строк и 3 (количество независимых переменных + 1) колонки пустых ячеек и вводим следующее выражение (рис. 3.15):

ln(Y)	ln(K)	ln(L)
11,3114	12,3804	11,1684
11,3509	12,4822	11,1781
11,3683	12,5299	11,1759
11,3756	12,5640	11,1780
11,3557	12,6324	11,1889
11,3576	12,7370	11,1899
11,3757	12,6942	11,1864
11,4034	12,7614	11,1893
11,4252	13,1210	11,1811
11,3980	13,1368	11,1628

=ЛИНЕЙН(F3:F12;G3:H12;1;1)

Рис. 3.15. Расчет неизвестных параметров

После нажатия клавиш «Ctrl+Shift+Enter» система заполнит выделенный диапазон следующими значениями (рис. 3.16):

[1] Строго говоря, для нашего примера возвращать дополнительную статистику не требуется, показатели качества модели и др. здесь не рассматриваются.

0,75131773	0,107053162	1,612531355
0,668755203	0,024314527	7,507181558
0,740680583	0,018373505	#Н/Д
9,996868219	7	#Н/Д
0,006749599	0,0023631	#Н/Д

Рис. 3.16. Результаты расчета

Значения коэффициентов располагаются слева направо, то есть A=1,61, α=0,107, β=0,751. Однако ввиду того, что при расчете мы приводили степенное выражение к натуральным логарифмам, нужно произвести обратное преобразование для коэффициента A. Для этого можно произвести следующие действия: число e (2,7182818284590452)[1] возвести в степень числа A:

$$2{,}7182818284590452^{1{,}61}.$$

Это действие можно произвести при помощи MS Excel (рис. 3.17):

	=I16^H16
1,612531355	2,718281828

Рис. 3.17. Расчет

Аналогичные результаты мы получим при использовании функции EXP, которая принимает только один аргумент – степень, в которую возводится основание e. Для этого в соседнюю ячейку введем следующее выражение (рис. 3.18):

=EXP(H16)	5,01549116
1,612531355	2,718281828

Рис. 3.18. Расчет

После нажатия клавиши «Enter» в ячейке появится число, идентичное рассчитанному ранее вручную.

Подставив рассчитанные значения в уравнение, получим следующее выражение: $Y = 5{,}01 \cdot K^{0{,}107} \cdot L^{0{,}751}$.

Рассчитаем модельные значения (рис. 3.19) и заполним следующую таблицу (таблица 3.2):

[1] Число e (число Эйлера) – это основание натурального логарифма, приблизительно равно 2,7182818284590452.

Рис. 3.19. Расчет модельных значений

Таблица 3.2. Исходные и модельные данные функции Кобба – Дугласа

	Валовой внутренний продукт (в ценах 2016 г., млрд руб.)	Модель
	Y	Y*
2011	81 750,6	83 193,5
2012	85 040,3	84 718,1
2013	86 533,1	85 014,7
2014	87 170,2	85 457,9
2015	85 450,6	86 794,1
2016	85 616,1	87 834,2
2017	87 179,3	87 205,4
2018	89 626,6	88 028,8
2019	91 596,7	90 922,9
2020	89 138,9	89 831,0

На следующем этапе построим диаграмму линейного типа с отображением фактических и модельных данных (рис. 3.20).

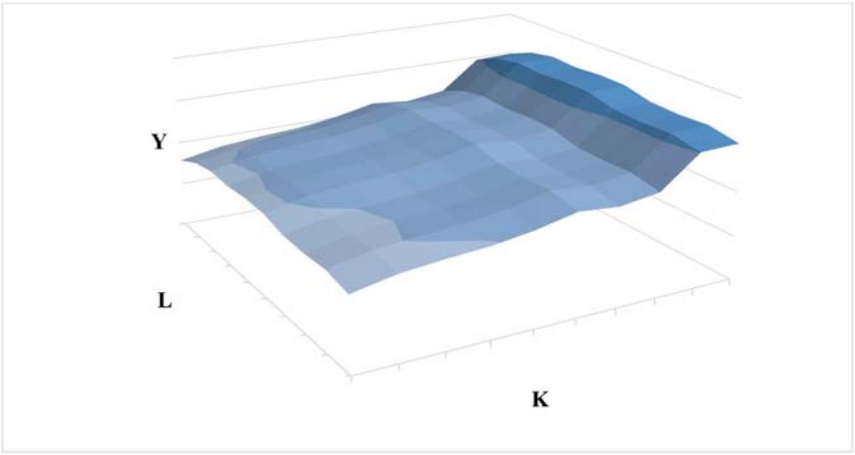


Рис. 3.20. Модель производственной функции Кобба – Дугласа

Рассмотрим процедуру построения полулогарифмической модели с использованием Calc LibreOffice, которая в общем виде выглядит как:

$$Y_i = \alpha + \beta \ln x_i + \varepsilon_i.$$

На первом этапе рассчитаем логарифм переменной x, для этого активируем первую ячейку [в нашем примере D2], и нажмем кнопку f_x. В появившемся окне «Мастера функций» находим в списке функцию LN, выбираем ее и нажимаем кнопку «Далее».

В следующем окне 

выбираем ячейку с переменной x [в нашем примере C2] и нажимаем «ОК». Используя функцию автозаполнения, аналогичную MS Office, заполняем диапазон данных [C2:C12] соответствующими значениями.

На следующем этапе нужно рассчитать неизвестные параметры модели α и β, для этого используем функцию «ЛИНЕЙН». В ячейках [B14] и [B15] будут располагаться, соответственно, значения α и β. Для того чтобы минимизировать выводимую информацию, используем функцию «ИНДЕКС» для извлечения из возвращаемого массива значений необходимые значения. В ячейку [B14] введем выражение «=ИНДЕКС(ЛИНЕЙН(B2:B12;D2:D12;1;0);1;2)», а в ячейку [B15] «=ИНДЕКС(ЛИНЕЙН(B2:B12;D2:D12;1;0);1;1)». Система рассчитает неизвестные значения параметров модели МНК α=-39745,5 и β=7585,1.

В первую ячейку [E2] диапазона ячеек [E2:E12] введем выражение «=$B$14+$B$15*D2». Напомним, что в ячейках [B14] и [B15] располагаются значения α и β. Используя функцию автозаполнения, заполняем диапазон данных [E2:E12] соответствующими значениями. Затем рассчитаем разницу между модельными и фактическими значениями в диапазоне ячеек [F2:F12], то есть из модельных значений [E2:E12] вычитаем [B2:B12]. Для этого введем в ячейку [F2] выражение «=E2-B2» и распространим введенное значение на весь диапазон. Должны получиться следующие данные, представленные в таблице 3.3.

Таблица 3.3. Данные полулогарифмической модели

№ п/п	y	x	ln (x)	Модель	Остаток
	1	**2**	**3**	**4**	**5**
1	4 473	354	5,86929691	4774,00852	301,00852

2	5 134	402	5,99645209	5738,49891	604,49891
3	5 632	413	6,02344759	5943,26371	311,26371
4	6 137	413	6,02344759	5943,26371	-193,73629
5	6 355	418	6,03548143	6034,54213	-320,45787
6	6 688	420	6,04025471	6070,74814	-617,25186
7	6 749	424	6,04973346	6142,64578	-606,35422
8	6 805	461	6,13339804	6777,25377	-27,74623
9	6 840	465	6,14203741	6842,78459	2,78459
10	6 979	488	6,19031541	7208,98020	229,98020
11	7 001	495	6,20455776	7317,01053	316,01053

На последнем этапе построим линейный график[1] с отображением исходных и модельных данных нашей модели. Рассмотрим подробную процедуру построения диаграммы с использованием Calc LibreOffice. Выбираем весь диапазон данных с заголовками [A1:F12], и после выбора пункта меню «Вставка-Диаграмма» или после нажатия кнопки  запустится окно следующего вида (рис. 3.21):

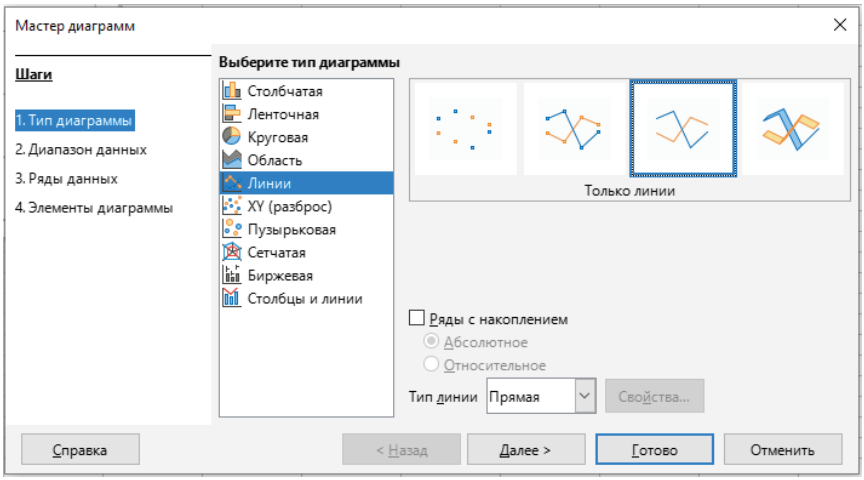


Рис. 3.21. Мастер диаграмм (тип диаграммы)

Выберем тип диаграммы, как указанно на рисунке, и нажмем кнопку «Далее». Появится окно следующего вида (рис. 3.22), где зададим диапазон данных, расположение ряда данных и наличие подписей.

[1] Calc LibreOffice (версия 7.3) не позволяет строить диаграммы поверхностного типа.

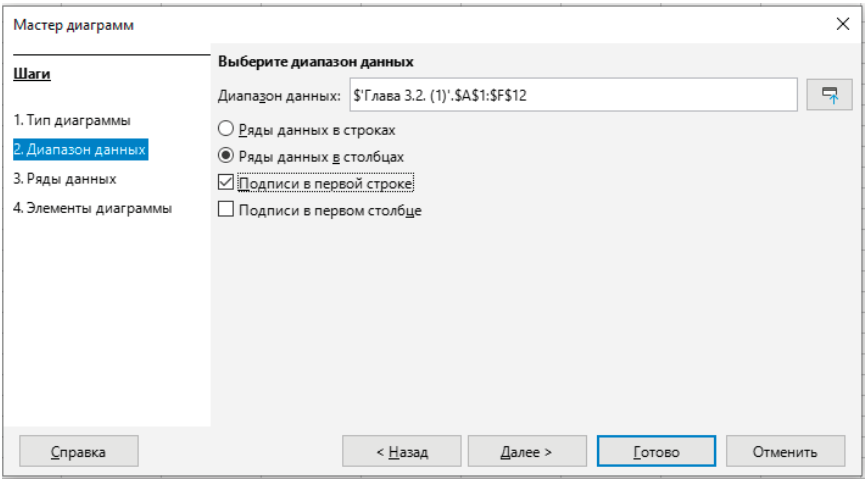


Рис. 3.22. Мастер диаграмм (диапазон данных)

После нажатия кнопки «Далее» появится окно, в котором можно настроить диапазоны данных для каждого ряда данных (рис. 3.23).

