Министерство науки и высшего образования
Российской Федерации
Забайкальский государственный университет

А. Н. Шеин
Е. Ю. Юдицких
В. В. Потапов

Статистические методы обработки геофизической информации

Учебное пособие

Чита
ЗабГУ
2023

УДК 550.3:311(075)
ББК 26.20:60.6я73
ББК 26.20:С6я73
Ш 395

Рекомендовано к изданию учебно-методическим советом
Забайкальского государственного университета

Рецензенты
***Д. Л. Авгулевич*,** канд. геол.-минерал. наук,
главный инженер проектов обособленного подразделения
в г. Чите, ООО «АйДи инжиниринг», г. Чита
***С. Ю. Левченко*,** канд. геол.-минерал. наук, гл. геолог,
ООО «Горнорудная Компания», г. Чита

Шеин, Александр Николаевич

Ш 395 Статистические методы обработки геофизической информации : учебное пособие / А. Н. Шеин, Е. Ю. Юдицких, В. В. Потапов ; Забайкальский государственный университет. – Чита : ЗабГУ, 2023. – 152 с.

ISBN 978-5-9293-3230-2

В учебном пособии рассмотрены основные вопросы статистической обработки геофизической информации. Изложены теоретические основы по каждому из рассматриваемых методов, а также предложены варианты решения данных задач с использованием пакета «Анализа данных» Microsoft Excel.

Издание предназначено для студентов специальности 21.05.03 *Технология геологической разведки,* специализация «Геофизические методы поиска и разведки месторождений полезных ископаемых», а также может быть полезно геофизикам-интерпретаторам и при изучении дисциплин «Статистическая обработка геофизической информации», «Комплексирование геофизических методов», «Разведочная геофизика».

УДК 550.3:311(075)
ББК 26.20:60.6я73
ББК 26.20:С6я73

ISBN 978-5-9293-3230-2

Оглавление

Введение

В последние годы непрерывно растёт объём ***геоинформации,*** что обусловлено, в первую очередь, переходом на цифровую регистрацию физических полей и стремительное развитие технологий. Для обработки поступающей **гео**физической, **гео**логической, **гео**графической ***информации*** в настоящее время применяются практически все разделы современной математики. Возрастающий объём цифровой ***информации*** и всё возрастающие требования к современным методам исследования Земли привели к тому, что применение современных методов обработки стало невозможным без использования вычислительной техники и современных прикладных программ. Тем не менее, человек (интерпретатор) по-прежнему остаётся основным действующим лицом в этом процессе.

Обработка ***геоинформации,*** в том числе и **гео**физической, – важнейший этап анализа экспериментальных данных. Основой получения геофизической информации являются ***измерения*** – нахождение значения физической величины опытным путём посредством технических средств. Такие измерения чаще всего называют ***данными.*** В геофизике предметом измерения являются физические свойства горных пород и физические поля, создаваемые горными породами. Цель обработки **гео**физических ***данных*** – извлечение полезной информации из результатов измерений отдельных геофизических методов и их комплексов.

Существуют два подхода к обработке и интерпретации результатов геоинформации: детерминированный и вероятностно-статистический.

Основой детерминированного подхода является применение аналитических методов, где решение прямых и обратных задач детерменировано уравнениями и/или функциональными зависимостями (*определены* – от англ. *determine* – «определять»).

Не менее важным является вероятностно-статистический подход. В этом случае полученные в отдельных точках данные рассматриваются как случайные события. Для исследователя

случайно и расположение геологических объектов, точек и площадей исследования. Кроме того, физическое поле также реализуется случайным образом из-за наложения помех различной природы. При вероятностно-статистическом подходе результатом решения является уже не число и не функция, а распределение вероятностей, заданное для возможных значений искомого параметра.

Для обработки геофизической информации применяются практически все разделы современной математики, которые, естественно, нельзя охватить в одном учебном издании. В работе рассматриваются основные статистические методы обработки геофизических данных.

Учебное пособие включает следующие разделы: «Информация и данные»; «Статистика и история её развития. Статистические данные»; «Статистические методы обработки данных»; «Формы представления статистических данных»; «Простейшие статистические показатели»; «Основные понятия математической статистики»; «Некоторые законы распределения случайной величины»; «Статистические оценки. Доверительный интервал»; «Ряды динамики. Анализ временных рядов»; «Двумерный статистический анализ и его применение»; «Система множества случайных величин и её статистические характеристики»; «Основы дисперсионного анализа». Это наиболее часто встречающиеся операции в обработке геофизической информации. В конце каждого раздела приведены перечень контрольных вопросов и заданий, список литературы. Наряду с теоретическими выкладками, которые сами по себе не являются оригинальными, рассматривается возможность получения различных статистических характеристик с помощью некоторых прикладных программ. Необходимость использования ЭВМ и новейших компьютерных технологий в обработке геофизической информации не подлежит сомнению. Соответственно, как нельзя кстати в работе рассматриваются возможности некоторых довольно популярных программ в статистическом анализе данных. Такое изложение материала позволяет студентам связать теорию с практическим получением результатов.

1. Информация и данные

В настоящее время нет единого мнения о том, что такое информация. Это понятие до сих пор остаётся дискуссионным в науке. Термин «информация» происходит от латинского слова *informatio,* что означает «сведения, разъяснения, изложение независимо от формы их представления». Существует множество определений информации. Приведём лишь некоторые из них.

Информация – это сведения об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, которые уменьшают имеющуюся о них степень неопределённости, неполноты знаний (Н. В. Макарова).

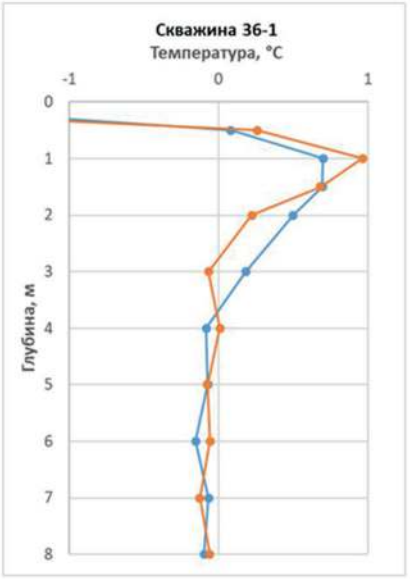
Информация – это обозначение содержания, полученного из внешнего мира в процессе нашего приспособления к нему и приспособления к нему наших чувств (Норберт Винер).

Другими словами, информация – это знания о предметах, фактах, явлениях, которыми могут обмениваться люди в процессе коммуникации. Применительно к геоинформации под информацией подразумеваются сведения о геофизических полях, знания о положении объектов, сведения об экологии или о погоде и др.

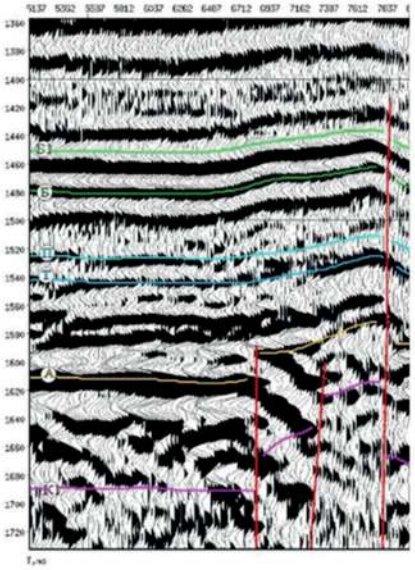
Несмотря на то, что информация и данные в неформальном контексте являются синонимами, в отношении второго понятия имеется большая определённость. Данные – это представление информации (знания о предметах, фактах, явлениях), фактов и идей в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе. Данные – это форма представления информации, приемлемая для её обработки и интерпретации человеком или с помощью автоматических/вычислительных средств.

Примером геоданных могут служить скважинные измерения (рис. 1*а*), сейсмический разрез (рис. 1*б*), табличное представление содержания железа в руде (рис. 1*в*) и др. Геоинформация представлена в разной форме, что и является в нашем случае геоданными.

а



б



в

Класс	*Частота*
30–32	*2*
32–34	*6*
34–36	*9*
36–38	*14*
38–40	*20*
40–42	*25*
42–44	*21*
44–46	*17*
46–48	*13*
48–50	*10*
50–52	*5*

Рис. 1. Геоданные – различные формы представления геоинформации:
а – термометрические скважинные данные; б – сейсмический разрез; в – содержание железа в руде

Данные – это величины, заданные заранее, вместе с условием задачи. Противоположность им – переменные величины. Данные – это зарегистрированные сигналы/показания приборов. Данные могут рассматриваться как записанные наблюдения, которые не используются, а пока хранятся. Если данные ориентированы на их понимание человеком непосредственно при их восприятии или после их некоторого преобразования, то они содержат в себе информацию. Возможна ситуация, когда данные содержат информацию, но на данный момент она не доступна человеку – не позволяет теория или вычислительные мощности и т. д.

Представление о данных может несколько меняться в зависимости от предметной области. Так, в настоящее время данные неразрывно связаны с информационными технологиями и программированием, где данные – это часть программы, совокупность значений определённых ячеек памяти, преобразование которых осуществляет код. С точки зрения компилятора, процессора, операционной системы, это совокупность ячеек памяти, обладающих определёнными свойствами.

Однако стоит сказать о том, что понятие о данных появилось намного раньше, чем возникли первые вычислительные машины. Термин «статистические данные» возник с появлением государственности.

Контрольные вопросы и задания

1. В чём принципиальное отличие между информацией и данными?
2. Приведите пример геоинформации, геофизической информации и данных.

Список литературы

1. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

2. Калинин А. Г. Обработка статистических данных. Чита: ЗИП СибУПК, 2010. 106 с.

3. Шауцукова Л. З. Информатика. 10–11-е кл. М.: Просвещение, 2000.

2. Статистика и история её развития. Статистические данные

Новый завет: Евангелие от Луки, Глава 2

[1]В те дни вышло от кесаря Августа повеление сделать перепись по всей земле.

…

[8]В той стране были на поле пастухи, которые содержали ночную стражу у стада своего.

Прочитав приведённые строки, можно заметить, что зарождение статистики возникло задолго до появления самой отрасли знаний, точкой отсчёта которой многие считают середину XVIII в.

Слово «статистика» происходит от латинского *status* – «состояние, положение вещей» с точки зрения закона, а первоначально *status* – «статистика» – употреблялось в значении «политическое состояние». В науку термин «статистика» ввёл немецкий учёный Готфрид Ахенваль в 1746 г., предложив заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины (рис. 2).

Рис. 2. Готфрид Ахенваль (нем. *Gottfried Achenwall*) (1719–1772), немецкий философ, статистик, экономист, педагог, историк, юрист и один из основоположников статистики

В настоящее время насчитывается около тысячи определений статистики. Дать определение статистики как науки пытались философы, математики, экономисты, социологи, государственные деятели и, конечно, сами статистики.

Приведём некоторые из определений.

Статúстика – это отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических (количественных или качественных) данных.

Статúстика – это наука, изучающая закономерности массовых явлений методом обобщающих показателей.

Статúстика – это одна из общественных наук, имеющая целью сбор, упорядочивание, анализ и сопоставление числового представления фактов, относящихся к самым разнообразным массовым явлениям. Вместе с тем статистика – это учение о системе показателей, т. е. количественных характеристик, дающих всестороннее представление об общественных явлениях, национальном хозяйстве в целом и отдельных его отраслях.

Статúстика – это эффективное орудие, инструмент познания, используемый в естественных и общественных науках для установления тех специфических закономерностей, которые действуют в конкретных массовых явлениях, изучаемых данной наукой.

Статúстика – это также одна из форм практической деятельности людей, целью которой являются сбор, обработка и анализ массовых данных о тех или иных явлениях. Когда мы говорим «государственная и ведомственная статистика РФ», «организация статистики в России», то имеем в виду особую форму практической деятельности людей.

Как уже было замечено в самом начале этого раздела, издавна в каждом государстве соответствующими органами власти собирались сведения о числе жителей по полу, возрасту, занятости в различных сферах труда, наличии различных воинов, вооружения, денежных средств, орудий труда, средств производства и т. д. Все эти и подобные им данные называются статистическими. Под статистическими данными в этом случае понимают совокупность количественных характеристик социально-экономических явлений и процессов, полученных в результате статистического наблюдения, их обработки или соответствующих расчетов.

С развитием государства и международных отношений возникла необходимость анализа статистических данных, их

прогнозирования, обработки, оценки достоверности основанных на их анализе выводов и т. п. К решению таких задач стали привлекаться математики. Таким образом, в математике сформировалась новая область – математическая статистика, изучающая общие закономерности статистических данных или явлений и взаимосвязи между ними.

Сфера применения математической статистики распространилась на многие, особенно экспериментальные, науки. Так, появились экономическая, медицинская, биологическая статистика, статистическая физика и т. д.

В настоящее время под термином «статистические данные» понимают все собранные сведения, которые в дальнейшем подвергаются статистической обработке. В различной литературе их еще называют переменными, вариантами, величинами, датами и т. д. Все статистические данные можно разделить на следующие виды:

– ***качественные***, труднодоступные для измерения (больше, меньше; сильно, слабо; красный, чёрный; мужской, женский и т. д.), и **количественные**, которые можно измерить и представить в виде числа общих мер (2 кг, 3 м, 15 с и т. д.);

– ***точные*** (6 человек, 5 столов, деревянный, металлический, мужской, женский и т. д.) и ***приближенные*** (все измерения: рост 170 см, вес 56 кг, результат бега на 100 м – 10,3 с, и т. д.);

– ***определённые (детерминированные)***, причины появления, не появления или изменения которых известны (2 + 3=5, подброшенный вверх камень обязательно будет иметь вертикальную скорость, равную 0, и т. д.), и ***случайные,*** которые могут появляться и не появляться или не все причины изменения которых известны (пойдёт дождь или нет, родится девочка или мальчик, команда выиграет или нет, в беге на 100 м – 12,2 с, принятая нагрузка вредна или нет).

Общее свойство, присущее нескольким статистическим данным, называют их статистическим признаком. Статистической совокупностью называют несколько статистических данных, объединённых в группу хотя бы одним статистическим признаком. Число данных в статистической совокупности называют её объемом и обозначают n. Различают следующие совокупности:

– ***бесконечные*** – $n \to \infty$, ***конечные*** – n – конечное число;

– ***большие*** – $n > 30$, ***малые*** – $n < 30$;

– ***генеральные*** – содержащие все данные, обусловленные постановкой задачи, ***выборочные*** – части генеральных совокупностей.

Стоит сказать о том, что какие бы ни были статистические данные, их нужно уметь правильно обработать. Частое неверное толкование статистических данных привело к широко распространённому ложному представлению о вседоказанности любого явления с помощью статистики. Небрежное применение законов статистики или умышленное опускание некоторых фактов может ввести людей в заблуждение. Иногда одно и то же явление даже квалифицированные специалисты, во всех тонкостях знающие статистику, могут объяснить по-разному, принять ложное утверждение и отвергнуть правильное. Соответственно, принятие утверждения как истинного в известной мере зависит от субъективных особенностей исследователя. Следовательно, выводы, которые делаются на основании статистических данных, не всегда однозначны.

Несмотря на существующую неоднозначность заключений при обработке статистических данных, сила статистики заключается и в том, что она на основе анализа разрозненных, как бы пестрящих случайностями данных помогает исследователю проникнуть в существо изучаемых явлений.

Контрольные вопросы и задания

1. Кратко охарактеризуйте историю развития статистики.
2. Что такое статистика как наука?
3. Что такое статистические данные?
4. Приведите классификацию статистических данных.

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.

2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

3. Шауцукова Л. З. Информатика. 10–11-е кл. М.: Просвещение, 2000.

3. Статистические методы обработки данных

С появлением быстродействующих ЭВМ возможность применения математической статистики в различных сферах деятельности человека постоянно возрастает. Тем не менее, по большому счёту, основным «устройством для обработки информации» до сих пор является сам человек. Следовательно, то, что называют современными «информационными технологиями», сводится к обработке данных человеком с помощью различных методов, включая применение современных компьютеров и программ для них, а также методы создания и издания: книг, фильмов, музыки, веб-сайтов, справочников, учебных пособий и т. п.

Для повышения качества представления человеку данные преобразуются из одного вида в другой с помощью методов обработки.

Типичными целями обработки данных являются:

– собрать всю доступную информацию, представленную в данных различной природы;

– отделить существенную информацию, представленную данными, от несущественной (шум, выбросы) для рассмотрения в данный момент;

– представить существенную информацию в виде, наиболее удобном для восприятия человеком.

Обработка данных включает такие операции, как:

1) ввод (сбор) данных – это накопление данных с целью обеспечения достаточной полноты для принятия решений, например ввод данных в различные информационные системы:

– автоматический ввод данных;

– ручной ввод данных;

2) формализация данных – это приведение данных к форме, удобной для восприятия человеком или устройством (таблицы, графики, разрезы, гистограммы, диаграммы и т. д.);

3) фильтрация данных – это отсеивание «лишних» данных (шум, выбросы), в которых нет необходимости для повышения достоверности и адекватности или которые даже могут привести к неверным выводам;

4) сортировка данных – это упорядочивание данных по заданному признаку с целью удобства использования;

5) архивация – это организация хранения данных в удобной и легкодоступной форме, в том числе:

– длительное хранение данных;

– надёжность хранения данных;

– учёт и инвентаризация данных;

6) защита данных – включает меры, направленные на предотвращение утраты, воспроизведения и модификации данных, контроль доступа к данным;

7) транспортировка данных – приём и передача данных между участниками информационного процесса;

8) преобразование данных – это перевод данных из одной формы в другую или из одной структуры в другую;

9) представление данных:

– текстовое представление данных;

– табличное представление данных;

– графическое представление данных;

– визуальное представление данных;

– форматы представления данных в различных информационных системах.

Можно выделить некоторые методы/методики обработки статистических данных:

1) диалектический метод познания – заключается в том, что общественные явления и процессы рассматриваются в развитии, взаимосвязи и причинной обусловленности;

2) метод статистического наблюдения – научно организованный сбор сведений, заключающийся в регистрации тех или иных фактов, признаков, относящихся к каждой единице изучаемой совокупности, который обеспечивает полноту, всеобщность и представительность полученной в результате исследования первичной информации об отдельных единицах изучаемого явления;

3) метод группировки и сведения материала – позволяет выделять в изучаемой совокупности подгруппы и обобщать данные статистического наблюдения;

4) научный анализ исследуемых явлений – применение научно-обоснованных методик расчёта в виде, например, средних, относительных величин, выявление закономерностей в распределениях, динамике показателей, оценивание возможности применения методик прогнозирования и пр.;

5) табличный, цифровой, интерактивный, графический и другие методы отображения информации – используются при представлении результатов и итогов статистического исследования явлений и объектов.

В следующем подразделе разберём некоторые методы обработки статистических данных.

Сводка и группировка

Прежде чем непосредственно перейти к простейшим методам обработки данных, введём некоторые понятия.

Объектом статистического исследования является статистическая совокупность – множество объектов или явлений (геоданных), объединённых качественной основой, но отличающихся отдельными признаками. В нашем случае это могут быть данные, полученные различными методами, или даже данные, полученные в различных областях геонаук: геофизика, метеорология, геология и др.

Единица совокупности – первичный элемент статистической совокупности, являющийся носителем признаков и основой ведущегося при исследовании счёта. Признак единицы совокупности – свойства единицы совокупности, которые могут отличаться способами их измерения и другими особенностями, что позволяет проводить их классификацию. Для геоданных ими являются физические параметры (например, сопротивление, плотность, температура), геофизические поля, геометрические или геохимические параметры и т. д.

Генеральная совокупность – полная совокупность объектов, имеющих отношение к изучаемой проблеме.

Выборочная совокупность (выборка) – часть генеральной совокупности, непосредственно участвующая в исследовании.

Репрезентативность – соответствие характеристик выборки характеристикам генеральной совокупности в целом. Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определённой выборки на всю генеральную совокупность, из которой она была сформирована. Репрезентативность можно определить и как свойство выборочной совокупности представлять параметры генеральной совокупности, значимые с точки зрения задач исследования. Другими словами, в репрезентативной выборке присутствуют все группы (признаки?), присутствующие в генеральной совокупности, как их процентное соотношение. Примеры неудачных и репрезентативных выборок показаны на рис. 3.

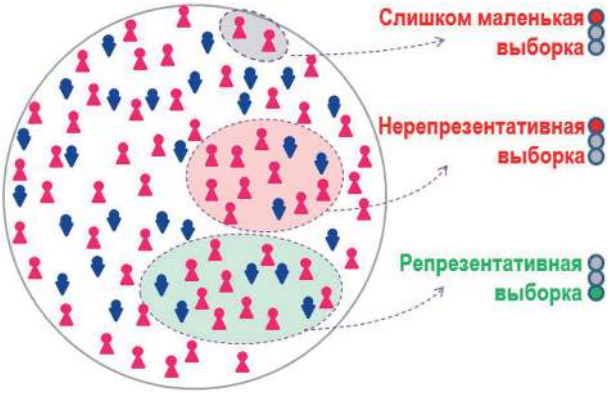


Рис. 3. Пример неудачных и репрезентативных выборок

Ошибка выборки (доверительный интервал) – отклонение результатов, полученных с помощью выборочного наблюдения, от истинных данных генеральной совокупности.

Сводка представляет собой второй этап любого статистического исследования после сбора первичных данных. Сводкой в статистике называется научно организованная (по заранее разработанной программе) обработка материалов наблюдения, которая включает, кроме обязательного контроля собранных данных, систематизацию, группировку, составление таблиц, получение итоговых и производных показателей (от-

носительных и средних величин). Целью сводки является получение обобщающих статистических показателей, отражающих сущность изучаемых процессов и явлений в целом и наличие определённых статистических тенденций/связей.

Одним из основных и наиболее распространённых методов обработки и анализа первичных статистических данных является ***группировка,*** которая позволяет представить в «сжатом» и обозримом виде большой объём первичных данных, собранных в ходе статистического наблюдения, и на этой основе судить о наличии статистических закономерностей. ***Группировка*** – это объединение единиц совокупности в некоторые группы, имеющие свои характерные особенности, общие черты и сходные размеры изучаемого признака.

Статистические группировки преследуют три основные цели:

1) выделение качественно однородных совокупностей;

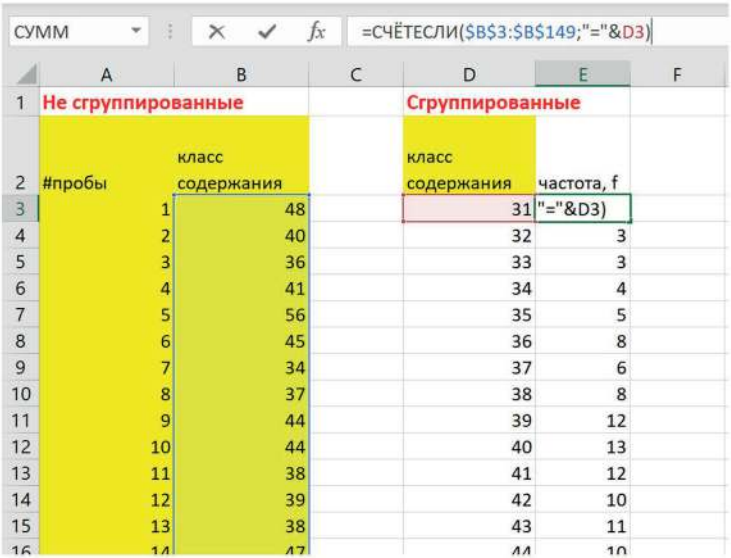
2) изучение структуры совокупности;

3) исследование существующих статистических зависимостей.

Каждой из приведённых целей соответствует особый вид группировки: типологическая, структурная, аналитическая (факторная). ***Типологическая группировка*** решает задачу выявления и характеристики различных типов (частных подсовокупностей). ***Структурная группировка*** даёт возможность описать составные части (структуру) совокупности, а также проанализировать структурные сдвиги, состоящие в изменении величин. ***Аналитическая (факторная)*** группировка позволяет оценивать статистические связи между взаимодействующими признаками изучаемой совокупности.

Рассмотрим простейший пример группировки статистических данных. В примере Ex1_task.xlsx на Листе1 (https://disk.yandex.ru/i/x9RFf9er29VRkA) представлены данные о содержании железа в руде с одного из месторождений для 147 проб. Часть таблицы с этими данными в не сгруппированном виде приведена на рис. 4. В первом столбце таблицы содержатся номера проб, а во втором – класс содержания руды, который варьируется от 31 до 56. Чтобы провести простей-

шую обработку, можно отсортировать данные по классу содержания в порядке возрастания либо сразу сгруппировать (пересчитать элементы) с одинаковым значением признака – класса содержания. Для этого нужно создать столбец со всеми имеющимися в данных классами содержания от 31 до 56 (рис. 4, столбец D от 3 до 28), а затем воспользоваться в MS Excel функцией СЧЁТЕСЛИ(), как это показано на рис. 4 в строке формул. В ячейке рядом с каждым классом (рис. 4, ячейка E3) в качестве диапазона – первый аргумент функции – указать столбец с классом содержания (рис. 4, столбец B от 3 до 149), а в качестве критерия – второй аргумент функции – указать равенство соответствующему классу. То же самое нужно сделать для каждого из классов в соседней ячейке столбца E. В результате таких действий получим простейшую группировку данных по классам содержания, где количество элементов данных с одинаковым значением признака (класс содержания) обозначается f и называется ***частота,*** причём $\Sigma f = n$, где n – число единиц совокупности (элементов выборки).