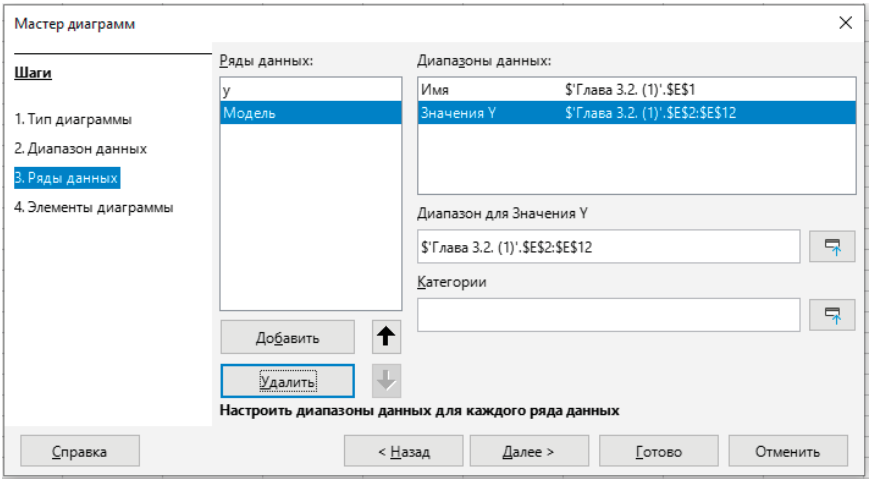


Рис. 3.23. Мастер диаграмм (ряды данных)

В данном окне можно удалить ненужные наборы данных, добавить ряды данных, добавить категории (подписи данных). После нажатия кнопки «Далее» запустится последнее окно «Мастера диаграмм» (рис. 3.24).

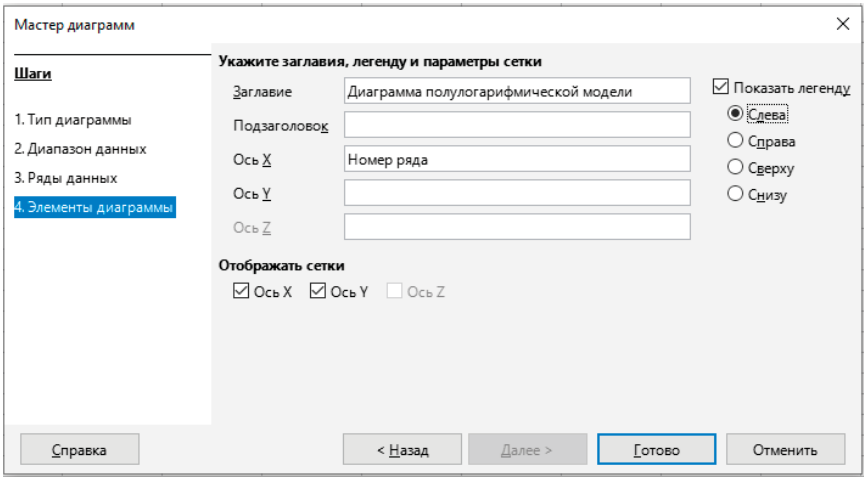


Рис. 3.24. Мастер диаграмм (элементы диаграммы)

Вводим данные, как показано на рисунке, и после нажатия кнопки «Готово» появится диаграмма линейного типа с отображением фактических и модельных данных (рис. 3.25).

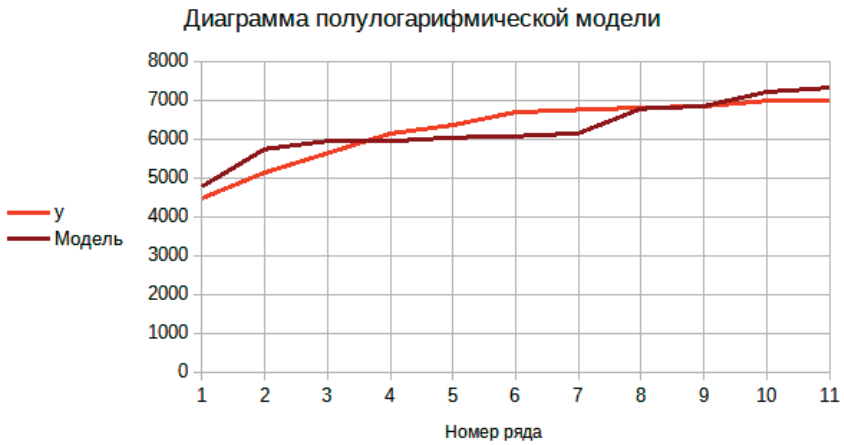


Рис. 3.25. Диаграмма полулогарифмической модели

Аналогичным образом можно построить регрессионные модели других видов. На следующем рисунке представлено сравнение

линейной, степенной и логарифмической моделей с фактическими данными (рис. 3.26).

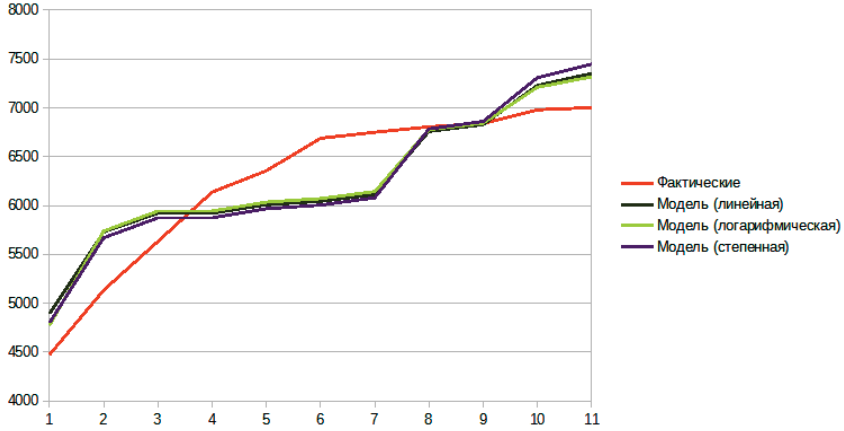


Рис. 3.26. Диаграмма моделей

Далее рассмотрим процедуру построения моделей регрессии со специфическими переменными (переменными, измеряемыми по дихотомической шкале).

3.3. Модели регрессии со специфическими переменными

В классическом регрессионном анализе значения представлены в основном в виде переменных, измеряемых по интервальной шкале. Однако зачастую возникает необходимость применения переменных, измеряемых по дискретной шкале. К данной категории и относятся номинальные переменные, которые могут принимать конечное число значений (например, Тяжесть преступления = {небольшой, средней, тяжкое, особо тяжкое}, Направленность преступления = {общеуголовная, экономическая} или Пол = {мужской, женский}). Причем переменная, принимающая два значения {0, 1}, называется дихотомической.

Модели регрессии с фиктивными переменными

Набор данных, находящихся у аналитика, чаще всего обладает несколькими характерными особенностями:

– различия, возникающие за счет влияния отдельных факторов при неизменном влиянии других;

– структурные различия объясняющих переменных, измеренных по интервальной или порядковой шкале.

Выявить их возможно за счет оценки влияния номинальной переменной на моделируемый показатель путем введения фиктивных переменных непосредственно в математическую модель[1]. Другими словами, они вводятся для исследования структуры данных и характера взаимосвязи переменных с другим уровнем измерения. Так, например, можно ожидать, что динамика преступности будет по-разному проявляться при изменении нормативной правовой базы, действующей в системе учета преступлений, а также варьироваться по регионам с разным уровнем социально-экономического развития либо природно-климатических особенностей.

Модель с фиктивной переменной выглядит следующим образом:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{i=1}^{m} c_i z_i + \varepsilon ,$$

где y – зависимая переменная; x – независимая переменная; z – фиктивная переменная, принимающая два значения [0, 1]. В графическом виде данную модель можно представить в виде следующего рисунка (рис. 3.27):

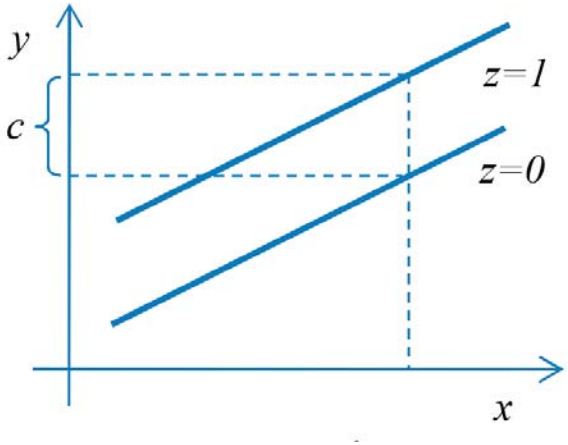


Рис. 3.27. Модель регрессии с фиктивной переменной сдвига

Рассмотрим следующий пример. В качестве независимой переменной (y) будут выступать преступления, совершенные в общественных местах, β_0, β_j, c_i – коэффициенты модели, z_i – фиктивная

[1] *Елисеева И. И., Курышева С. В.* Фиктивные переменные в анализе данных // Социология: 4М. 2010. № 30. С. 43–63.

переменная, а в качестве зависимых переменных (x_i) – численность сотрудников дорожно-патрульной службы полиции. Модель множественной регрессии, учитывающей два регрессора, будет представлена в виде:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z + \varepsilon,$$

где ε – случайная компонента, z – переменная, отражающая период до вступления в силу Федерального закона «О полиции» и после [0 – период с 1997 г. по 2010 г.; 1 – период с 2011 г. по 2019 г.]. Заполненная таблица представлена ниже (рис. 3.28).

	A	B	C	D
1		Совершено в общественных местах	Численность ДПС	Фиктивная переменная
2	1997	4544	1666	0
3	1998	3796	1781	0
4	1999	3588	1742	0
5	2000	3427	1638	0
6	2001	3424	1867	0
7	2002	3797	1602	0
8	2003	3186	1933	0
9	2004	4137	1943	0
10	2005	5062	1995	0
11	2006	5652	1607	0
12	2007	5702	1577	0
13	2008	6164	1649	0
14	2009	6160	1653	0
15	2010	7078	1491	0
16	2011	8122	1658	1
17	2012	8336	1824	1
18	2013	8208	1821	1
19	2014	10860	1765	1
20	2015	11021	1558	1
21	2016	9822	1572	1
22	2017	9000	1561	1
23	2018	9910	1395	1

Рис. 3.28. Таблица данных

Рассчитаем регрессию[1] и получим следующие результаты (рис. 3.29):

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,917636
R-квадрат	0,842056
Нормированный R-квадрат	0,82543
Стандартная ошибка	1092,123
Наблюдения	22

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*значимость F*
Регрессия	2	1,21E+08	60409364	50,64791	2,43E-08
Остаток	19	22661900	1192732		
Итого	21	1,43E+08			

	Коэффициенты	*артная ошибка*	*t-статистика*	*Значение*	*ижние 95%*	*рхние 95%*	*ижние 95,0*	*рхние 95,0*
Y-пересечение	11593,48	2748,236	4,218517	0,000465	5841,357	17345,61	5841,357	17345,61
Численность ДПС	-4,00065	1,584563	-2,52477	0,020633	-7,31718	-0,68412	-7,31718	-0,68412
Фиктивная переменная	4394,466	500,4851	8,780412	4,09E-08	3346,938	5441,993	3346,938	5441,993

Рис. 3.29. Регрессия

Полученное уравнение регрессии представим в виде $y = 11593.5 - 4 \cdot x + 4394.5$ и на ее основе построим прогноз на 2019–2020 гг. (рис. 3.30).

23	2018	9910	1395	1	10408
24	2019		1395	1	=11593,5-4*C24+4394,5*D24
25	2020		1395	1	10408

Рис. 3.30. Прогноз

Полученная диаграмма прогноза представлена на рис. 3.31.

Рис. 3.31. Диаграмма прогноза

[1] См. параграф 3.1. главы 3.

Модели дискретного выбора

В предыдущем разделе была рассмотрена технология построения регрессионных моделей с фиктивными переменными. В случае если дискретные переменные выступают в качестве зависимых, их называют моделям дискретного выбора. Зависимая переменная в них является дискретной, то есть может принимать бинарные и множественные значения.

Одной из наиболее распространенных моделей является Logit-модель, или модель логистической регрессии. Построение модели логистической регрессии в MS Excel включает в себя 7 последовательных этапов:

1. Упорядочение данных. Используя инструмент MS Excel «Сортировка», производится первичная сортировка данных по зависимой переменной (рис. 3.32).

y	x_1	x_2
0	536	116
0	766	230
0	963	102
0	301	470
0	593	980
0	221	824
0	630	133
1	367	149
1	235	177
1	787	539
1	634	207
1	150	392
1	539	733
1	120	342

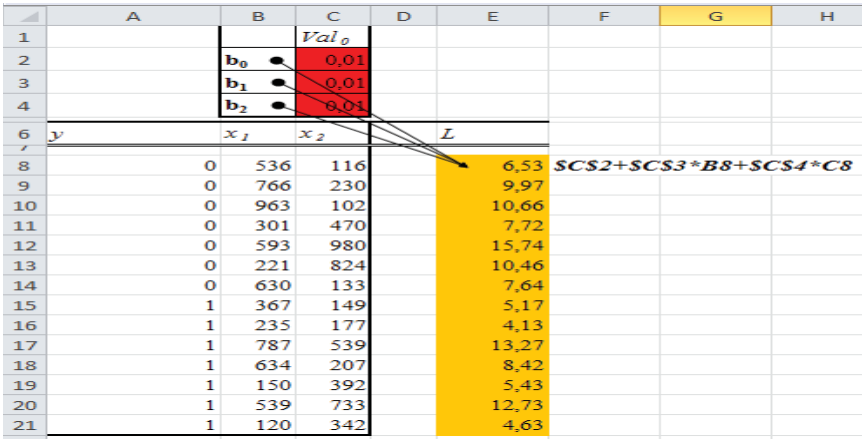
Рис. 3.32. Первичная сортировка данных

2. Расчет значений Logit-модели для набора входящих параметров $x_1,x_2,...,x_k$:

$$LogReg=L=b_0+b_1x_1+ b_2x_2+... +b_kx_k.$$

При помощи встроенного инструмента MS Excel «Поиск решения» оптимизируются коэффициенты $b_0,b_1,b_2,...,b_k$, которые произвольно устанавливаются со значением 0,01 (рис. 3.33).

Для примера рассмотрим процедуру построения Logit-модели с двумя независимыми переменными x_1, x_2.

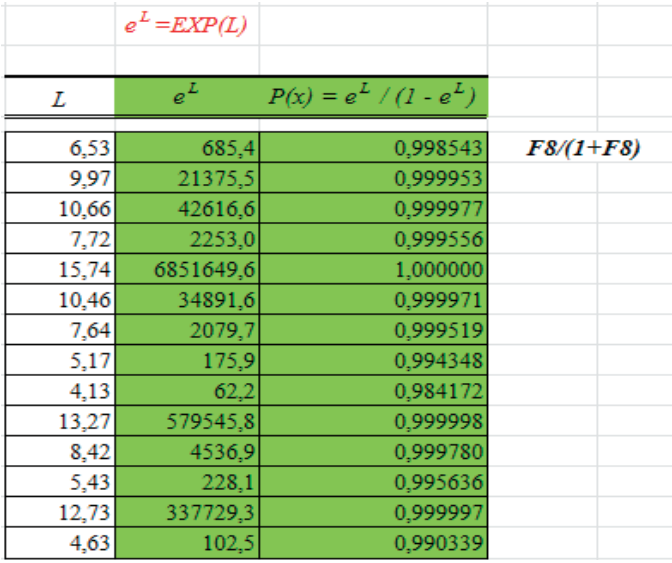


	A	B	C	D	E	F	G	H
1			Val_0					
2		b_0	0,01					
3		b_1	0,01					
4		b_2	0,01					
6	y	x_1	x_2		L			
8	0	536	116		6,53	SCS2+SCS3*B8+SCS4*C8		
9	0	766	230		9,97			
10	0	963	102		10,66			
11	0	301	470		7,72			
12	0	593	980		15,74			
13	0	221	824		10,46			
14	0	630	133		7,64			
15	1	367	149		5,17			
16	1	235	177		4,13			
17	1	787	539		13,27			
18	1	634	207		8,42			
19	1	150	392		5,43			
20	1	539	733		12,73			
21	1	120	342		4,63			

Рис. 3.33. Расчет первичных значений модели

3. Расчет значения e^L. Число e является основанием натурального логарифма и приблизительно равно 2,71828. Данное число рассчитывается для каждого значения данных e^L (рис 3.34).

4. Расчет вероятности события $P(x)$ по формуле $P(x) = \frac{e^L}{1+e^L}$ представлен на следующем рисунке (рис. 3.34):



	e^L =EXP(L)		
L	e^L	$P(x) = e^L / (1 - e^L)$	
6,53	685,4	0,998543	F8/(1+F8)
9,97	21375,5	0,999953	
10,66	42616,6	0,999977	
7,72	2253,0	0,999556	
15,74	6851649,6	1,000000	
10,46	34891,6	0,999971	
7,64	2079,7	0,999519	
5,17	175,9	0,994348	
4,13	62,2	0,984172	
13,27	579545,8	0,999998	
8,42	4536,9	0,999780	
5,43	228,1	0,995636	
12,73	337729,3	0,999997	
4,63	102,5	0,990339	

Рис. 3.34. Расчет значения eL и вероятности события P(x)

5. Расчет функции правдоподобия *PP*. Вероятность $Pr\ (Y_i=y_i|X_{1i}, X_{2i},..., X_{ki})$ – вероятность того, что предсказанная зависимая переменная y_i равна значению Y_i с учетом значений независимых переменных $X_{1i}, X_{2i},..., X_{ki}$. В сокращенном виде данное выражение записывается как $Pr(Y=y|X)$ и рассчитывается по формуле: $Pr\ (Y=y|X)=P(X)^{Y*}[1-P(X)](1-Y)$. Прологарифмировав обе части, получим: $ln[Pr(Y=y|X)]=y*ln[P(X)]*(1-y)*ln[[1-P(X)]]$. Функция правдоподобия (*PP*) представляет собой сумму значений $ln[Pr(Y=y|X)]$ и рассчитывается так:

$$PP = \sum Y_i \cdot P(X_i) + (1 - Y_i)(1 - P(X_i)) \quad .$$

Расчет *PP* в MS Excel представлен на следующем рисунке:

L	e^L	$P(x) = e^L / (1 - e^L)$	$y*ln[P(X)]*(1-y)*ln[[1-P(X)]]$	
6,53	685,3982115	0,99854312	-6,531457943	*A8*LN(G8)+(1-A8)*LN(1-G8)*
9,97	21375,48535	0,99995322	-9,970046781	
10,66	42616,63717	0,999976536	-10,66002346	
7,72	2252,959581	0,999556336	-7,720443762	
15,74	6851649,608	0,999999854	-15,74000015	
10,46	34891,55145	0,999971341	-10,46002866	
7,64	2079,743817	0,999519403	-7,640480713	
5,17	175,9148375	0,994347563	-0,005668473	
4,13	62,17792293	0,984171686	-0,015954919	
13,27	579545,8161	0,999998275	-1,72549E-06	
8,42	4536,903455	0,999779634	-0,00022039	
5,43	228,1492454	0,995636032	-0,004373518	
12,73	337729,3115	0,999997039	-2,96095E-06	
4,63	102,5140641	0,990339477	-0,009707489	
		СУММА	-68,75841095	*СУММ(H8:H21)*

Рис. 3.35. Расчет функции правдоподобия

6. Расчет функции максимального правдоподобия. Используется «Поиск решения» MS Excel для нахождения коэффициентов $b_0,b_1,b_2,...,b_k$, который максимизирует функцию PP в ячейке «СУММА». Данный инструмент настраивает числа в определенных ячейках, в которых находятся значения коэффициентов $b_0,b_1,b_2,...,b_k$ для оптимизации (максимизации или минимизации) целевой функции. Эти ячейки будут скорректированы таким образом, чтобы максимизировать *PP*, который находится в ячейке «СУММА» (рис. 3.36). Для оптимизации целевой функции используем «Поиск решения» MS Excel для нелинейных и гладких задач методом обобщенного приведенного градиента (ОПГ).

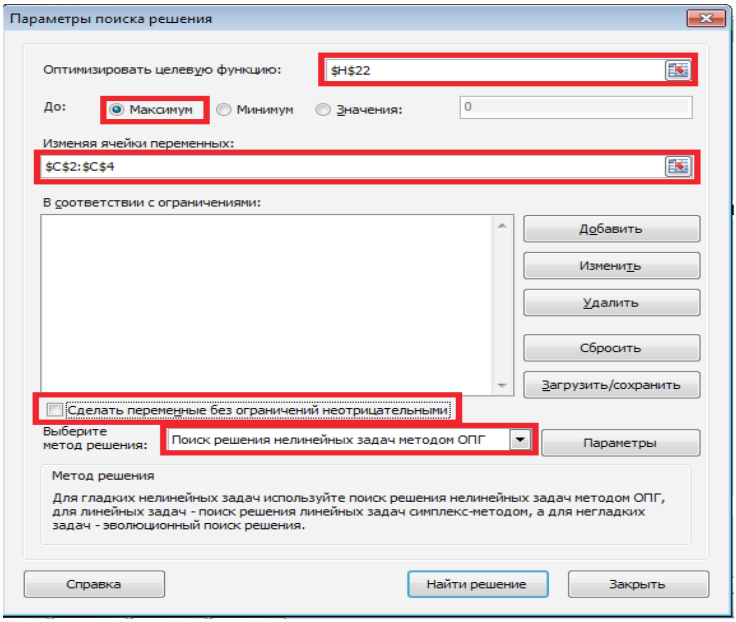


Рис. 3.36. Оптимизация целевой функции

Результаты работы приведены на следующем рисунке (рис. 3.37).

	A	B	C	D	E	F	G	H
1			Val_0					
2		b_0	0,009984089					
3		b_1	-0,000847362					
4		b_2	0,000443785					
5								
6	y	x_1	x_2		L	e^L	$P(x) = e^L / (1 - e^L)$	$y*\ln[P(X)]+(1-y)*\ln[1-P(X)]$
8	0	536	116		-0,39272	0,675215698	0,403061945	-0,515941932
9	0	766	230		-0,53702	0,584484529	0,368879922	-0,460259136
10	0	963	102		-0,76076	0,46731124	0,318481333	-0,383431637
11	0	301	470		-0,03649	0,964164672	0,490877718	-0,675067052
12	0	593	980		-0,05759	0,944034283	0,485605779	-0,664765341
13	0	221	824		0,188396	1,207311003	0,54696008	-0,791775034
14	0	630	133		-0,46483	0,628241392	0,38584045	-0,487500532
15	1	367	149		-0,23487	0,790670497	0,441549966	-0,817464092
16	1	235	177		-0,1106	0,895300226	0,472379106	-0,749973425
17	1	787	539		-0,41769	0,658566275	0,397069617	-0,923643657
18	1	634	207		-0,43538	0,647018631	0,392842326	-0,934346952
19	1	150	392		0,056843	1,058489961	0,514207007	-0,665129358
20	1	539	733		-0,12145	0,885635297	0,469674755	-0,755714834
21	1	120	342		0,060075	1,061916149	0,515014226	-0,663560756
22							СУММА	-9,488573738

Рис. 3.37. Результаты оптимизации

Значения функции максимального правдоподобия и коэффициентов $b_0, b_1, ..., b_k$ равны: $MPP=-9,488$; $b_0=0,009$; $b_1=-0,000847362$; $b_2=0,000443785$.