	A	B	C	D	E	F
1	Не сгруппированные			Сгруппированные		
2	#пробы	класс содержания		класс содержания	частота, f	
3	1	48		31	"="&D3)	
4	2	40		32	3	
5	3	36		33	3	
6	4	41		34	4	
7	5	56		35	5	
8	6	45		36	8	
9	7	34		37	6	
10	8	37		38	8	
11	9	44		39	12	
12	10	44		40	13	
13	11	38		41	12	
14	12	39		42	10	
15	13	38		43	11	
16	14	47		44	10	

Рис. 4. Пример простейшей группировки данных

Мы рассмотрели простейшую группировку, где количество групп совпадает с числом различных признаков (классов содержания), однако чаще наборы данных имеют значительно больше элементов, соответственно, возникает необходимость расчёта интервального шага, построения соответствующих интервалов и определения удельных весов (частот) или средних значений признака в группе.

Интервалы группировки в зависимости от их величины бывают равные и неравные. Чаще всего используют равные интервалы. Тогда величина интервала определяется по простой формуле:

$$h = \frac{\left(x_{max} - x_{min}\right)}{k},$$

где x_{max} и x_{min} – максимальное и минимальное значения признака в совокупности;

k – число групп.

Полученную величину округляют, и она будет являться шириной интервала.

Разберём группировку с равными интервалами на примере Ex1_task.xlsx, Лист2 (https://disk.yandex.ru/i/x9RFf9er29VRkA). Здесь у нас x_{max}=56, x_{min}=32. Проведём группировку на 13 интервалов и по формуле получим h =2. Далее определяем границы интервалов (групп) – столбцы D и E на рис. 5. Для подсчёта количества проб, попадающих в каждую группу, воспользуемся функцией MS Excel СЧЁТЕСЛИМН(), как это показано в строке формул на рис. 5. В ячейке рядом с каждым интервалом (рис. 5, ячейка F3) в качестве диапазона – первый аргумент функции – указать столбец с классом содержания (рис. 5, столбец B от 3 до 149), а в качестве критерия – второй аргумент функции – указать нестрогое неравенство больше или равно левого края первого интервала. Третий аргумент – также столбец с классом содержания (рис. 5, столбец B от 3 до 149), а в качестве критерия – четвёртый аргумент функции – нестрогое неравенство меньше или равно правого края первого интервала. То же самое нужно сделать для каждой из групп в соседней ячейке столбца F. В результате таких действий получим простейшую группировку данных с использованием

равных интервалов. Для проверки себя можно просуммировать количество частот $\Sigma f = 147$, где **147** – число единиц совокупности (элементов выборки – количество проб).

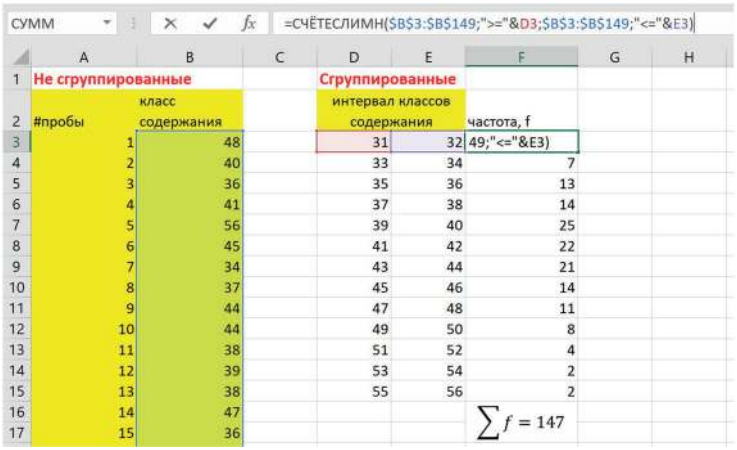


Рис. 5. Пример группировки данных
с использованием равных интервалов

Ранее нами был разобран пример группировки данных с равными интервалами, причём интервалы были выбраны произвольно. Однако считается, что существует оптимальное число групп (интервалов), которое может быть определено по ***Формуле Стёрджесса***:

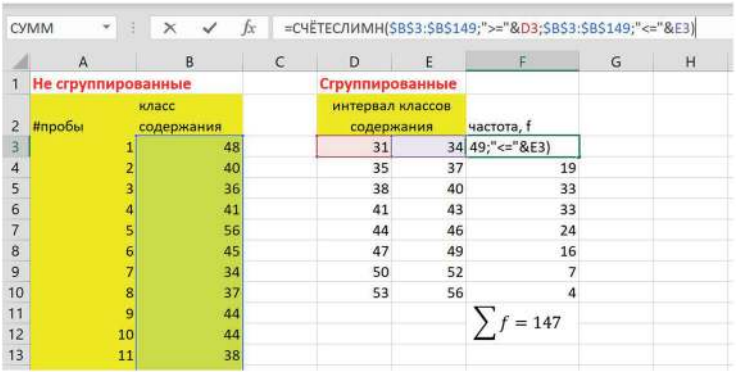
$$k = 1 + (3.332 \times \lg n) \text{ или } k = 1 + (1.4 \times \lg n).$$

Применим эту формулу для нашей выборки, где $n = 147$. Тогда

$$k = 1 + (3.332 \times \lg 147) \approx 8.$$

Теперь, используя алгоритм группировки, разобранный в прошлом примере, можно разбить наши данные на 8 интервалов. В связи с тем что 147 не делится нацело на 8, возьмём первый и последний интервал 4, а остальные 3. Тогда, используя функцию MS Excel СЧЁТЕСЛИМН(), получим результат, представленный на рис. 6 и в примере Ex1_task.xlsx, Лист3

(https://disk.yandex.ru/i/x9RFf9er29VRkA). Для проверки суммируем количество частот $\Sigma f = 147$.



	A	B	C	D	E	F	G	H
1	Не сгруппированные			Сгруппированные				
2	#пробы	класс содержания		интервал классов содержания		частота, f		
3	1	48		31	34	49;"<="&E3)		
4	2	40		35	37	19		
5	3	36		38	40	33		
6	4	41		41	43	33		
7	5	56		44	46	24		
8	6	45		47	49	16		
9	7	34		50	52	7		
10	8	37		53	56	4		
11	9	44				$\sum f = 147$		
12	10	44						
13	11	38						

Рис. 6. Пример группировки данных

Среди простых группировок особо выделяют ряды распределения. Ряд распределения – это группировка, в которой для характеристики групп (упорядоченно расположенных в соответствии со значением изучаемого признака) применяется один показатель – абсолютная частота или численность группы. Другими словами, это ряд чисел, показывающий, как распределяются единицы некоторой совокупности по отношению к изучаемому признаку. Именно такие ряды были получены после группировки данных в разобранных примерах (рис. 4–6), где в качестве признака для формирования групп используется класс содержания руды, а характеризуется каждая группа абсолютной частотой – численность группы.

Ряды распределения, построенные по количественному признаку, называются ***вариационными рядами.*** Любой вариационный ряд состоит их двух элементов: вариантов и частот. Вариантами считаются отдельные значения признака, которые он принимает в вариационном ряду, т. е. конкретное значение варьирующего признака. Частоты – это численности отдельных вариантов или каждой группы вариационного ряда, т. е. это числа, показывающие, как часто встречаются те или иные варианты в ряду распределения. Сумма всех частот определяет численность всей совокупности, её объём.

Нередко частоты заменяют на частости – частоты, выраженные в долях единицы или процентах к итогу. Соответственно, сумма всех частостей равна 1 или 100 %. Для их вычисления в долях нужно поделить частоты на общее количество единиц совокупности (в разобранных примерах Σf=147. Чтобы получить проценты, полученные доли нужно умножить на 100. Таблица с вычисленными частостями в долях и процентах для разобранного примера Ex1_task.xlsx, Лист4 (https://disk.yandex.ru/i/x9RFf9er29VRkA), приводится на рис. 7.

интервал классов содержания		частота, f	частость в долях	частость в %	накопленная частота	накопленная частость
31	32	4	0,03	2,7	4	0,03
33	34	7	0,05	4,8	11	0,07
35	36	13	0,09	8,8	24	0,16
37	38	14	0,10	9,5	38	0,26
39	40	25	0,17	17,0	63	0,43
41	42	22	0,15	15,0	85	0,58
43	44	21	0,14	14,3	106	0,72
45	46	14	0,10	9,5	120	0,82
47	48	11	0,07	7,5	131	0,89
49	50	8	0,05	5,4	139	0,95
51	52	4	0,03	2,7	143	0,97
53	54	2	0,01	1,4	145	0,99
55	56	2	0,01	1,4	147	1,00
		$\sum f = 147$	$\sum = 1$	$\sum = 100$		

Рис. 7. Пример группировки данных пересчётом частот в частости

При изучении вариационных рядов наряду с понятием частоты используется понятие накопленной частоты (рис. 7). Накопленная частота показывает, сколько наблюдалось вариантов со значением признака, которое было меньшим, чем рассматриваемое. Отношение накопленной частоты к общему числу наблюдений называют накопленной частостью. Накопленные частоты (частости) для каждого интервала находятся последовательным суммированием частот (частостей) всех предшествующих интервалов, включая приведённый на рис. 7.

Статистические группировки и классификации преследуют три основные цели:

1) выделение качественно однородных совокупностей;

2) изучение структуры совокупности;

3) исследование существующих статистических зависимостей.

Для достижения указанных целей вариационные ряды подвергают дальнейшему статистическому анализу, в частности вычисляют простейшие статистические показатели – средние значения.

Контрольные вопросы и задания

1. Что такое вариационный ряд?
2. Что такое сводка и группировка статистических данных? Чем отличаются частота и частость?
3. Перечислите формы представления статистических данных.
4. Дайте определение выборочной и генеральной совокупности, репрезентативности.
5. Для чего нужна Формула Стёрджесса?

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.

2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

3. Калинин А. Г. Обработка статистических данных. Чита: ЗИП СибУПК, 2010. 106 с.

4. Формы представления статистических данных

Сгруппированные статистические данные можно представить в различных формах: текстовых, табличных и графических. Текстовые и табличные формы используются для аналитического анализа данных, в то время как графическое представление позволяет наглядно показать закономерность – вариационные ряды.

Ряды распределения показывают закономерность изменения изучаемого признака. Для наглядности выражения закономерностей принято изображать вариационные ряды графически в виде гистограммы, полигона частот, кумуляты, огивы, круговой диаграммы. Разберём каждое из этих представлений.

Гистограмма

Гистогра́мма – способ представления табличных данных в графическом виде – в виде столбчатой диаграммы. Интервальный ряд изображается столбиковой диаграммой, в которой основания столбиков, расположенные по оси абсцисс, – это интервалы значений варьируемого признака, а высоты столбиков – частоты, соответствующие масштабу по оси ординат.

Разберём построение гистограммы в MS Excel. Для этого воспользуемся любым из вариационных рядов, например Ex1_task.xlsx, Лист1 (https://disk.yandex.ru/i/x9RFf9er29VRkA). Выделим два столбца сгруппированного ряда D и E. Затем на вкладке «Вставка» в разделе «Диаграммы» выберем пункт «Рекомендуемые диаграммы» или «Диаграммы». В открывшемся окне найдём и выберем гистограмму с группировкой. При неудачном выборе можно кликнуть по появившейся диаграмме правой кнопкой мыши и в появившемся списке выбрать пункт «Изменить тип диаграммы», где ещё раз найти и выбрать гистограмму с группировкой. Верная диаграмма изображена на рис. 8. После построения нужно добавить название осей: по вертикали «Частота» – высота столбиков, по горизонтали «Класс содержания» – интервалы значений варьируемого признака.

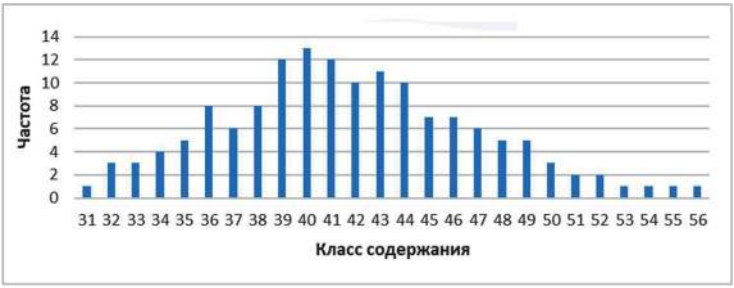


Рис. 8. Пример гистограммы, построенный при использовании пункта «Вставка» в разделе «Диаграммы» MS Excel

Другой способ построения гистограммы – с помощью модуля анализа данных в MS Excel. Для активации модуля необходимо перейти в пункт «Файл→Параметры». Выбрать пункт «Надстройки», где внизу окна напротив строки «Управление» нажать кнопку «Перейти…». В появившемся списке поставить галочки напротив пунктов «Пакет анализа» и «Пакет анализа – VBA». Теперь на вкладке «Данные» появилось поле «Анализ» с пунктом «Анализ данных» – выберем его, кликнув левой кнопкой мыши. В появившемся списке нужно найти и выбрать строку «Гистограмма». Откроется диалоговое окно, где необходимо ввести «Входной интервал» – это столбец B не сгруппированных данных. Интервал карманов – это столбец D – интервалы, например, полученные в примерах Ex1_task.xlsx, Лист2 (https://disk.yandex.ru/i/x9RFf9er29VRkA) и Лист3, причём нужно выбрать правые концы интервалов. В области настроек параметров вывода можно выбрать вывод на новом листе или в выбранном интервале, ниже поставить галочку напротив пункта «Вывод графика» для построения гистограммы. В результате сформируются таблица с карманами (интервалами) и соответствующая гистограмма (рис. 9).