7. Тестирование результатов (рис. 3.38).

	Coeff
b_0	0,009984089
b_1	-0,000847362
b_2	0,000443785

x_1	161
x_2	949
$L=$	0,294710395
$P(x)=$	57%

Рис. 3.38. Тестирование произвольных данных

Задавая произвольные значения x_1 и x_2, получаем соответствующее значение $P(x)$.

Глава 4. Методы анализа анкетных данных

В главе 4 рассматриваются особенности формирования анкет в процессе исследования и методика создания анкет с учетом особенностей их подачи в электронном виде, а также технология создания сводных таблиц. Рассматриваются методы использования компьютерных технологий для анализа результатов анкетирования[1].

4.1. Методика сбора результатов анкетных опросов

Первичная компьютерная обработка результатов опроса – это последовательное заполнение каждого документа (анкеты, формы и т. д.), заполнение всех ответов на все вопросы в единую матрицу. Результатом являются возможность получения данных (как в абсолютных числах, так и в процентах) для каждой позиции, в каждой группе и для всей таблицы (обычно это называется линейным распределением); данные о связях между определенными ответами и той или иной парой (произвольной) или даже несколькими вопросами (эти данные принято называть корреляциями); разные типы коэффициентов и др.

При редактировании исследовательских работ вручную особенно важно знать (и лучше подумать об этом при создании исследовательской программы), нужна ли Вам просто информация о том, как все респонденты ответят на заданный вопрос, или Вас интересуют подробные ответы от представители той или иной группы.

Однако логика поиска часто требует дальнейшего исследования, установления связи между двумя или более характеристиками респондента. В общем случае мы говорим о том, что именно зависит от появления или распространения такого-то ответа на такой-то вопрос. Практика показывает, что определяющая и объяснительная характеристика в основном социально-демографическая – возраст, пол, образование, место работы или учебы, трудовой стаж и т. д.

На первом этапе исследования необходимо выбрать метод сбора информации от респондентов, наиболее часто используемым является анкетирование. Для этого этапа можно использовать открытые инструменты Yandex Forms или Google Forms. Рассмотрим процедуру создания анкеты при помощи Yandex Forms. Переходим

[1] Математические методы исследования социальных систем: курс лекций / И. В. Горошко, Б. А. Торопов, Ш. Х. Гонов. Москва: Академия управления МВД России, 2019. 80 с.

по ссылке (https://forms.yandex.ru/)[1], на появившейся странице мы можем создать форму из шаблонов (рис. 4.1) либо выбираем пункт «Создать форму» и создаем форму без заданных полей.

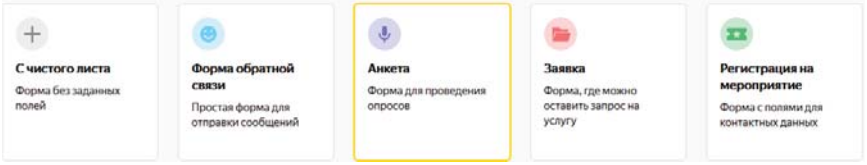


Рис. 4.1. Выбор шаблона

После создания формы в меню Yandex Forms имеется несколько разделов. Раздел «Конструктор» позволяет создавать, редактировать анкеты (рис. 4.2).

Рис. 4.2. Настройки теста

В разделе «Темы» можно выбирать как готовую тему, так и создавать свою, выбирая цветовую схему, шрифты, отображение текста и т. п. В разделе «Интеграция» можно задавать действие (перечень действий), выполняющееся всегда или при выполнении определенного условия (группы условий). В разделе «Настройки» можно задавать параметры, относящиеся ко всей форме. Например, задавать события после отправки или ограничить время доступа к форме (рис. 4.3).

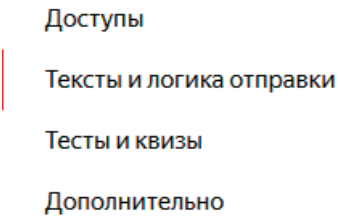


Рис. 4.3. Настройки формы

Подчеркнем, что, на наш взгляд, цель анкетирования заключается, прежде всего, в получении нового знания, установлении скрытых взаимосвязей в изучаемом процессе, объекте или явлении, взаимосвя-

[1] Для работы с опросными формами потребуется аккаунт в поисковых системах «Yandex»или «Google».

зей, которые, несомненно, должны содержаться в ответах на вопросы, сформулированные в анкете[1].

Поставленная цель достигается, если в основе проводимого анкетирования лежит определенная технология, базирующаяся на общих и специальных принципах. Общие принципы характерны для любого вида социологического исследования, в том числе и прикладного, специальные – относятся только к анкетированию.

Общими принципами опросных методов социологического исследования предлагаем считать принципы квантификации, репрезентативности и необходимой верификации[2].

Квантификация предполагает обязательное преобразование результатов опросов в цифровые данные для анализа и сравнения.

Принцип репрезентативности означает обязательность сопоставимости характеристик выборки (группы опрашиваемых респондентов) и характеристик изучаемой генеральной совокупности (например, определенной категории прокурорских работников).

Соблюдение принципа необходимой верификации обеспечивается в ходе итерационного проведения процедуры анкетирования, на каждом этапе которой возможна корректировка вопросов, уточнение параметров выборочной совокупности, переформулирование гипотез в зависимости от полученных результатов.

Что касается специальных принципов, то к ним относятся:

– обязательный учет специфических особенностей опрашиваемой аудитории (возраста, стажа работы, уровня образования, направленности деятельности и т. п.);

– обеспечение точности и корректности формулируемых вопросов, логической стройности и непротиворечивости построения анкеты;

– обязательное использование уточняющих и контрольных вопросов, ответы на которые должны подтверждать ответственное отношение опрашиваемого к заполнению анкеты;

– обеспечение толерантности сформулированных вопросов: респонденты, понимая общую постановку задачи, не должны догадываться о собственной позиции анкетера;

– чередование уровня сложности вопросов с постепенным его повышением;

---

[1] *Горошко И. В.* Технология проведения анкетных опросов // Вестник Университета прокуратуры Российской Федерации. № 3 (77). 2020. С.72–77.

[2] *Ядов В. А.* Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности. Москва: Омега-Л, 2007. 567 с.

– разумное ограничение объема анкеты: заполнение анкеты и нахождение ответов на поставленные в ней вопросы не должны казаться утомительной процедурой и отнимать слишком много времени.

Анкета чаще всего начинается с вступительного обращения к респондентам. Здесь указываются цель исследования, выходные результаты, общая методика опроса и т. п. В качестве примера ниже приведен фрагмент анкеты, разработанной авторским коллективом Академии управления МВД России в рамках научно-исследовательской работы по теме: «Разработка проекта концепции использования искусственного интеллекта в деятельности подразделений МВД России».

Уважаемые коллеги!

Авторский коллектив Академии управления МВД России в рамках научно-исследовательской работы по теме: «Разработка проекта концепции использования искусственного интеллекта в деятельности подразделений МВД России» проводит опрос сотрудников МВД России. Нам важно знать Ваше мнение по ряду вопросов...

1. Укажите Ваш пол.

– мужской ☐

– женский ☐

2. К какой возрастной группе Вы относитесь?

– Менее 28 лет ☐

– 29–38 лет ☐

– 39– 48 лет ☐

– Более 48 лет ☐

5. Насколько, на Ваш взгляд, целесообразно внедрение систем искусственного интеллекта в различные направления деятельности органов внутренних дел? (оцените каждый вариант по шкале от 1 до 5, где 1 – абсолютно нецелесообразно, 5 – необходимо)

– Расследование преступлений ☐

– Оперативно-розыскная деятельность ☐

– Экспертно-криминалистическая деятельность ☐

– Охрана общественного порядка ☐

– Организационно-аналитическая деятельность

– Информационно-аналитическая деятельность

– Кадровая работа

– Материально-техническое обеспечение

– Другое (укажите)

6. С каким родом деятельности связано Ваше место службы?

– Расследование преступлений

– Оперативно-розыскная деятельность

– Экспертно-криминалистическая деятельность

– Охрана общественного порядка

– Организационно-аналитическая деятельность

– Информационно-аналитическая деятельность

– Кадровая работа

– Материально-техническое обеспечение

– Другое (укажите)

9. На Ваш взгляд, готовы ли сотрудники и работники подразделения, где Вы проходите службу, по своей подготовке и квалификации для использования систем искусственного интеллекта в своей оперативно-служебной деятельности?

Оцените по шкале от 1 до 5, где 1 – абсолютно не готовы, а 5 – полностью готовы

Рассмотрим технологию создания указанной выше анкеты при помощи Yandex Forms. Вначале добавим вопросы. Для этого выбираем пункт «Один вариант»[1] (рис. 4.4).

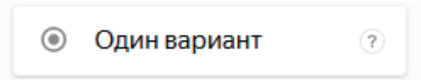


Рис. 4.4. Выбор варианта

[1] Данный элемент пользовательского интерфейса называется переключателем, или радиокнопкой (*англ. RadioButton*).

В появившейся форме (рис. 4.5) вводим информацию о первом вопросе и вариантах ответа.

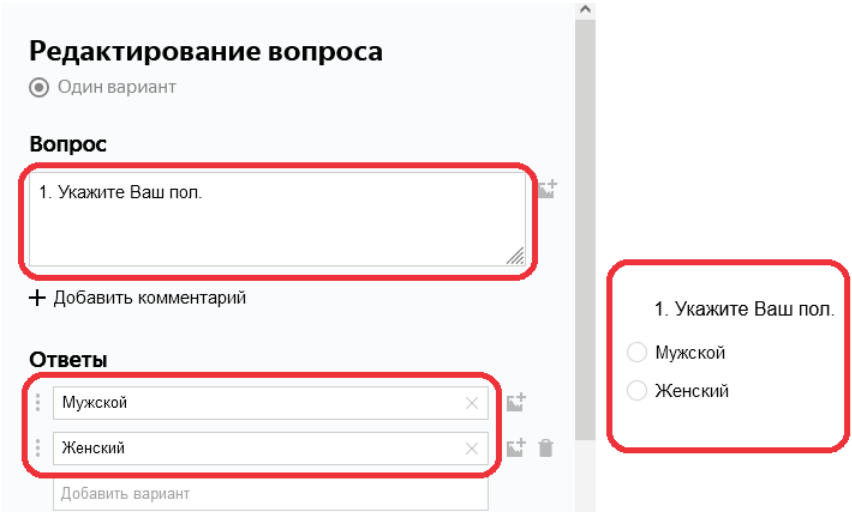


Рис. 4.5. Редактирование вопроса

Ниже в форме (рис. 4.6) можно задать дополнительные настройки вопроса (обязательность[1], скрытость вопроса, варианты сортировки ответов).