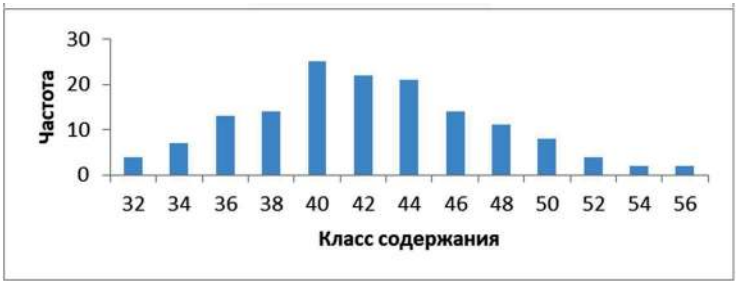


Рис. 9. Пример гистограммы, построенный с помощью модуля анализа данных в MS Excel

Гистограмма – это инструмент, позволяющий зрительно оценить распределение статистических данных, сгруппированных по частоте попадания данных в определённые, заранее установленные интервалы. Если соединить вершины столбиков гистограммы, то получим ещё одну форму представления полученных данных – полигон.

Полигон

Полигон – графическое изображение ряда в виде графика, который получается соединением точек с координатами x_i и f_i. Гистограмма и полигон часто совмещаются (рис. 10). Чтобы совместить полигон и построенную выше гистограмму в MS Excel, нужно кликнуть по гистограмме правой кнопкой мыши и в появившемся списке выбрать пункт «Выбрать данные…». В открывшемся диалоговом окне «Выбор источника данных» кликнуть кнопу «Добавить», после чего в следующем окне в строке «Значение» ввести столбец частот E из примера Ex1_task.xlsx (https://disk.yandex.ru/i/x9RFf9er29VRkA) нашего вариационного ряда. В области диаграммы появится вторая идентичная первой гистограмма другим цветом. Для изменения гистограммы на полигон нужно кликнуть правой кнопкой мыши на гистограмму и в появившемся списке выбрать пункт «Изменить тип диаграммы для ряда…». В появившемся окне для «Ряд2» поменять «Гистограмму с группировкой» на «График» либо «График с областями». В результате получим полигон, совмещённый с гистограммой (рис. 10).

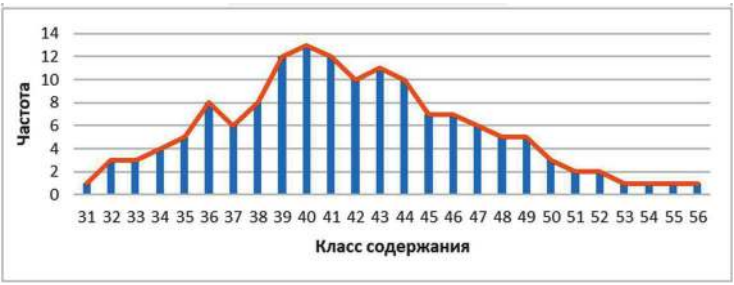


Рис. 10. Пример полигона с наложением на гистограмму, построенные при использовании пункта «Вставка» в разделе «Диаграммы» MS Excel

Кумулята

Кумулята – графическое изображение ряда в виде гистограммы или графика, который получается соединением точек с координатами интервалов варьируемого признака x_k и накопленных частот $\sum_{i=1}^{k} f_i$. Строится аналогично гистограмме и полигону, но вместо частот используются накопленные частоты (рис. 11).

Для примера используем Ex1_task.xlsx, Лист4 (https://disk.yandex.ru/i/x9RFf9er29VRkA).

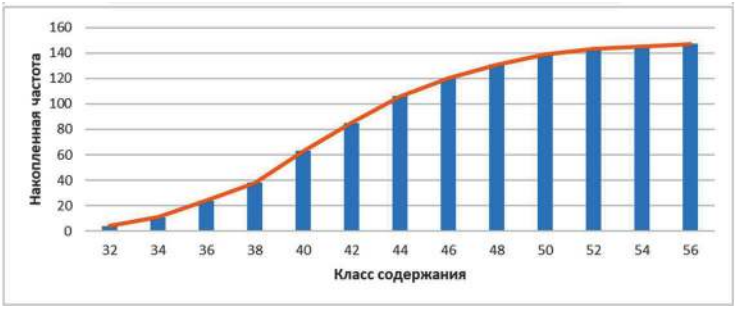


Рис. 11. Пример кумуляты, построенной при использовании пункта «Вставка» в разделе «Диаграммы» MS Excel

В таблице данных примера Ex1_task.xlsx, Лист4, выберем правые концы интервалов и столбец накопленных частот в

столбцах E и I соответственно. Далее нужно выполнить действия как при построении гистограммы и полигона. В результате получим кумуляту в виде гистограммы и графика (рис. 11).

Огива

Огива – графическое изображение ряда в виде гистограммы или графика, который получается соединением точек с координатами накопленных частот $\sum_{i=1}^{k} f_i$ и интервалов варьируемого признака x_k. Иными словами, огива – это ***кумулята*** с развёрнутыми осями. В MS Excel огиву можно построить, выбрав в списке рекомендуемых диаграмм линейчатую диаграмму с группировкой. Для примера будем использовать Ex1_task.xlsx, Лист4 (https://disk.yandex.ru/i/x9RFf9er29VRkA), где выберем правые концы интервалов и накопленные частоты в столбцах E и I соответственно. В результате получим огиву, представленную на рис. 12.

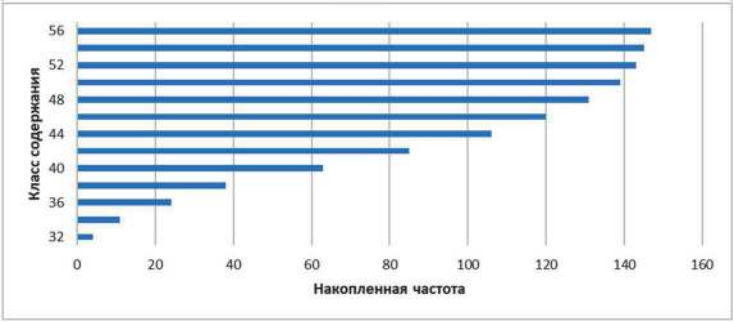


Рис. 12. Пример огивы, построенной при использовании пункта «Вставка» в разделе «Диаграммы» MS Excel

Круговая диаграмма

Круговая диаграмма – это круговая статистическая диаграмма, которая разделена на срезы, чтобы проиллюстрировать числовую пропорцию. На круговой диаграмме длина дуги каждого среза пропорциональна величине, которую он

представляет. Например, можно сгруппировать образцы по типу грунта и выразить это в процентном отношении. Например, если 60 % образцов – песок, 25 % – суглинок, 15 % – глина, то в MS Excel можно сформировать два столбца с типом грунта и процентной долей, как это сделано в Ex1_task.xlsx, Лист5 (https://disk.yandex.ru/i/x9RFf9er29VRkA). Выделив эти два образца, нужно в пункте главного меню «Вставка» выбрать пункт «Рекомендуемые диаграммы» или «Круговая диаграмма». После этого появится круговая диаграмма, которая будет наглядно демонстрировать полученное соотношение (рис. 13).

Рис. 13. Пример круговой диаграммы, построенной при использовании пункта «Вставка» в разделе «Диаграммы» MS Excel

Глядя на все графические представления, можно дать описание ряда на качественном уровне, однако существуют численные статистические показатели или характеристики, которые вместе или вместо графического изображения характеризуют анализируемые данные.

Контрольные вопросы и задания

1. Назовите формы представления статистических данных.
2. Что такое гистограмма, полигон, огива, кумулята, круговая диаграмма? Приведите примеры.
3. Как построить гистограмму, полигон, огиву, кумуляту, круговую диаграмму?

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.

2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

3. Розенцвайг А. К., Исавнин А. Г. Статистика. Сводка и группировка данных статистического наблюдения: учеб.-метод. пособие. Набережные Челны: Изд-во Набережночелнинского института КФУ, 2019. 29 с.

4. Савинский И. Д. Таблицы вероятностей подсечения эллиптических объектов прямоугольной сетью наблюдений. М.: Недра, 1964. 86 с.

5. Простейшие статистические показатели

Простейшими статистическими показателями, которые применяют повсеместно при исследовании вариационных рядов являются средни***Е*** значени***Я***. Ещё раз акцентируем внимание на том, что средних значений существует несколько, а выбор средней величины зависит от содержания осредняемого признака и конкретных данных, по которым её приходится вычислять.

Среднее значение – это обобщающий показатель, характеризующий значение признака, вокруг которого концентрируются наблюдения. Различают следующие структурные (медиана, мода) и аналитические средние величины (средняя арифметическая, средняя гармоническая, средняя геометрическая, средняя квадратическая, средняя степенная).

Медиана (Me) – величина варьирующего признака, делящая совокупность на две равные части – со значением признака меньше медианы и со значением признака больше медианы. Другими словами, ***медиана (Me)*** – вариант, приходящийся на середину вариационного ряда. Далее приводится формула, по которой можно найти медиану, а в табл. 1 представлен вариационный ряд, где жирным цветом выделяется медиана: вариант $x_i = 17$, с частотой $f_i = 15$.

$$Me = \begin{cases} x_j, & \text{если } n = 2j - 1; \\ \frac{1}{2}\left(x_j + x_{j+1}\right), & \text{если } n = 2j. \end{cases}$$

Таблица 1

Пример медианы вариационного ряда

x_i	15	16	**17**	18	19
f_i	5	10	**15**	20	12

Медианный интервал (Me) – интервал, где накопленная частота равна или превышает полусумму всех частот ряда. Для вариационного ряда, приведённого в табл. 2 жирным

шрифтом, выделен медианный интервал, вычисление которого будет выглядеть следующим образом:

$$Me = x_{\min} + d\frac{\frac{1}{2}\sum_{i=1}^{k} f_i + \sum_{i=1}^{i_{Me}-1} f_i}{f_{Me}} = 250 + 50\frac{\frac{510}{2} - 170}{115} = 286.96,$$

где $x_{\min}$ – левая граница интервала;

$$d = x_{\max} - x_{\min}.$$

Таблица 2

Пример медианного интервала вариационного ряда

Интервал		***Частоты***	***Накопленные частоты***
10	150	20	20
150	200	50	70
200	250	100	170
250	300	115	285
300	350	180	465
350	400	45	510

Значения признака, делящие ряд на 4 равные части, называют ***квартилями,*** на 5 равных частей – **квинтилями**, на 10 частей – ***децилями,*** на 100 частей – ***перцентилями.*** Эти показатели вычисляются аналогично медиане, только вместо двух частей ряд делят на соответствующее количество интервалов.

Мода (Mo) – величина признака, которая встречается в изучаемом ряду распределения чаще всего. В дискретном ряду ***мода*** (***Mo***) определяется без вычисления как значение признака с наибольшей частотой (табл. 3).

Таблица 3

Пример моды вариационного ряда

x_i	15	16	17	**18**	19
f_i	5	10	15	**20**	12

Для интервального ряда мода вычисляется по приведённой далее формуле, где вычисляется мода для ряда из табл. 4.

$$Mo = x_{\min} + d\frac{(f_{Mo} - f_{Mo-1})}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} =$$

$$= 140 + 20\frac{(18-11)}{18-11+18-4} = 146.7.$$

Таблица 4

Пример моды интервального вариационного ряда

x_i	100–120	120–140	**140–160**	160–180
f_i	6	11	**18**	4

Прежде чем перейти к вычислению аналитических средних значений, разберём простой пример: в одну сторону машина двигалась 6 ч со скоростью 40 км/ч, в обратную сторону – 4 ч со скоростью 60 км/ч. Расстояние между пунктами – 240 км (рис. 14). Найдём среднюю скорость движения машины.

Рис. 14. Пример для демонстрации вычисления средних величин

Вначале воспользуемся известной всем формулой средней арифметической и получим среднюю скорость 50 км/ч:

$$v_{срарифм} = \frac{v_1 + v_2}{2} = \frac{40+60}{2} = 50\ км/ч.$$

Если же рассуждать немного по-другому, то машина проехала два раза по 240 км и затратила на это 6 и 4 ч. Таким образом, средняя скорость будет вычисляться по формуле средней гармонической:

$$v_{сргармонич} = \frac{240+240}{\frac{240}{60}+\frac{240}{40}} = \frac{1+1}{\frac{1}{60}+\frac{1}{40}} = 48\ км/ч.$$

Видно, что разные подходы при вычислении среднего значения дают разную среднюю скорость, причём распространённое среднеарифметическое значение в данном случае даёт неправильный результат. Соответственно, при вычислении среднего нужно принимать во внимание содержание осредняемого признака и конкретных данных, по которым её приходится вычислять.