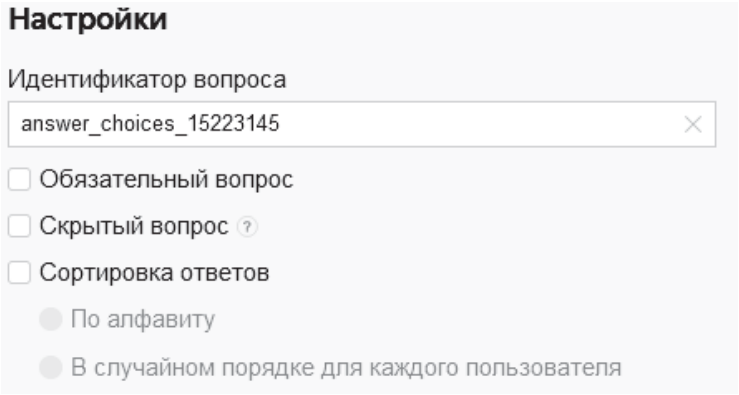


Рис. 4.6. Настройки вопроса

[1] При включении данной опции рядом с вопросом появляется красная звездочка, которая и сигнализирует об обязательности ответа на данный вопрос.

После нажатия кнопки «Сохранить» вопрос добавляется на страницу (рис. 4.7), в которой можно: 1) перемещать; 2) копировать; 3) удалять и 4) создавать условия показа.

Рис. 4.7. Добавление вопроса

Рассмотрим, как создаются условия показа вопросов. Например, следующий вопрос: «К какой возрастной группе Вы относитесь?», не задается респондентам, выбравшим в первом вопросе ответ «Женский». После нажатия кнопки 4) активируется следующее окно (рис. 4.8), в котором активируем пункт «При условии», указываем вопрос, математический оператор (равно, не равно) и варианты вопроса. Теперь если респондент выберет вариант ответа «Женский», то вопрос о возрасте будет ему недоступен.

Для обработки вопроса № 5 анкеты нам потребуется вариант «Оценка по шкале» (рис. 4.9), в котором задаем критерии и ответы (рис. 4.10).

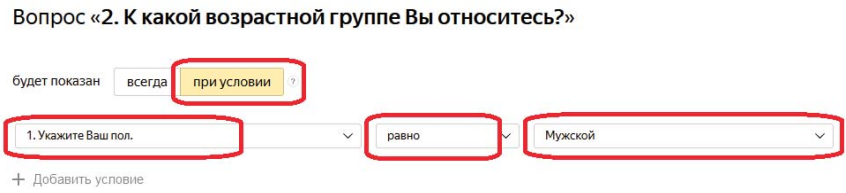


Рис. 4.8. Условия показа

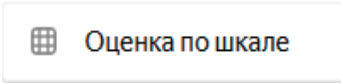


Рис. 4.9. Оценка по шкале

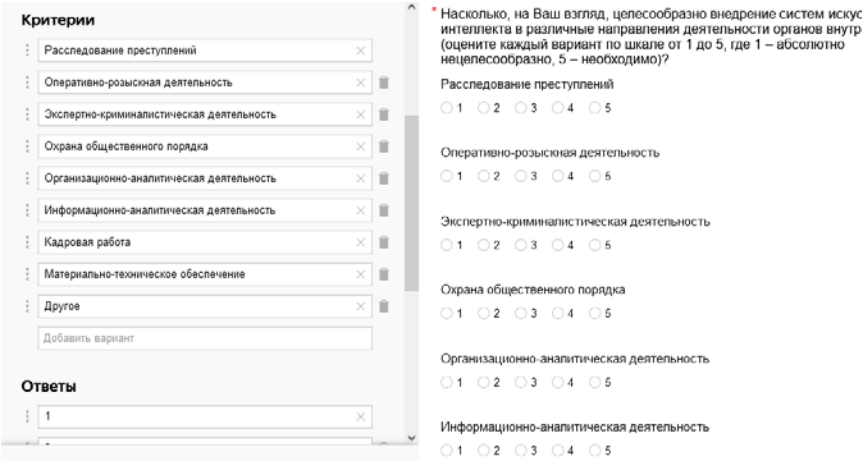


Рис. 4.10. Оценка по шкале

Следующий вопрос анкеты (№ 6) предполагает множественный выбор ответов[1] (рис. 4.11).

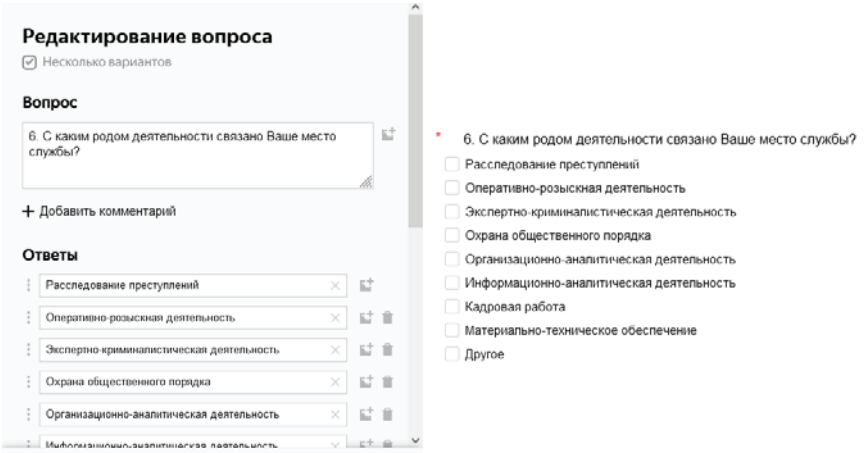


Рис. 4.11. Множественный выбор

Для вопроса № 11 анкеты указываем вариант «Длинный текст» или «Короткий текст».

После того, как анкета создана, можно посмотреть, как она выглядит за счет кнопки «Предпросмотр», и нажатием кнопки

[1] Данный элемент пользовательского интерфейса называется флажком, галочкой или чекбоксом *(англ. CheckBox)*.

«Публикация» дать к ней доступ для заполнения респондентами[1]. В окне, которое появляется после (рис. 4.12), можно поделиться ссылкой, отправить через соцсети и т. д., также можно снять форму с публикации.

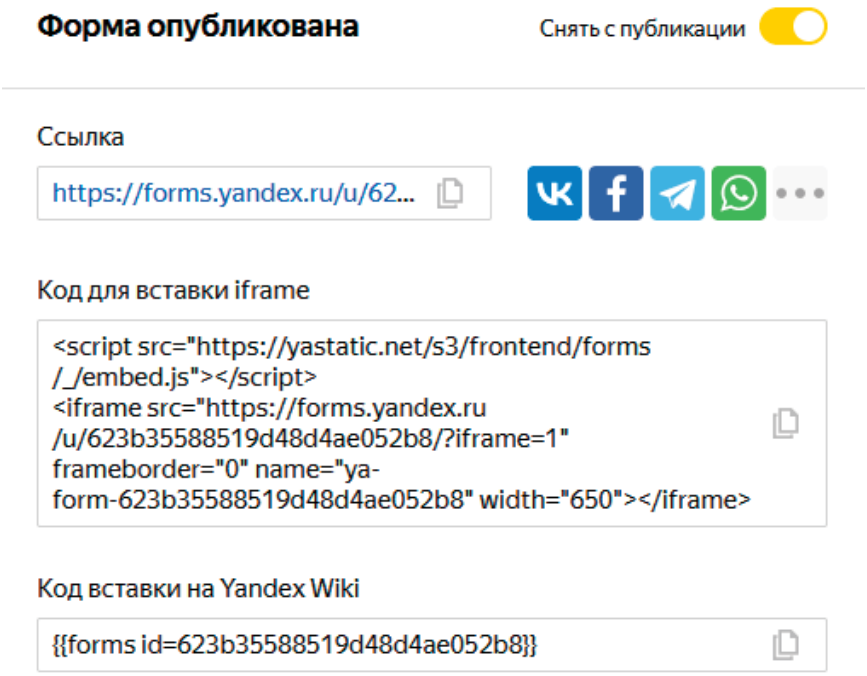


Рис. 4.12. Публикация формы

После публикации форма становится доступной для заполнения. После завершения процедуры сбора анкет автор может просматривать и обрабатывать результаты в разделе «Ответы», где будет указано количество поступивших ответов, а ниже в подменю «По ответам» можно просмотреть каждый ответ по отдельности. Более расширенные возможности предоставляет подменю «Сводка», где можно обрабатывать агрегированные данные: указывать интересующие вопросы, скачивать в разных форматах (xlsx, csv, json), фильтровать по дате ответа, а также включать дополнительные реквизиты в ответ (даты создания, обновления и т. п.). Кроме этого, в дан-

[1] Ссылка для созданной формы: https://forms.yandex.ru/u/623b35588519d48d4ae052b8/.

ном пункте можно просмотреть визуальные данные о статистике ответов (рис. 4.13).

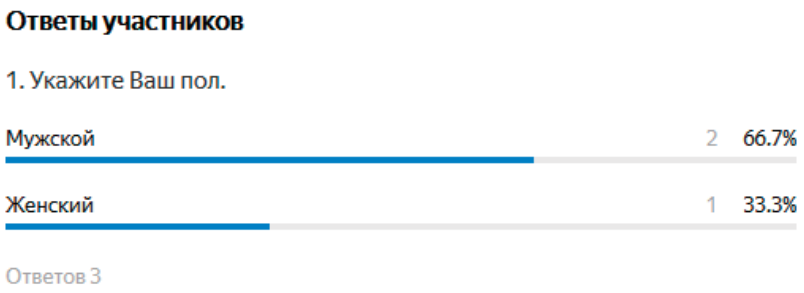


Рис. 4.13. Статистика ответов

Выбрав формат (xlsx) и нажав кнопку «Скачать», мы получим файл с ответами, который в дальнейшем можно обработать. Ответы в файле представлены в виде текстовых значений, которые нужно преобразовать в числовые значения, соответствующие номеру варианта[1].

Для этого создаем в файле лист («Описание»), в котором будет располагаться анкета с кодами ответов (рис. 4.14), также переименуем лист с ответами в («Данные»).

	A	B
1	**1. К какой возрастной группе Вы относитесь?**	
2	Мужской	**1**
3	Женский	**2**
5	**2. К какой возрастной группе Вы относитесь?**	
6	Менее 28 лет	**1**
7	29 - 38 лет	**2**
8	39 - 48 лет	**3**
9	Более 48 лет	**4**
11	**3. К какой группе Вы относитесь по стажу службы в органах внутренних дел?**	
12	Менее 5 лет	**1**
13	5 – 10 лет	**2**
14	11 – 15 лет	**3**
15	16-20 лет	**4**
16	Более 20 лет	**5**

Рис. 4.14. Фрагмент описания анкеты

[1] Для расчета корреляций текстовые значения не подходят, их нужно преобразовать в числовые.

Причем варианты должны строго соответствовать ответам, указанным в самой форме.

Создадим еще один лист, на котором и будут располагаться обработанные данные («Сборка»). В ячейку [A2] на листе «Сборка» введем следующее выражение: (=ВПР(Данные!A2;Описание!$A$2:$B$3;2;ЛОЖЬ)). Функция ВПР позволяет найти искомое значение [A2] в крайнем левом столбце таблицы [Описание.A2:B5] и возвращает значение, находящееся во [2-й] ячейке. Аналогичным образом вводим выражение для второго вопроса (=ВПР(Данные!B2;Описание!$A$6:$B$9;2;ЛОЖЬ)) и т. д.

4.2. Технология обработки результатов анкетных опросов

Корреляция (*от лат. correlatio* – отношение) или корреляционная зависимость – это статистическая связь двух или более случайных величин (или величин, которые могут рассматриваться как таковые с некоторой приемлемой степенью точности). В этом случае изменения значений одной или нескольких из этих переменных сопровождаются систематическим изменением значений той или иной переменной.