Средняя арифметическая простая (невзвешенная) – вычисляется, когда каждый вариант совокупности встречается только один раз:

$$\overline{X} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Средняя арифметическая (взвешенная) – варианты повторяются различное число раз, при этом число повторений вариантов называется частотой, или статистическим весом. Название «вес» выражает тот факт, что разные значения признака имеют неодинаковую «важность» при расчёте средней величины.

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + ... + x_k f_k}{f_1 + f_2 + ... + f_k} = \frac{\sum_{i=1} x_i f_i}{\sum_{i=1}^{k} f_i}.$$

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменную сумму квадратов исходных величин, то средняя будет являться квадратической средней (средневзвешенной) величиной:

$$\overline{X}_{кв} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}}\ ;\ \overline{X}_{кв} = \sqrt{\frac{\sum_{i=1}^{k} X_i^2 f_i}{\sum_{i=1}^{k} f_i}}.$$

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменным произведение индивидуальных величин, то следует применить геометрическую среднюю (средневзвешенную) величину, имеющую следующий вид:

$$\overline{X}_{геом} = \sqrt[N]{X_1 \cdot X_2 \cdot \ldots \cdot X_N};$$

$$\overline{X}_{геом} = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_k^{f_k}}.$$

Средняя геометрическая используется для анализа динамики явлений и позволяет определить средний коэффициент роста. При расчёте средней геометрической индивидуальные значения признака обычно представляют собой относительные показатели динамики, построенные в виде цепных величин как отношение каждого уровня ряда к предыдущему уровню.

Средняя гармоническая вычисляется в тех случаях, когда приходится суммировать не сами варианты, а обратные им величины. Средняя гармоническая и средняя гармоническая взвешенная вычисляются по следующим формулам:

$$\overline{X}_{гарм} = \frac{1+1+\ldots+1}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots\frac{1}{x_N}} = \frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}};$$

$$\overline{X}_{гарм} = \frac{f_1+f_2+\ldots+f_N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\ldots\frac{f_k}{x_k}} = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k}\frac{f_i}{x_i}}.$$

Соотношения между средней, модой и медианой характеризуют вариационный ряд и распределение в нём элементов:

$x_{ср} = \boldsymbol{Mo} = \boldsymbol{Me}$, то распределение симметрично;

$\boldsymbol{Me} < x_{ср}$ характерно при небольшой группе с большими числами;

$x_{ср} < \boldsymbol{Me}$ соответствует большой концентрации данных и не очень большим числам;

$\boldsymbol{Mo} < x_{ср}$, если совокупность неоднородна;

Мо $> x_{ср}$, если совокупность небольшая и мода отчётливо выражена.

Средние арифметическая, гармоническая, геометрическая и квадратическая, рассчитанные для одного и того же ряда вариантов, отличаются друг от друга. Их численное значение возрастает с ростом показателя степени в формуле степенной средней по правилу мажорантности средних А. Я. Боярского, т. е.

$$\overline{X}_{гарм} < \overline{X}_{геом} < \overline{X}_{ариф} < \overline{X}_{квадр}.$$

Контрольные вопросы и задания

1. Сколько простейших статистических показателей средних значений Вам известно?
2. Дайте определение медианы, моды, среднего.
3. Всегда ли равны различные средние между собой и что такое «правило мажорантности средних А. Я. Боярского»?

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.
2. Иваненкова А.П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.
3. Розенцвайг А. К., Исавнин А. Г. Статистика. Сводка и группировка данных статистического наблюдения: учеб.-метод. пособие. Набережные Челны: Изд-во Набережночелнинского института КФУ, 2019. 29 с.
4. Савинский И. Д. Таблицы вероятностей подсечения эллиптических объектов прямоугольной сетью наблюдений. М.: Недра, 1964. 86 с.
5. Статистические характеристики процессов. URL: https://www.moodle.kstu.ru/pluginfile.php/383238/mod_resource/content/1/ЦМХТП_Т2_Статистические%20характеристики%20процессов_ЛР4.pdf (дата обращения: 17.02.2023). Текст: электронный.

6. Основные понятия математической статистики

6.1. Событие и вероятность

Для более глубокого статистического анализа вариационных рядов вспомним элементы теории вероятности и основные понятия математической статистики.

Наблюдаемые при проведении экспериментов события делятся на достоверные, невозможные и случайные. Если при конкретных условиях проведения эксперимента событие может как произойти, так и не произойти, оно называется случайным. Достоверным называется событие, которое при том же комплексе условий эксперимента обязательно происходит, а невозможным – то, которое заведомо не может произойти при данных условиях.

Теория вероятностей изучает закономерности случайных событий во времени и пространстве и приёмы их количественного описания.

При геофизических, геологических и других геонаблюдениях полученные в отдельных точках данные целесообразно рассматривать именно как случайные события, т. к. геологические объекты, обуславливающие появление конкретных значений поля, и точки наблюдений по площади исследований расположены случайным образом. Случайно и наложение помех, вызванных разнообразными причинами.

Событием в геоданных можно считать: появление конкретного значения физического параметра или физического поля, появление аномалии какого-либо поля, факт соответствия определённых значений поля конкретному типу горных пород и т. д. Именно поэтому большинство инструментов статистики можно применять для геоданных.

Два события называются ***несовместными***, если появление одного из них исключает появление другого при одном и том же эксперименте.

Суммой событий называется событие, состоящее в появлении хотя бы одного из этих событий.

Произведением событий называется событие, состоящее в совместном появлении всех этих событий.

События A_1, A_2 ,..., A_n образуют ***полную группу событий,*** если они попарно несовместны, а в сумме образуют ***достоверное событие,*** т. е. какое-либо из них обязательно происходит, причём только одно.

Противоположными событиями называются два несовместных события, образующие полную группу.

Критической мерой степени объективной возможности того или иного события A служит ***вероятность события*** $P(A)$, которая измеряется отношением числа m благоприятствующих событию A исходов к общему числу n всех равновозможных исходов экспериментов: $P(A)=m/n$. Это классическое определение вероятности.

На практике чаще всего используют статистическое определение вероятности, при котором вероятностью события называют относительную частоту его появления при многократном воспроизведении комплекса условий эксперимента. При большом числе опытов частота события A стремится к вероятности $P(A)$ в её классическом определении.

Вероятность события A, вычисленная при условии, что произошло событие B, называется условной вероятностью $P(A/B)$ события A.

Два ***события*** называются ***независимыми,*** если появление одного из них не изменяет вероятности появления другого, т. е. для независимых событий $P(A/B)=P(A)$, а для зависимых $P(A/B)\neq P(A)$.

Виды различных событий, соответствующие им вероятности и свойства этих вероятностей сведены в табл. 5. Следствием этих свойств вероятностей является формула полной вероятности, на основе которой определяется вероятность события A, происходящего вместе с одним из событий H_1, H_2, ..., H_n, образующих полную группу. События H_1, H_2, ..., H_n называют гипотезами, а формула полной вероятности выглядит следующим образом:

$$P(A)=\sum_{i=1}^{n} P(H_i)P(A/H_i).$$

Таблица 5

Виды событий и соответствующие им вероятности

№ п/п	*Событие*	*Вероятность события*	*Свойства вероятности*
1	Достоверное U	$P(U)$	$P(U)=1$
2	Невозможное V	$P(V)$	$P(V)=0$
3	Случайное A	$P(A)$	$0 \le P(A) \le 1$
4	Противоположное $\bar{A}$	$P(\bar{A})$	$P(\bar{A}) = 1 - P(A)$
5	Произведение двух событий (пересечение) AB ($A \cap B$)	$P(AB)$	Независимые события $P(AB) = P(A)P(B)$ Зависимые события $P(AB) = P(A/B)P(B) = P(B/A)P(A)$
	Произведение n событий $\prod_{i=1}^{n} A_i$	$P(\prod_{i=1}^{n} A_i)$	Независимые события $P(\prod_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$

Окончание табл. 5

№ п/п	***Событие***	***Вероятность события***	***Свойства вероятности***
6	Сумма двух событий (объединение) $A+B$ (A ∪ B)	$P(A+B)$	Несовместные события $P(A+B)=P(A)+P(B)$ Совместные события $P(A+B)=P(A)+P(B)-P(AB)$
	Сумма n событий $\sum_{i=1}^{n} A_i$	$P(\sum_{i=1}^{n} A_i)$	Полная группа $P(\sum_{i=1}^{n} A_i)=\sum_{i=1}^{n} P(A_i)=1$ Любая группа $P\left(\sum_{i=1}^{n} A_i\right)=\sum_{i=1}^{n} P(A_i)-$ $\sum_{i,j}^{n} P(A_i A_j)+\sum_{i,j,k}^{n} P(A_i A_j A_k)+$ $+(-1)P^{n-1}(A_1, A_2,\ldots, A_n)$

Следствием свойств умножения и формулы полной вероятности является формула Бейеса, или теорема гипотез, или формула обратных вероятностей. Формула Бейеса играет важную роль. На её основе решаются задачи выделения сигналов на фоне помех, обработки данных комплекса геофизических полей, определения параметров источников аномалий при количественной интерпретации и др.

Понятие вероятности события допускает также простую геометрическую трактовку. Геометрическая вероятность обобщает классическое определение на бесконечное множество экспериментов. К понятию геометрической вероятности приводит задача о бросании точки в некоторую ограниченную область G, содержащую меньшую по размерам область g, причём все положения падения точки в область G считаются равновозможными.

Если событие A состоит в попадании точки в область g, то $P(A) = мера\ g/мера\ G$, где под мерами областей g и G можно понимать длины отрезков, размеры площадей или объёмов.

Понятие геометрической вероятности используется при решении задачи Бюффона, заключающейся в определении вероятности пересечения наугад брошенной иглы длиной $2l$ одной из параллельных прямых, отстоящих друг от друга на плоскости на расстоянии $2a$, причём $l < a$. Эта вероятность равна $P = 2l/\pi a$. Важность этой задачи по определению вероятности пересечения рудной жилы длиной $2l$ одним из профилей площадной геофизической или геологической съёмки совершенно очевидна. На основе задачи Бюффона И. Д. Савинским (1964 г.) были рассчитаны «Таблицы вероятностей пересечения эллиптических объектов прямоугольной сетью наблюдений», а также решены и другие задачи выбора оптимальных сетей геофизических наблюдений.

Расчёт сетей при таком подходе ориентируется лишь на геометрию объектов и основывается на предположении о том, что объекты чётко фиксируются аномалиями в физических полях. В то же время при слабой дифференциации по физическим свойствам вмещающих пород и искомых объектов последние даже при значительных размерах могут не выделять-

ся аномальными значениями поля. Для уверенного выделения аномалий от таких объектов обычно требуется накопление аномального эффекта по большому числу точек наблюдений. Таким образом, в общем случае расчёт сетей геофизических наблюдений следует проводить с учётом необходимости накопления аномального эффекта.

6.2. Случайная величина

Случайной называют величину, принимающую в результате эксперимента только одно возможное значение, причём заранее не известно, какое именно, зависящее от случайных причин, которые не могут быть учтены.

Случайные величины бывают ***непрерывные*** и ***дискретные.***

Непрерывные принимают значения на числовой оси в рабочем диапазоне прибора. Например, запись полного вектора магнитного поля в аэромагнитной станции или сейсмическая запись в аналоговых станциях.

Дискретная случайная величина принимает вполне определённые значения $x_1, x_2, \ldots, x_n$ с вероятностями $p_1, p_2, \ldots, p_n$. Все возможные n значений случайной величины при этом образуют полную группу событий, т. е.

$$\sum_{i=1}^{n} p_i = 1,$$

где n – конечно или бесконечно. Например, данные измерений физических свойств горных пород, значения физических полей, разбитых на n градаций.

При компьютерной обработке непрерывные случайные величины преобразуются в дискретные, при этом соотношения, устанавливающие связь между возможными значениями случайной величины и соответствующими им вероятностями, могут задаваться в виде таблиц, графиков или формул. Такие соотношения называются ***законами распределения случайной величины***.

Универсальной характеристикой случайной величины является функция распределения $F(x)$. Она определяет для каждого значения x на числовой оси вероятность того, что случайная величина X примет значение, которое меньше x, т. е. $F(x)=P(X<x)$. Эта функция существует как для непрерывных, так и для дискретных величин.