Выбор респондентами определенных ответов на разные вопросы также может быть взаимозависимым, то есть коррелированным. Числовая характеристика такой корреляции выражается коэффициентом корреляции (непараметрический коэффициент корреляции Фехнера, коэффициент ранговой корреляции Кендалла и Спирмена).

Для проверки статистических гипотез Карл Пирсон в 1900 г. предложил простой, универсальный и эффективный способ проверки соответствия между предсказаниями модели и экспериментальными данными. Его критерий хи-квадрат является наиболее важным и широко используемым статистическим тестом. С его помощью можно решить большинство проблем, связанных с оценкой неизвестных параметров модели и проверкой согласованности модели и экспериментальных данных.

Рассмотрим практическую реализацию авторской технологии обработки результатов анкетных опросов с использованием инструментария MS Excel.

Имеются данные о результатах анкетного опроса 343 респондентов по анкете[1], состоящей из трех вопросов: Предпочтение внедорожников? («Да»|«Нет»); Размер семьи («Больше 2 детей»|«Не больше 2

[1] Данные взяты с сайта Финансового университета при Правительстве Российской Федерации (http://www.fa.ru/org/dep/findata/Pages/Stat-book.aspx). Анализ данных

детей»); Доходы («Высокие»|«Низкие»). Необходимо выяснить, имеется ли взаимообусловленность между ответами респондентов:

а) на вопрос о количестве детей и на вопрос о предпочитаемом автомобиле;

б) на вопросы о количестве детей и о доходах.

На первом этапе формулируется нулевая гипотеза (H0) о том, что ответы респондентов не взаимообусловлены между собой. Также формулируется альтернативная гипотеза (H1) о том, что имеется взаимообусловленность между ответами респондентов. Задача исследования заключается в том, чтобы подтвердить или опровергнуть гипотезу.

На следующем этапе рассчитаем описательную статистику. Удобнее всего это сделать при помощи функций MS Excel СЧЁТ или в случае подсчета непустых записей СЧЁТЗ. Пример расчета и соответствующая формула представлены в таблице.

Таблица 4.1. Пример расчета описательной статистики

ВСЕГО	343	СЧЁТЗ($A$2:$A$344)
Предпочитают внедорожники		
- Да	148	СЧЁТЕСЛИ($A$2:$A$344;"=Да")
- Нет	195	СЧЁТЕСЛИ($A$2:$A$344;"=Нет")
Размер семьи		
- Больше 2 детей	186	СЧЁТЕСЛИ($B$2:$B$344;"=Больше 2 детей")
- Меньше 2 детей	157	СЧЁТЕСЛИ($B$2:$B$344;"=Не больше 2 детей")
Доходы		
- Высокие	246	СЧЁТЕСЛИ($C$2:$C$344;"=Высокие")
- Низкие	97	СЧЁТЕСЛИ($C$2:$C$344;"=Низкие")

Однако проще и быстрее реализовать данную функцию через инструмент MS Excel «Сводная таблица». Для этого открываем меню «Вставка» и выбираем элемент «Сводная таблица». Появляется следующее окошко (рис. 4.15), в котором выбираем «Таблица или диапазон»[1] и указываем, куда нужно поместить отчет сводной таблицы. Это может быть или новый лист (установлен по умолча-

в экономике: теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel: учебник / В. И. Соловьев. Москва: КноРус, 2018. 500 с.

[1] Чаще всего MS Excel правильно определяет диапазон данных, который необходимо включить в сводную таблицу, но иногда лучше выделить необходимый диапазон самостоятельно.

нию), или существующий лист, в этом случае необходимо указать диапазон ячеек, куда будет помещен отчет.

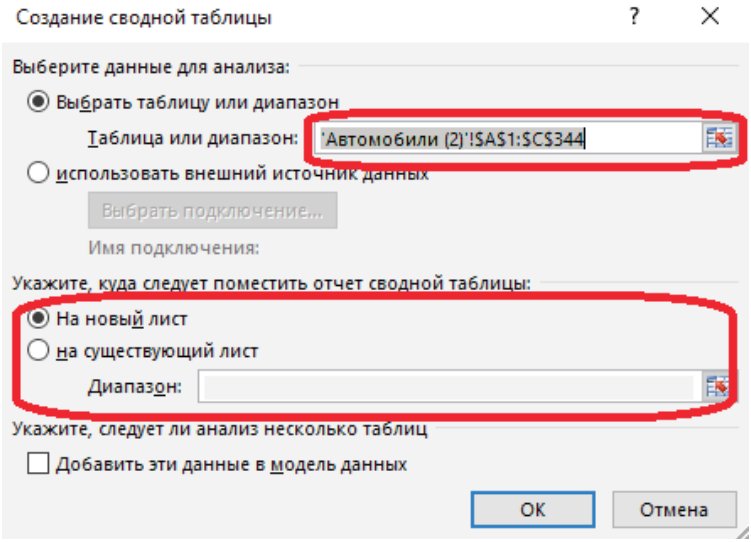


Рис. 4.15. Создание сводной таблицы MS Excel

Для удобства работы выберем ячейку на том же листе, на котором и находятся наши исходные данные. После нажатия кнопки «ОК» MS Excel поместит отчет сводной таблицы на лист с результатами анкетных опросов[1].

Для построения отчета нужно выбрать поля из списка полей сводной таблицы, который появляется справа (рис. 4.16).

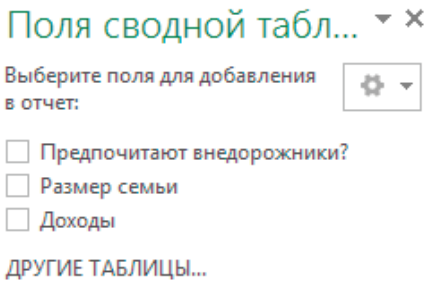


Рис. 4.16. Поля сводной таблицы

[1] На этом этапе отчет сводной таблицы не содержит данных.

Для этого нужное поле «перетаскиваем»[1] мышью в соответствующую область. Например, поле «Размер семьи» «перетаскиваем» в область «Строки», а поле «Предпочитают внедорожники» в область «Колонны». Кроме этого нам нужно указать и итоговые значения. Для этого одно из выбранных полей «перетаскиваем» в область «∑ Значения»[2], например, поле «Размер семьи», как это представлено на следующем рисунке (рис. 4.17).

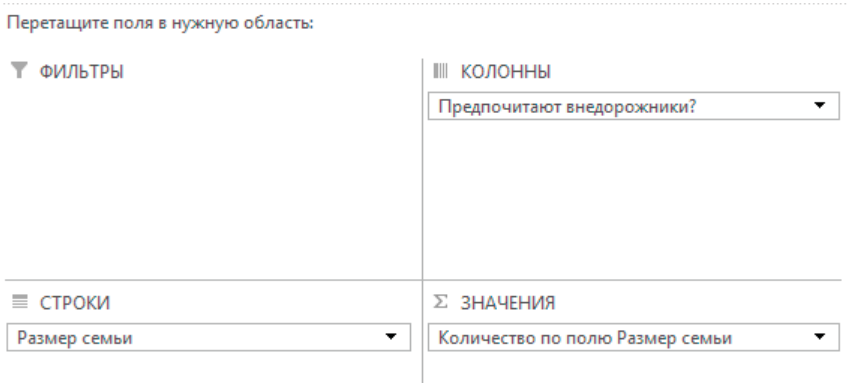


Рис. 4.17. Итоговые значения

На листе MS Excel появится таблица следующего вида (рис. 4.18):

Количество по полю Размер семьи	Названия столбцов		
Названия строк	Да	Нет	Общий итог
Больше 2 детей	138	48	186
Не больше 2 детей	10	147	157
Общий итог	**148**	**195**	**343**

Рис. 4.18. Сводная таблица

Для удобства дальнейших расчетов скопируем данные в отдельные ячейки листа. Примечание: вставлять нужно только значения[3], в противном случае скопируется макет сводной таблицы. Таким образом, у нас получается таблица с эмпирическим (наблюдаемым) распределением ответов.

[1] Более корректное выражение на английском языке звучит как «Drag&Drop» (бери-и-брось).

[2] Знак ∑ в математике означает сумму.

[3] Вставка – только значения.

На следующем этапе рассчитаем ожидаемые частоты (теоретические частоты), то есть теоретическое распределение ответов при полной независимости ответов на два вопроса. Для расчета теоретических частот можно использовать следующее выражение:

$$\sum \text{Col} * \sum \text{Row} / \sum \text{All},$$

где ∑ Col – количество по колонке, ∑ Row – количество по строке, ∑ All – всего анкет (респондентов).

Для нашего примера: Количество предпочитающих внедорожники*Количество «Больше 2 детей» / Общее количество респондентов. Подставляя полученные значения, получаем: 148*186/343.

Данные расчеты проще реализовать в MS Excel, где введенная формула будет представлена в следующем виде (рис. 4.19):

Наблюдаемые частоты			
	Да	Нет	Общий итог
Больше 2 детей	138	48	186
Не больше 2 детей	10	147	157
Общий итог	148	195	343
Ожидаемые частоты			
	Да	Нет	
Больше 2 детей	=G23*I21/I23		
Не больше 2 детей			

Рис. 4.19. Расчет ожидаемых частот

Введенное в ячейку выражение лучше написать следующим образом: =G$23*$I21/$I$23. То есть зафиксировать ячейки для того, чтобы можно было использовать функцию автозаполнения.

Далее рассчитаем наблюдаемое значение Хи-квадрат Пирсона по следующей формуле:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4.1),$$

где χ^2 – Хи-квадрат, O_i – наблюдаемые частоты, E_i – ожидаемые частоты.

Данный расчет реализуем в MS Excel. Введенное выражение будет выглядеть следующим образом (рис. 4.20):

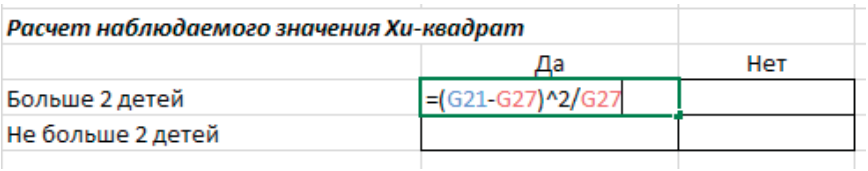


Рис. 4.20. Расчет наблюдаемого значения Хи-квадрат

Сумма полученных результатов и будет являться наблюдаемым значением Хи-квадрат (рис. 4.21).

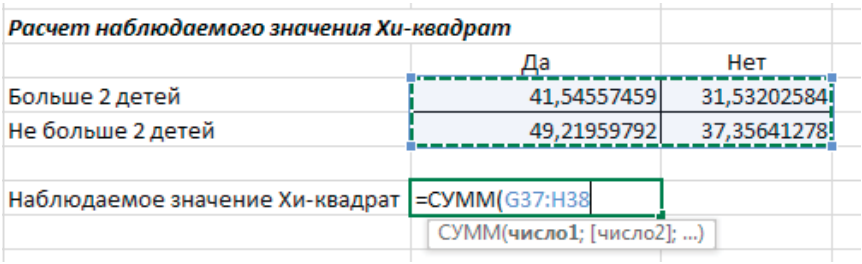


Рис. 4.21. Хи-квадрат

На последнем этапе сравнивается наблюдаемое значение χ^2 с критическим, для расчета которого используется функция MS Excel ХИ2.РАСП.ПХ, которая возвращает правостороннюю вероятность распределения χ^2. Данная функция принимает два аргумента, каждый из которых является обязательным:

1) уровень значимости. Для гуманитарных исследований обычно принимается равным 0,05;

2) число степеней свободы рассчитывается как df=(Row-1)*(Col-1).

Для нашего примера первый аргумент равняется 0,05, а второй – (2-1)*(2-1)=1. Таким образом, ХИ2.РАСП.ПХ(0,05;1) (рис. 4.22).

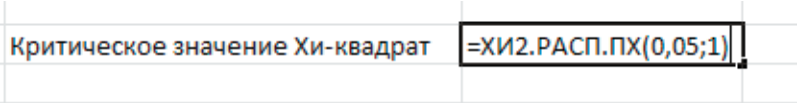


Рис. 4.22. Хи-квадрат

Так как критическое значение Хи-квадрат меньше, чем расчетное, нулевая гипотеза о невзаимозависимости ответов отвергается.

Рассмотрим процедуру создания сводной таблицы средствами LibreOffice Calc. Она очень похожа на методику, рассмотренную ранее с помощью MS Excel. Здесь также выбирается диапазон данных и пункт меню «Вставка» – «Сводная таблица» (рис. 4.23), но указывать на данном этапе, куда поместить отчет сводной таблицы, нельзя.

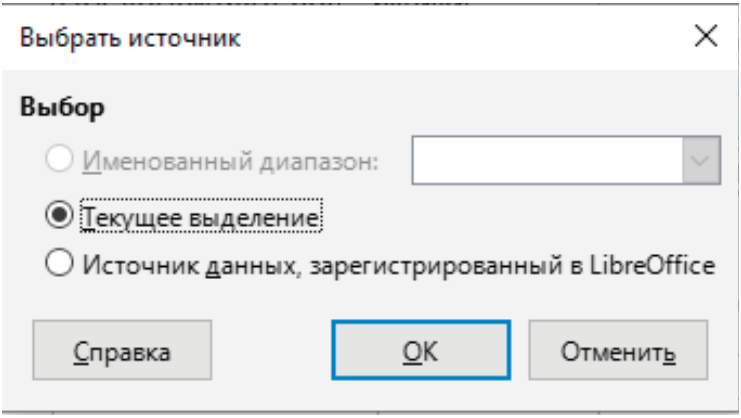


Рис. 4.23 Создание сводной таблицы LibreOffice Calc

После нажатия кнопки «ОК» активируется следующая форма (рис. 4.24). Для построения отчета также нужно выбрать поля из списка полей сводной таблицы, который находится справа. На рис. 4.24 номерами обозначены примерные этапы:

1) доступное поле «Предпочитают внедорожники» разместим в полях столбцов;

2) поле «Размер семьи» разместим в полях строк;

3) одно из доступных полей (в нашем примере – «Размер семьи») разместим в полях данных. По умолчанию в данном поле активна функция «Сумма». Это значит, что в итогах посчитается сумма значений данного поля. Для того чтобы переключиться на другую функцию, на поле данных на значении «Сумма – размер семьи» производим двойной щелчок мышью[1]. Активируется окно с дополнительными настройками для поля данных (рис. 4.25), в котором выбираем функцию «Количество»;

[1] Или при активном поле нажать кнопку «Enter».

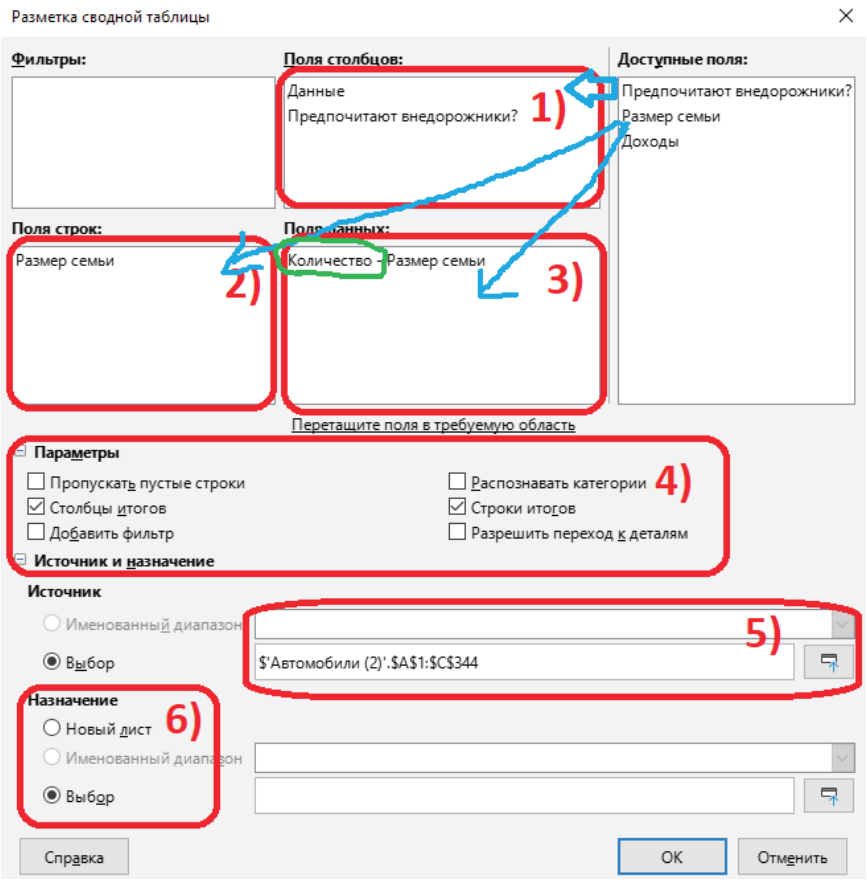


Рис. 4.24. Разметка сводной таблицы LibreOffice Calc

4) после нажатия на значок «+» в пункте «Параметры» активируются дополнительные параметры, в которых можно установить «птички» напротив нужных пунктов;

5) после нажатия на значок «+» в пункте «Источник» активируются параметры выбора источника данных. Здесь можно поменять источник;

6) кроме этого, можно задать параметры назначения, то есть куда будут выводиться результаты (сводная таблица). Есть возможность вывести на новый лист, именованный диапазон[1] и определенный диапазон (пункт «Выбор»).

[1] Здесь не рассматривается.

После нажатия кнопки «ОК» система разместит в указанном в п. 6 месте сводную таблицу следующего вида (таблица 4.2):

Таблица 4.2. Сводная таблица

Количество и размер семьи	Данные		
Размер семьи	Да	Нет	**Итог Результат**
Больше 2 детей	138	48	**186**
Не больше 2 детей	10	147	**157**
Итог Результат	**148**	**195**	**343**

Дальнейшая процедура расчета идентична расчету с использованием MS Excel, и на ней мы останавливаться не будем.

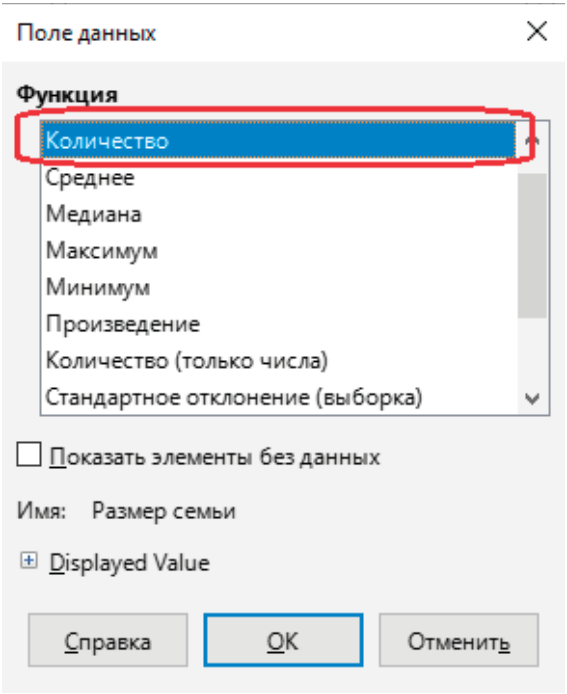


Рис. 4.25. Поля данных сводной таблицы LibreOffice Calc

В следующем параграфе рассмотрим вопросы оценки согласованности мнений респондентов (экспертов).

4.3. Методы оценки согласованности мнений респондентов

Прикладное социологическое исследование в области правоохранительной деятельности направлено на решение задач совершенствования системы управления органами внутренних дел, создания стабильного правоохранительного пространства, оценки и улучшения результатов деятельности органов внутренних дел.

Одной из задач научного исследования является оценка согласованности мнений респондентов. Данная оценка чаще всего проводится по вопросам, предполагающим ранжирование. Наиболее простым и распространенным методом выступает оценка за счет расчета коэффициента конкордации Кендалла (W), показывающего степень согласованности мнений экспертов (респондентов) и рассчитываемого как:

$$W = \frac{S}{\frac{1}{12}m^2(n^3 - n) - m\sum T_i} \quad (4.2),$$

где S – разность между суммой квадратов рангов по каждому признаку и средним квадратом суммы рангов по каждому признаку, m – число респондентов, n – число признаков. S и T_i – рассчитываются как:

$$S = \sum P^2 - \frac{(\sum P)^2}{n} \quad (4.3),$$

$$T_i = \frac{1}{12}\sum(t_i^3 - t_i) \quad (4.4),$$

где P – ранги, t_i – число повторений каждого ранга в i-ом ряду.

Для оценки статистической значимости коэффициента конкордации применяется критерий χ-квадрат по формуле:

$$\chi^2 = \frac{S}{\frac{1}{12}mn(n+1) + \frac{1}{n-1}\sum T_i} \quad (4.5).$$

Рассмотрим технологию оценки ответов экспертной группы на вопросы анкеты, предложенной в параграфе 1 главы 4 настоящего пособия. В данной анкете вопрос № 5: «Насколько, на Ваш взгляд, целесообразно внедрение систем искусственного интеллекта в различные направления деятельности органов внутренних дел», предполагает ранжирование.

Приведем фрагмент листа с данными (рис. 4.26), которые располагаются на листе «Данные».

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		1. Пол	2. Возрастная группа	3. Стаж службы в органах внутренних дел	4. Познания о технологиях и системах искусственного интеллекта	5 - Расследование преступлений	5 - Оперативно-розыскная деятельность	5 - Экспертно-криминалистическая деятельность	5 - Охрана общественного порядка	5- Организационно-аналитическая деятельность	5 -Информационно-аналитическая деятельность	5 - Кадровая работа	5 - Материально-техническое обеспечение
2	1	1	3	5	3	5	5	5	5	4	5	3	1
3	2	2	3	4	2	5	5	5	5	5	5	4	5
4	3	2	3	5	2	5	5	5	5	3	4	3	3
5	4	1	3	5	2	5	5	5	3	3	3	4	4
6	5	1	3	5	2	5	5	5	3	4	4	1	4
7	6	2	3	5	2	5	5	5	5	5	5	5	5
8	7	2	3	4	2	5	4	4	3	4	4	2	2

Рис. 4.26 Фрагмент листа с данными

На отдельном листе создадим таблицу (рис. 4.27) с названием «Расчет».