Функции распределения обладают следующими свойствами:

$$\lim_{x\to-\infty} F(x)=0;\ F(-\infty)=P(x<-\infty)=0;$$

$$\lim_{x\to+\infty} F(x)=1;\ F(+\infty)=P(x<+\infty)=1;$$

$$0\le F(x)\le 1.$$

$F(x_2)\ge F(x_1)$ при $x_2> x_1$ – неубывающая функция.

Вероятность того, что случайная величина X примет значение, заключённое в интервале (x_1, x_2), равна приращению функции распределения на этом интервале:

$$P(x_2<X<x_1)=F(x_2)-F(x_1).$$

Аргумент функции распределения – это значение случайной величины, для которого функция распределения принимает конкретно заданное значение. Он называется ***квантилью распределения.***

Пользуясь последним свойством функции распределения, можно проделать следующие действия:

$$P(x<X<x+\Delta x)=F(x)-F(x+\Delta x);$$

$$\lim_{\Delta x\to 0}\left[F(x+\Delta x)-F(x)\right]/\Delta x=F'(x)=f(x).$$

Функция $f(x)=F'(x)$ характеризует плотность, с которой распределены значения случайной величины в данной точке, и называется плотностью распределения. Эта не менее важная характеристика случайной величины также имеет ряд свойств:

– *f(x)* определена при тех же значениях, что и $F(x)$, за исключением тех точек, где $F'(x)$ не существует;

$$-f(x)\ge 0;\ \lim_{\Delta x\to\pm\infty} f(x)=0;\ \int_{-\infty}^{\infty} f(x)dx=1.$$

$$- p(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx.$$

Пример функции распределения случайной величины и соответствующей плотности распределения приведён на рис. 15. Ещё раз, теперь геометрически, можно проследить, что $F(X)$ есть вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки X (рис. 15, пересечение пунктирных линий). При этом плотность можно истолковать так: вероятность того, что непрерывная случайная величина примет значение, лежащее левее X, равна площади криволинейной трапеции, ограниченной осью Ox, кривой плотности $f(x)$ и прямой $x = X$ (рис. 15, голубая область).

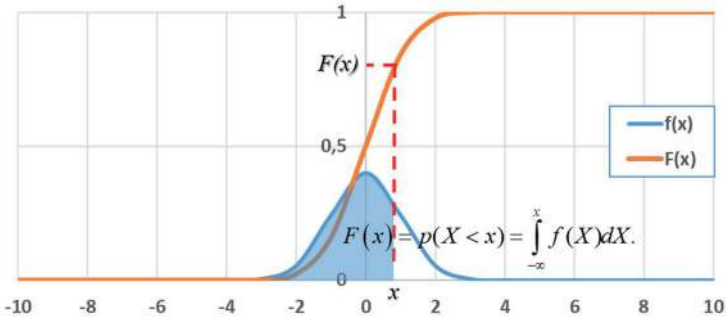


Рис. 15. Пример функции распределения случайной величины и её плотности

6.3. Описательная статистика

Описательная статистика – это традиционный набор основных статистических показателей совокупности случайных величин (или выборки).

Основные числовые характеристики случайной величины, некоторые из которых уже приводились ранее, представлены в табл. 6.

Таблица 6

Основные числовые характеристики случайной величины

Характеристика (параметр) случайной величины	***Формула для нахождения характеристики случайной величины***
γ-квантиль t_γ (квантиль порядка γ)	$P\ (x< t_\gamma)=F(t_\gamma)=\gamma\ \ t_\gamma=F^{-1}(\gamma)$, где $F^{-1}(\gamma)$ – функция, обратная функции распределения $F(x)$
Математическое ожидание MX (или среднее значение)	$MX = \bar{x} = \sum_{i=1}^{n} x_i p_i$
Мода Mo	Mo – такое значение X, при котором плотность распределения $f(x)$ максимальна
Медиана Me (или квантиль порядка 0,5)	$P(X<Me)=P(X>Me)=F(Me)=0.5$
Дисперсия DX	$DX = M(X-MX)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 p_i$
Среднее квадратичное отклонение (стандарт) σ	$\sigma = \pm\sqrt{DX}$
Коэффициент вариации V	$V = \sigma/MX = \sqrt{DX}/MX$
Центрированное и нормированное отклонение	$\tilde{x}_i = (x_i - MX)/\sqrt{DX}$
Начальный момент k-го порядка α_k	$\alpha_k = MX^k = \sum_{i=1}^{n} x_i^k p_i$
Центральный момент k-го порядка μ_k	$\mu_k = M(X-MX)^k = \sum_{i=1}^{n}(x_i-MX)^k p_i$
Асимметрия распределения A	$A = \mu_3/\sigma^3$
Эксцесс распределения E	$E = \mu_4/\sigma^4 - 3$

Все указанные параметры случайной величины используются при обработке экспериментальных данных. Приведём краткое описание некоторых из них.

Среднее значение, математическое ожидание. Из рассмотренных статистических показателей среднее значение, медиана и мода характеризуют центр статистического распреде-

ления, являясь мерами положения. Выборочное среднее значение, или математическое ожидание (арифметическое среднее), – наиболее часто применимая характеристика центра распределения случайной величины. Математическое ожидание можно рассматривать как центр тяжести распределения:

$$MX = \overline{x} = \sum_{i=1}^{n} x_i p_i.$$

В анализе данных применяются и другие виды средних, краткий обзор которых приводится в разделе 5 учебного пособия (см. с. 32–37).

Медиана – числовая характеристика непрерывно распределённой случайной величины, определяемая условием, что случайная величина с вероятностью 0,5 принимает значения как большие, так и меньшие медианы. Медиана разделяет выборку на две равные по числу значений части. Медиана – это число, которое является серединой множества чисел, т. е. половина чисел имеют большие значения, чем медиана, а половина чисел – меньшие значения, чем медиана.

Мода – числовая характеристика распределения случайной величины – точка максимума эмпирической функции распределения. Это наиболее часто встречающееся или повторяющееся значение в массиве или интервале данных. Как и медиана, мода является мерой взаимного расположения значений. Формула для вычисления медианы, моды и простейшие примеры приводятся в разделе 5 учебного пособия (см. с. 32–37).

Среднее квадратичное (стандартное) отклонение – это мера того, насколько широко разбросаны точки данных относительно их среднего. Используется как мера качества статистических оценок:

$$\sigma = \pm\sqrt{D(X)} = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 p_i}.$$

Дисперсия характеризует степень отклонения случайных величин данной совокупности от среднего в абсолютных числах, иначе меру разброса (расстояния) распределения относительно среднего значения, и равна квадрату ***среднеквадратичного отклонения:***

$$\sigma^2 = D(X) = M(X - M(X))^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 p_i.$$

Коэффициент вариации представляет собой характеристику рассеяния распределения вероятностей случайной величины Он имеет важное значение для установления степени выравненности совокупности по тому или иному признаку. Кроме того, он применяется при проверке репрезентативности (достаточности объёма) выборки.

$$V = \sigma / M(X) = \sqrt{D(X)} / M(X).$$

Моментами распределения называют средние значения степеней отклонений значений выборки:

– от среднего значения, тогда моменты называются центральными

$$\mu_k = M(X - MX)^k = \sum_{i=1}^{n} (x_i - MX)^k p_i;$$

– от произвольного числа C (например, медианы), тогда моменты называются условными;

– от нуля, тогда моменты называются начальными.

Условные и начальные моменты иначе называются нецентральными. Порядок момента равен степени k, в которую возводятся отклонения. На практике моменты выше 4 используются редко.

Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение. Отрицательный эксцесс обозначает относительно сглаженное распределение. ***Эксцесс*** находит применение в различных способах оценки отклонения распределения от нормальности

$$E = \mu_4 / \sigma^4 - 3.$$

Асимметрия характеризует степень несимметричности распределения относительно его среднего. Положительная асимметрия указывает на отклонение распределения в сторону положительных значений. Отрицательная асимметрия указывает на отклонение распределения в сторону отрицательных значений.

$$A = \mu_3 / \sigma^3.$$

Перечисленные характеристики описывают случайную величину и могут быть вычислены как с помощью приведённых формул, так и с использованием функций и модулей MS Excel. Разберём возможности Excel на примере данных, приведённых на рис. 16, столбец A, в примере Ex2_task.xlsx (https://disk.yandex.ru/i/2PEhbS7rr_HX5A). В столбце C приводится соответствующая статистическая характеристика, а напротив её – значение в столбце C для случайной величины *x*. Далее, в каждой строке, приводятся функции MS Excel, которые позволяют вычислить соответствующее значение. В некоторых случаях необходимо выбрать случай генеральной совокупности или выборки (СТАНДОТКЛОН.Г или СТАНДОТКЛОН.В) в зависимости от используемого набора данных.

	A	B	C	D	E
1	***Случайная величина x***		***Характеристики случайной величины***		***Функции Excel***
2	2,96		Среднее	3,026923	СРЗНАЧ
3	2,96		Медиана	3,03	МЕДИАНА
4	3		Мода	2,96	МОДА
5	3,03		Стандартное отклонение	0,047131	СТАНДОТКЛОН
6	3,03		Дисперсия	0,002221	ДИСП
7	2,98		Эксцесс	-0,28217	ЭКСЦЕСС
8	3,04		Асимметрия	0,449626	СКОС
9	3				
10	3,12				
11	3,1				
12	3,04				
13	3,07				
14	3,02				

Рис. 16. Описательная статистика – основной набор статистических показателей случайной величины и соответствующие функции Excel

Помимо функций в MS Excel можно воспользоваться надстройкой «Анализ данных», которую мы уже активировали при построении гистограммы в п. 4, где есть возможность создания одномерного статистического отчёта, содержащего информацию о центральной тенденции и изменчивости входных данных. Чтобы создать отчёт, необходимо перейти на вкладку

«Данные», кликнуть левой кнопкой мыши по пункту «Анализ данных» и в появившемся окне выбрать из «Инструментов анализа» пункт «Описательная статистика». В результате возникнет соответствующее диалоговое окно (рис. 17), где нужно выбрать необходимые элементы.

В области «Входные данные» нужно настроить: «Входной интервал» – это ссылка на диапазон, содержащий анализируемые данные, которые расположены по строкам или столбцам. Мы используем тот же пример Ex2_task.xlsx, тогда данные будут расположены по столбцам, а входной интервал $A$1:$A$14. При этом в указанный диапазон входит текстовый заголовок набора данных, поэтому нужно поставить галочку в поле «Метки в первой строке». В этом случае заголовок будет выведен в выходном интервале, иначе заголовок будет создан автоматически. Для верного анализа необходимо установить в пункте «Группирование» переключатель в положение «По столбцам» или «По строкам» в зависимости от расположения данных во входном диапазоне. В нашем случае используется одна случайная величина, расположенная в столбце.

В области «Параметры вывода» нужно настроить:

1) «Выходной интервал» – адрес верхней левой ячейки диапазона, в который будут выведены статистические показатели. Этот инструмент анализа выводит два столбца сведений для каждого набора данных. Левый столбец содержит метки статистических данных (названия), а правый столбец – статистические данные. Состоящий из двух столбцов диапазон статистических данных будет выведен для каждого столбца или для каждой строки входного диапазона в зависимости от положения переключателя «Группирование»;

2) «Новый рабочий лист» – выбирается, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя;

3) «Новая рабочая книга» – устанавливается переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге;

4) «Итоговая статистика» – выбирается для получения в выходном диапазоне по одному полю для каждого из следующих видов статистических данных: среднее, стандартная ошибка (среднего), медиана, мода, стандартное отклонение, дисперсия выборки, эксцесс, асимметричность, интервал, минимум, максимум, сумма, счёт;

5) «Уровень надёжности» – устанавливается флажок, если в выходную таблицу необходимо включить строку для уровня надёжности (или доверительный интервал). В поле введите требуемое значение. Например, значение 95 % вычисляет уровень надёжности среднего со значимостью 0.05. Этот параметр будет разобран позже;

6) «*K*-ый наибольший» и «*K*-ый наименьший» – устанавливается флажок, если в выходную таблицу необходимо включить строку для *k*-го наибольшего/наименьшего значения для каждого диапазона данных. В соответствующем окне введите число *k*. Если *k* равно 1, эта строка будет содержать максимум/минимум из набора данных.

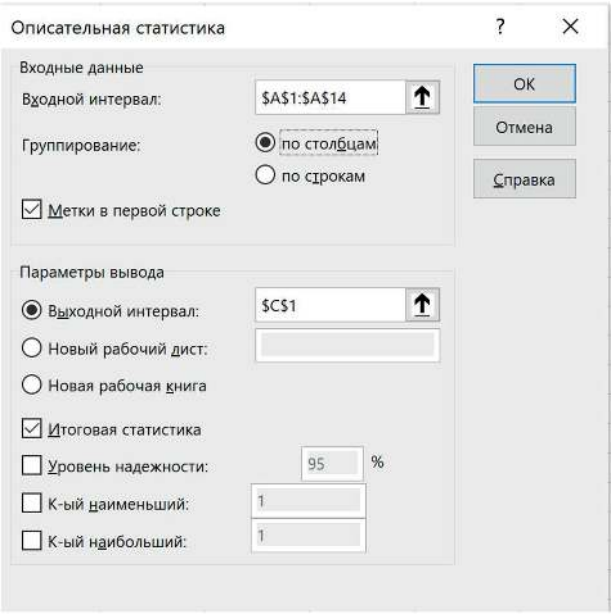


Рис. 17. Диалоговое окно «Описательная статистика»

Используя указанные настройки инструмента Excel «Описательная статистика» и данные примера Ex2_task.xlsx в столбце A (рис. 16), получится результат, представленный на рис. 18.