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		АНК5_1	АНК5_2	АНК5_3	АНК5_4	АНК5_5	АНК5_6	АНК5_7	АНК5_8	ОБР5_1	ОБР5_2	ОБР5_3	ОБР5_4	ОБР5_5	ОБР5_6	ОБР5_7	ОБР5_8	T.(ОБР5)
2	1																	
3	2																	
4	3																	
5	4																	
6	5																	
7	6																	
8	7																	
9	8																	
10	9																	

Рис. 4.27 Фрагмент листа с расчетами

В колонке [A], начиная со строки [2], будут располагаться номера анкет, а в первой строке наименования переменных (АНК5_1... АНК5_8) в диапазоне колонок [B1...I1] – это варианты ответа на соответствующие пункты вопроса № 5 анкеты. Переменные в первой строке (ОБР5_1...ОБР5_8), соответственно, диапазон колонок [J1...Q1] – это расчетные значения.

На первом этапе необходимо рассчитать средний ранг оценки по конкретному вопросу в диапазоне рангов. Для это-

го в ячейку [B2] вводим следующее выражение: «=РАНГ.СР(Данные!F2;Данные!$F2:$M2)»[1], и распространяем введенное значение на весь диапазон ячеек (рис. 4.28).

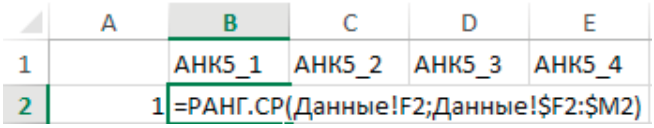


Рис. 4.28 Ввод формулы

Функция «РАНГ.СР» возвращает ранг оценки конкретного вопроса в диапазоне оценок, выставленных респондентом, то есть его величину относительно других значений в списке. В отличие от «РАНГ» функция «РАНГ.СР» возвращает среднее, если несколько оценок имеют одинаковый ранг[2].

На втором этапе рассчитаем сумму рангов. Для этого в ячейку под рассчитанными значениями [B618] введем следующее выражение: «=СУММ(B2:B617)» (рис. 4.29).

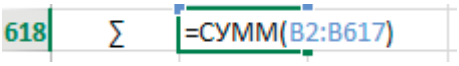


Рис. 4.29 Ввод формулы

Распространим введенную формулу на остальные ячейки [C618:I618], в результате получим следующие значения:

3103,5	2830	2238,5	2676,5	2279	2047	3590	3411,5

В следующей ячейке рассчитаем сумму квадратов. Самый простой способ – предыдущий рассчитанный показатель возвести в квадрат. Для этого в ячейке [B619] введем следующее выражение: «=B618^2» (рис. 4.30).

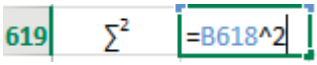


Рис. 4.30 Ввод формулы

[1] Чтобы иметь возможность использовать автозаполнение, как по вертикали, так и по горизонтали фиксируем диапазон ячеек только для колонок, то есть знак $ стоит только перед буквой.

[2] Используем функцию «РАНГ.СР» вместо «РАНГ», так как респондент может выставить одинаковые оценки для двух и более вариантов.

Далее рассчитаем показатель S по формуле (4.3). Для этого в ячейку [B621] введем следующее выражение: «=СУММ(B619:I619)-((СУММ(B618:I618)^2)/8)». Однако существует более простой способ расчета с использованием функции «КВАДРОТКЛ». Для этого в ячейку [B621] введем следующее выражение: «КВАДРОТКЛ(B618:I618)». На рисунке ниже представлены ячейки с расчетом и формулами (рис. 4.31).

620	S=	2253757	СУММ(B619:I619)-((СУММ(B618:I618)^2)/8)
621	S=	2253757	КВАДРОТКЛ(B618:I618)

Рис. 4.31 Ввод формулы

Обратим внимание, что результаты абсолютно идентичны. Таким образом, для расчета целесообразно использовать наиболее быстрый второй метод.

На третьем этапе рассчитаем показатель t_i. Это число повторений каждого ранга в *i*-ом ряду. Для этого в ячейку [J2] вводим выражение «=СЧЁТЕСЛИ(Данные!$F2:$M2;1)», в ячейку [K2] «=СЧЁТЕСЛИ(Данные!$F2:$M2;2)» и т. д. Данная операция призвана рассчитать, какую оценку для каких вариантов указал респондент. Распространяем введенное значение на весь диапазон ячеек (рис. 4.32).

J	K	L	
ОБР5_1	ОБР5_2	ОБР5_3	О
=СЧЁТЕСЛИ(Данные!$F2:$M2;1)			

Рис. 4.32 Ввод формулы

Далее рассчитаем T_i по формуле (4.4). Для этого в ячейку [R2] введем следующее выражение: «=((J2^3-J2)+(K2^3-K2)+(L2^3-L2)+(M2^3-M2)+(N2^3-N2)+(O2^3-O2)+(P2^3-P2)+(Q2^3-Q2))/12». А в ячейке [R618] рассчитаем сумму. Для этого введем следующее выражение: «=СУММ(R2:R617)».

На четвертом этапе рассчитаем коэффициент конкордации (W) по формуле (4.2). Для этого необходимо определить значения m – число респондентов (617), n – число признаков (8). Подставляя полученные значения в формулу (4.2), получаем:

$$W = \frac{2253757}{\frac{1}{12}616^2(8^3 - 8) - 616 \cdot 9808{,}5} \approx 0{,}228.$$

Эти же расчеты можно произвести в MS Excel. Для этого в ячейку [B622] введем следующее выражение: «=B621/((1/12*(616^2)*((8^3)-8))-616*R618)».

Однако для повышения универсальности наших расчетов произведем следующие действия. В ячейку [B623] поместим значение переменной m, а в ячейку [B624] значение переменной n. Для этого в соответствующие ячейки введем следующие выражения [B624] («=СЧЁТ(B2:B617)») и [B624] («=СЧЁТ(B617:I617)»).

Попробуем теперь ввести формулу со ссылкой на соответствующие ячейки (рис. 4.33).

622	W=	0,227764586	B621/((1/12*(616^2)*((8^3)-8))-616*R618)
623	m=	616	СЧЁТ(B2:B617)
624	n=	8	СЧЁТ(B617:I617)
625	W=	0,227764586	B621/((1/12*(B623^2)*((B624^3)-B624))-B623*R618)

Рис. 4.33 Ввод формулы

Обратите внимание, что результаты, рассчитанные двумя подходами, абсолютно идентичны.

На последнем этапе произведем оценку статистической значимости полученного коэффициента конкордации. Для этого рассчитаем значение критерия χ-квадрат по формуле (4.5). Введем в ячейку [B626] следующее выражение: «B621/((1/12*B623*B624*(B624+1))+((1/(B624-1))*R618))», а в ячейку [B627] критическое значение критерия χ-квадрат «ХИ2.ОБР.ПХ(0,05;B624-1)» (рис. 4.34).

626	$\chi^2_{расч}$	442,1546503	B621/((1/12*B623*B624*(B624+1))+((1/(B624-1))*R618))
627	$\chi^2_{крит}$	14,06714045	ХИ2.ОБР.ПХ(0,05;B624-1)

Рис. 4.34 Ввод формулы

Так как расчетное значение критерия χ-квадрат (442,2) больше критического (14,1), то полученные результаты статистически значимы. Однако само значение W = 0,228 говорит о слабой степени согласованности мнений экспертов.

Заключение

В учебном пособии авторами предпринята попытка дать обобщенное описание основных методов и моделей анализа статистических данных, характеризующих результаты научно-исследовательской работы. Очевидно, что успех применения математических моделей в аналитической работе во многом зависит от правильного выбора метода исследования. Для обоснования необходимости разработки и использования моделей того или иного типа необходимо обладать соответствующими знаниями в области теории управления, системного анализа и эконометрики. Изучение состояния процесса (объекта, явления), его динамики и структурных сдвигов должно производиться с применением современных методов визуализации и последующего моделирования. Так, рассмотренные способы и методы визуализации играют важную роль не только на этапе исследования данных, но и на этапе презентации данных.

Предлагаемое пособие не претендует на полное описание инструментария моделирования социально-правовых и экономических процессов (явлений), но при этом рассматриваемые методы и модели являются эффективным практическим инструментом как в научно-исследовательской так и в информационно-аналитической работе штабных и информационных подразделений.

Рассмотренные методы и модели могут быть интересны не только адъюнктам, слушателям и курсантам образовательных организаций системы МВД России, но и сотрудникам информационно-аналитических и штабных подразделений.

Список использованной литературы

Горошко И. В. Технология проведения анкетных опросов // Вестник Университета прокуратуры Российской Федерации. № 3 (77). 2020. С. 72–77.

Елисеева И. И., Курышева С. В. Фиктивные переменные в анализе данных // Социология: 4М. 2010. № 30. С. 43–63.

Инструменты для качественной визуализации данных: искусство использования диаграмм. Копенгаген: Европейское региональное бюро ВОЗ, 2021. Лицензия: CC BY-NC-SA 3.0 IGO.

Информационные технологии в науке и образовании: учебное пособие / И. В. Горошко, Б. А. Торопов. Москва: Академия управления МВД России, 2021. 76 с.

Информационные технологии управления и организация защиты информации: учебник / В. В. Баранов и др. Москва: Академия управления МВД России, 2018. 456 с.

Кравченко Ю. А. Работа полицейского в MS Excel 2013: практическое пособие. Москва, 2016. 320 с.

Математические методы исследования социальных систем: курс лекций / И. В. Горошко, Б. А. Торопов, Ш. Х. Гонов. Москва: Академия управления МВД России, 2019. 80 с.

Новиков Д. А., Новочадов В. В. Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: ВолГМУ, 2005. 84 с.

Новиков Д. А. Статистические методы в педагогических исследованиях (типовые случаи). Москва: МЗ-Пресс, 2004. 67 с.

Носко В. П. Эконометрика для начинающих. Москва, 2005. 379 с.

Торопов Б. А., Гонов Ш. Х. Статистические методы принятия управленческих решений: сборник задач (задачник). Москва: Академия управления МВД России, 2019. 76 с.

Фадеева Л. Н. Теория вероятностей и математическая статистика: учебное пособие / Л. Н. Фадеева, А. В. Лебедев. 2-е изд., перераб. и доп. Москва: Эксмо, 2010. 496 с.

Шеффе Г. Дисперсионный анализ. Москва: Наука; Главная редакция физико-математической литературы, 1980. 512 с.

Ядов В. А. Стратегия социологического исследования. Описание, объяснение, понимание социальной реальности. Москва: Омега-Л, 2007. 567 с.

Яу Н. Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами. Москва: Манн, Иванов и Фербер, 2013. 352 с.

ДЛЯ ЗАМЕТОК

ДЛЯ ЗАМЕТОК

Учебное издание

Шамиль Хасанович Гонов
кандидат технических наук
(Академия управления МВД России)

Игорь Владимирович Горошко
доктор технических наук, профессор
(Академия управления МВД России,
Университет прокуратуры Российской Федерации)

АКТУАЛЬНЫЕ ВОПРОСЫ АНАЛИЗА ДАННЫХ, ХАРАКТЕРИЗУЮЩИХ РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ВНУТРЕННИХ ДЕЛ

Учебное пособие

Редактор *Я. В. Артемьева*
Верстка *С. Н. Портновой*

Подписано в печать _._.2022. Формат 60 × 84 $^{1}/_{16}$.
Усл. печ. л. 7,44. Уч.-изд. л. 3,34. Тираж 65 экз. Заказ № 37у

Отделение полиграфической и оперативной печати РИО
Академии управления МВД России
125171, Москва, ул. Зои и Александра Космодемьянских, д. 8

ISBN 978-5-907530-05-8