	A	B	C	D
1	***Случайная величина x***		***Случайная величина x***	
2	2,96			
3	2,96		Среднее	3,026923
4	3		Стандартная ошибка	0,013605
5	3,03		Медиана	3,03
6	3,03		Мода	2,96
7	2,98		Стандартное отклонение	0,049055
8	3,04		Дисперсия выборки	0,002406
9	3		Эксцесс	-0,28217
10	3,12		Асимметричность	0,449626
11	3,1		Интервал	0,16
12	3,04		Минимум	2,96
13	3,07		Максимум	3,12
14	3,02		Сумма	39,35
15			Счет	13

Рис. 18. Описательная статистика – основной набор статистических показателей случайной величины, полученный с помощью модуля «Анализ данных»

Контрольные вопросы и задания

1. Дайте определение невозможным, случайным и достоверным событиям.
2. Что такое несовместные, противоположные и независимые события?
3. Что такое сумма и произведение событий?
4. Дайте классическое определение вероятности.
5. Что такое условная вероятность?
6. Что такое полная вероятность?
7. Что определяют на основе формулы Бейеса?
8. Приведите понятие геометрической вероятности.
9. Дайте определение случайной величины.

10. Что такое непрерывная и дискретная случайная величина?
11. Что такое описательная статистика?
12. Что такое медиана и мода?
13. Что такое среднеквадратичное отклонение и дисперсия?
14. Что такое эксцесс и асимметрия?

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.

2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

3. Розенцвайг А. К., Исавнин А. Г. Статистика. Сводка и группировка данных статистического наблюдения: учеб.-метод. пособие. Набережные Челны: Изд-во Набережночелнинского института КФУ, 2019. 29 с.

4. Савинский И. Д. Таблицы вероятностей подсечения эллиптических объектов прямоугольной сетью наблюдений. М.: Недра, 1964. 86 с.

5. Статистические характеристики процессов. URL: https://www.moodle.kstu.ru/pluginfile.php/383238/mod_resource/content/1/ЦМХТП_Т2_Статистические%20характеристики%20процессов_ЛР4.pdf (дата обращения: 17.02.2023). Текст: электронный.

7. Некоторые законы распределения случайной величины

7.1. Формула Бернулли. Биномиальный закон распределения

В данном разделе разберём наиболее часто встречающиеся на практике законы распределения, два из которых – для дискретных случайных величин, остальные – для непрерывных.

Если производится несколько испытаний (опытов), причём вероятность события A в каждом испытании не зависит от исходов других испытаний, то такие испытания называются независимыми относительно события A.

В схеме Бернулли рассматривается серия, состоящая из n независимых испытаний, каждое из которых имеет лишь два исхода: наступление какого-то события A (успех) или его не наступление $\overline{A}$ (неудача). При этом вероятность успеха при одном испытании $P(A)=p$ постоянна и не зависит от номера испытания. Тогда вероятность неуспеха $P(\overline{A})=1-p=q$ тоже постоянна. Вероятность того, что при n испытаниях событие A осуществится ровно k раз, следовательно, не осуществится $(n-k)$ раз описывается ***формулой Бернулли и называется биномиальным законом распределения***:

$$P_n(k)=C_n^k p^k q^{n-k}=\frac{n!p^k q^{n-k}}{k!(n-k)!}.$$

Для графической иллюстрации приведём несколько случаев биномиального закона распределения (рис. 19). С увеличением вероятности успеха среднее случайной величины $\overline{x}=np$ возрастает вместе с дисперсией $\sigma^2=npq$, что можно отследить на графиках: пик плотности смещается, а сам график выполаживается.

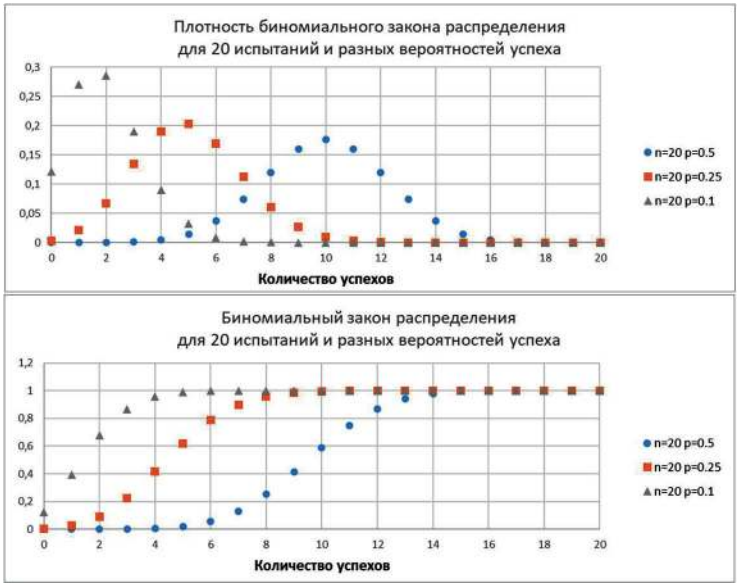


Рис. 19. Графики плотности и вероятности биномиального распределения для разных параметров

Графики этих распределений и соответствующие плотности можно вычислить по приведённой формуле или с использованием функций Excel (пример Ex3_task.xlsx, https://disk.yandex.ru/i/Pd0TSq9eK4rmPg). Для этого нужно задать столбец с количеством успехов от 0 до 20, соответственно, общее количество попыток/испытаний $n=20$. Тогда для построения графика плотности биномиального закона можно использовать приведённую формулу, а график закона можно получить, суммируя все значения плотности, полученные до заданного количества успехов. Такие же графики можно получить, используя встроенную функцию MS Excel БИНОМ.РАСП(k; n; p; ЛОЖЬ/ИСТИНА). Если указать ЛОЖЬ, то рассчитается плотность распределения, если ИСТИНА – вероятность.

7.2. Распределение Пуассона. Редкие события

Если количество испытаний достаточно велико ($n>30$), а вероятность появления события в отдельно взятом испытании p весьма мала (0,1 и меньше), то вероятность того, что в данной серии испытаний событие появится ровно k раз, можно приближенно вычислить по ***формуле Пуассона***:

$$P_\lambda(k)=\frac{\lambda^k}{k!}e^{-\lambda},\quad \lambda=np.$$

Графики плотностей и вероятностей распределения Пуассона для различных λ приводятся на рис. 20.

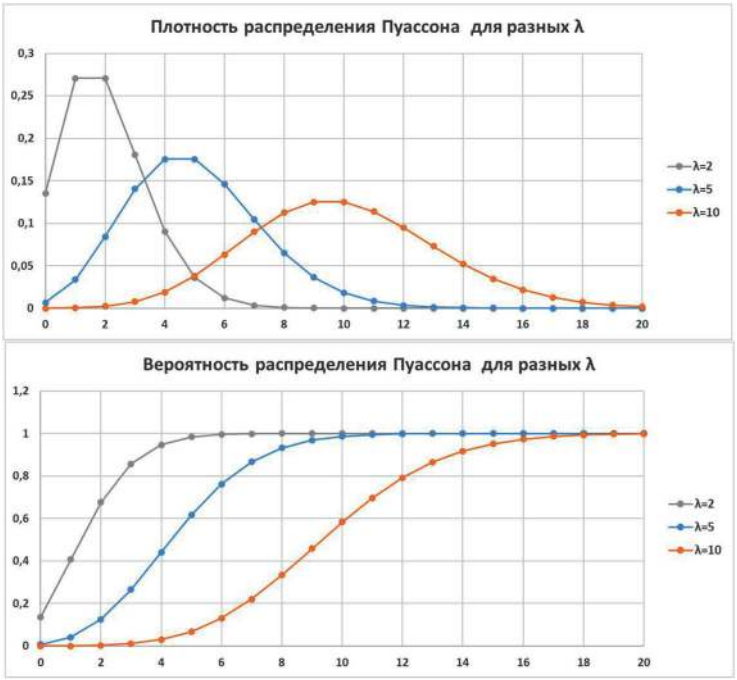


Рис. 20. Графики плотности и вероятности распределения Пуассона для разных параметров

Аналогично прошлому примеру эти графики можно построить, используя приведённую формулу или встроенную функцию MS Excel ПУАССОН.РАСП(*k*; *пр*; ЛОЖЬ/ИСТИНА). Если указать ЛОЖЬ, то рассчитается плотность распределения, если ИСТИНА – вероятность.

Можно проверить, что для больших $n>30$ распределение Пуассона и биномиальное совпадают, при этом вычисление по формуле Пуассона значительно проще.

Продемонстрируем обе формулы на задаче со следующими условиями:

На 1кг в среднем приходится 1 г полезного материала. Найти вероятность того, что если мы обработаем 9 кг, то получим 10 г полезного вещества.

Если привести всё к граммам, то вероятность $p=0{,}001$, соответственно, $q=0{,}999$. Для биномиального распределения будет формула с возведением в большие степени, что может вызвать в некоторых случаях затруднения:

$$P_{9000}(10)=C_n^k p^k q^{n-k}=C_{9000}^{10}\,0{,}001^{10}0{,}999^{9000-10}.$$

В данном случае мы можем применить формулу Пуассона, т. к. $n=9000$. Тогда без потери качества мы сможем значительно упростить вычисления:

$$P_9(10)=\frac{\lambda^k}{k!}e^{-\lambda}=\frac{9^{10}}{10!}e^{-9}.$$

Можно назвать много примеров, где используется формула Пуассона. Например, по закону Пуассона распределены автомашины на шоссе вдали от светофоров, капли дождя на асфальте (или шляпе), опечатки в книге, бактерии на питательной среде, моменты поломки сложных приборов, число посетителей в системе массового обслуживания, звезды в старых шаровых скоплениях, число радиоактивных распадов в куске радиоактивного вещества и т. д.

Несмотря на элементарность формулы Бернулли, как мы убедились ранее, при большом числе испытаний вычисление по ней может быть связано с вычислительными погрешностями. Соответственно, помимо формулы Пуассона разрешить эту проблему помогает ***локальная теорема Муавра-Лапласа:***

$$npq > 20 \Rightarrow P_n(k) = \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right), \text{ где } \varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Здесь φ(*x*) – это функция Гаусса или плотность нормального распределения.

7.3. Нормальное распределение. Распределение Гаусса

Нормальный закон распределения или закон Гаусса играет важную роль в статистике и занимает особое положение среди других законов.

Наблюдая за различными объектами и процессами окружающего мира, мы часто сталкиваемся с тем, что чего-то бывает мало, а что-то бывает нормой. Это, например, рост, вес людей, умственные способности и т. д. Из физики можно вспомнить молекулы воздуха: среди них есть медленные, есть быстрые, но большинство двигаются со «стандартными» скоростями. Для приведённых примеров и многих других природных явлений существует «основная масса» и имеются отклонения в обе стороны. Такие случайные величин *x* распределены по ***нормальному закону***, который имеет функцию плотности $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$ и определяется такими параметрами, как $\bar{x}$ – среднее нормального распределения, σ – дисперсия. При этом $\bar{x}$ определяет смещение кривой по оси X, а σ – размытость. Данная функция получила фамилию Гаусса, как и распределение. По определению функция распределения вероятностей нормального закона будет представлена следующим выражением:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-\bar{x})^2}{2\sigma^2}}dt.$$

Если $\bar{x}$ =0, а σ=1, то ***распределение*** называют ***нормальным стандартным.***

Чтобы проиллюстрировать нормальное распределение, можно в MS Excel задать столбец x значений от -10 до 10 с

равномерным арифметическим шагом, например 0,1 (пример Ex3_task.xlsx, https://disk.yandex.ru/i/Pd0TSq9eK4rmPg). В соседнем столбце можно воспользоваться формулой плотности нормального распределения или функцией Excel НОРМ.РАСП($x; \bar{x}; \sigma;$ЛОЖЬ). Для вычисления функции вероятности нужно заменить ЛОЖЬ на ИСТИНА, а в случае прямого вычисления – применить численное интегрирование.

Графики плотности и вероятности нормального распределения для различных средних и дисперсий приведены на рис. 21. Можно заметить, что от среднего зависит положение максимума распределения, а от дисперсии – размазанность, т. е чем больше, тем шире распределение.

Рис. 21. Графики плотности и вероятности нормального распределения для разных параметров

Синяя кривая на рис. 21 демонстрирует форму стандартного нормального распределения, которое используется во многих прикладных задачах, в том числе для вычисления доверительных интервалов.

7.4. Логарифмически нормальное распределение

Логарифмически нормальный (или логнормальный) закон описывает ситуацию, когда нормальному распределению подчиняются логарифмы значений случайной величины. При расчётах вначале находят ln$x$ – логарифмы значений случайной величины. Далее вся работа ведётся с логарифмами.

Случайная величина в логнормальном законе, в отличие от нормального, имеет область существования от 0 до $+\infty$. Плотность и вероятность выражаются по соответствующим формулам:

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{\left(\ln x-\overline{\ln x}\right)^2}{2\sigma^2}};\qquad F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{\left(\ln x-\overline{\ln x}\right)^2}{2\sigma^2}}d\ln t.$$

Чтобы проиллюстрировать логнормальное распределение, можно в MS Excel задать столбец x значений от 0,1 до 10 с равномерным арифметическим шагом, например 0,1 (пример Ex3_task.xlsx, https://disk.yandex.ru/i/Pd0TSq9eK4rmPg). В соседнем столбце можно воспользоваться формулой плотности логнормального распределения или функцией Excel ЛОГНОРМ.РАСП ($x;\ \bar{x};\sigma;$ЛОЖЬ). Для вычисления функции вероятности нужно заменить ЛОЖЬ на ИСТИНА.

Графики плотности и вероятности нормального распределения для различных средних и дисперсий приведены на рис. 22. Можно заметить, что от среднего зависит положение максимума распределения, а от дисперсии – размазанность, т. е. чем больше, тем шире распределение.

При малой дисперсии кривые плотности вероятности логнормального и нормального законов близки между собой, но статистические характеристики этих распределений будут

разными. Это можно наблюдать на рис. 22, где форма кривой становится схожей с плотностью нормального распределения, при уменьшении среднеквадратического отклонения (рис. 22, серая кривая).

Логарифмически нормальное распределение встречается в ряде технических задач. Оно даёт распределение размеров частиц при дроблении, содержаний элементов в минералах в изверженных горных породах, численности рыб в море и т. д.

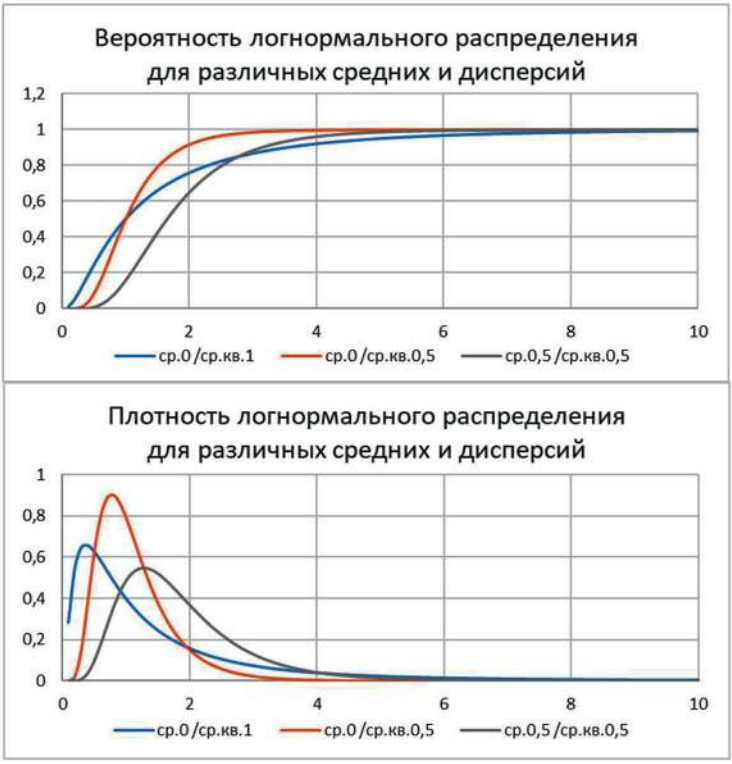


Рис. 22. Графики плотности и вероятности логнормального распределения для разных параметров

7.5. Распределение Стьюдента (*t*-распределение)

Ещё один важный закон, который следует упомянуть, – это распределение Стьюдента. В теории вероятностей и математической статистике распределение Стьюдента – семейство непрерывных одномерных распределений с одним параметром – числом степеней свободы. Формула распределения Стьюдента имеет сложный вид и выражается через функции Эйлера. Кривая плотности и вероятности распределения Стьюдента практически повторяет нормальное распределение. Отличием является то, что «хвосты» распределения Стьюдента медленнее стремятся к нулю – больше дисперсия, чем «хвосты» нормального распределения (рис. 23). При этом чем больше число степеней свободы, тем ближе распределение Стьюдента к нормальному, тем меньше дисперсия (рис. 23).

Чтобы проиллюстрировать распределение Стьюдента, зададим в MS Excel столбец x значений от -10 до 10 с равномерным арифметическим шагом, например 1 (пример Ex3_task.xlsx, https://disk.yandex.ru/i/Pd0TSq9eK4rmPg). В соседнем столбце можно воспользоваться функцией Excel СТЬЮДЕНТ.РАСП(*x*;k;ЛОЖЬ). Для вычисления функции вероятности нужно заменить ЛОЖЬ на ИСТИНА.

Приведём графики плотности и вероятности нормального распределения при разных степенях свободы и стандартное нормальное распределение (рис. 23, жёлтая кривая). Можно заметить, что чем больше степеней свободы, тем ближе распределение Стьюдента к стандартному нормальному.

Обычно распределение Стьюдента применяется в задачах, связанных с оценкой математического ожидания нормально распределённых случайных величин в условиях, когда объём выборки не велик. В анализе данных распределение Стьюдента используется для проверки гипотез о значимости моделей регрессии.

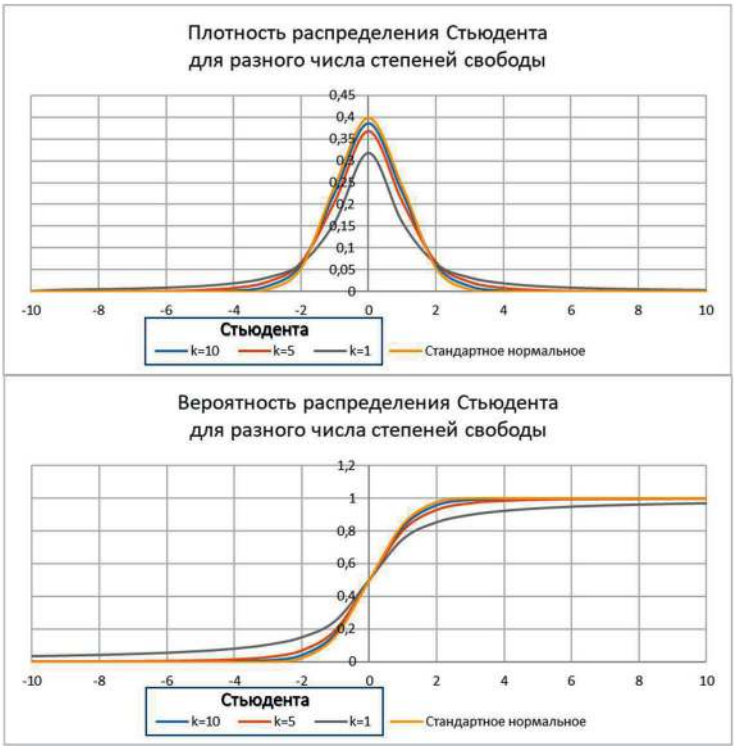


Рис. 23. Графики плотности и вероятности распределения Стьюдента для разных параметров

В задачах обработки геоданных распределения Стьюдента, как и стандартный нормальный закон, используются для оценки статистической значимости разности двух выборочных средних, при построении доверительного интервала для математического ожидания нормальной совокупности при неизвестной дисперсии, а также в линейном регрессионном анализе.

Контрольные вопросы и задания

1. Что характеризует функция распределения?
2. Что такое квантиль распределения?
3. Что характеризует плотность распределения?

4. Что такое описательная статистика?

5. Перечислите известные вам законы распределения точки. Какие из них применимы для дискретных случайных величин, а какие – для непрерывных?

6. Охарактеризуйте нормальное распределение.

7. Охарактеризуйте логарифмически нормальное распределение.

8. Охарактеризуйте распределение Стьюдента (t-распределение).

9. Чем отличаются законы распределения Пуассона и Бернулли?

10. Сформулируйте локальную теорему Муавра-Лапласа.

11. Как связаны нормальное и логнормальное распределения между собой? Когда их графики совпадают?

12. Когда распределение Стьюдента приближается к нормальному закону?

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.

2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

3. Розенцвайг А. К., Исавнин А. Г. Статистика. Сводка и группировка данных статистического наблюдения: учеб.-метод. пособие. Набережные Челны: Изд-во Набережночелнинского института КФУ, 2019. 29 с.

4. Савинский И. Д. Таблицы вероятностей подсечения эллиптических объектов прямоугольной сетью наблюдений. М.: Недра, 1964. 86 с.

8. Статистические оценки. Доверительный интервал

Не будем приводить всю теорию, связанную с доверительными интервалами. Обозначим лишь те моменты, которые могут понадобиться для обработки геоданных.

Представьте, что нам необходимо оценить долю или содержание какого-либо металла на рудном месторождении. При этом мы оперируем выборкой в количестве N образцов. Допустим, интересующая нас доля равна α. Устроит ли нас такая оценка? С одной стороны, устроит: выборка хорошая и достаточно большая. С другой стороны, какая бы выборка не была, мы не можем распространить найденное значение на всё месторождение, потому что, когда мы оцениваем долю по выборке, мы получаем значение с некоторой погрешностью. Что же мы можем в таком случае сделать? Зафиксировать ***уровень уверенности*** в наших расчётах и вместо одного значения для доли определить интервал, в пределах которого эта доля находится. Другими словами, мы можем построить ***доверительный интервал.***

Доверительный интервал – вычисленный на основе выборки интервал значений признака, который с известной вероятностью содержит оцениваемый параметр генеральной совокупности. Он является показателем точности измерений, а также показателем того, насколько стабильна полученная величина, т. е. насколько близкую величину вы получите при повторении измерений: «Мы на 95 % уверены, что Класс содержания/параметр между $-\Delta$ и Δ» или «Класс содержания/параметр находится между $-\Delta$ и Δ с 95 % вероятностью».

Построение доверительного интервала обычно выглядит следующим образом. У нас есть параметр Θ (например, доля содержания металла), который мы не знаем, но хотим оценить по выборке $\hat{\Theta}$ и делаем это с какой-то погрешностью. Мы допускаем, что при оценивании параметра по выборке, максимум, что мы можем позволить, – это отклониться от истинного значения параметра на некоторую величину, которая называется ***предельной ошибкой выборки*** ε***:***

$$\left|\Theta - \hat{\Theta}\right| < \varepsilon \Rightarrow \hat{\Theta} - \varepsilon < \Theta < \hat{\Theta} + \varepsilon \cdot$$

Значение предельной ошибки зависит от ***уровня доверия*** β, который мы выбираем. Обычно в исследованиях используется уровень доверия не менее 90 %. Что означает ***уровень доверия?*** Степень уверенности в оценках, которые мы будем получать на наших данных. Если мы будем повторять аналогичное исследование много раз, независимо друг от друга, в 95 % случаев истинное значение параметра будет попадать в доверительный интервал.

При работе с доверительными интервалами часто используют два термина. ***Уровень значимости*** α – это вероятность, с которой значение параметра не попадает в доверительный интервал. Уровень доверия $\beta = 1 - \alpha$ – это вероятность того, что доверительный интервал накрывает значение параметра. Обычно уровень значимости равен 0.01, 0.05, 0.1, что соответствует уровню доверия 0.99, 0.95, 0.9. Очень часто уровни значимости и доверия измеряются в процентах, т. е. уровень доверия 0.99 и 99 % – это одно и то же.

8.1. Правило 3σ

Рассмотрим самый простой и распространённый случай построения доверительного интервала – правило трёх сигма, которое справедливо для нормального распределения F, при этом нам известны среднее значение $\bar{x}$ и дисперсия σ случайной величины X. Тогда X с вероятностью 0.997 попадают в интервал $[\bar{x} - 3\sigma; \bar{x} + 3\sigma]$. Можно записать это в виде следующего равенства:

$$P(\bar{x} - 3\sigma < X < \bar{x} + 3\sigma) = 0{,}997.$$

Мы знаем, что вероятность того, что случайная величина X примет значение, заключённое в интервале (x_1, x_2), равна приращению функции распределения на этом интервале, поэтому:

$$P(\bar{x} - 3\sigma < X < \bar{x} + 3\sigma) = F(\bar{x} + 3\sigma) - F(\bar{x} - 3\sigma),$$

где F – нормальное распределение.

Для расчётов воспользуемся функцией MS Excel НОРМ. РАСП($\bar{x} \pm 3\sigma$; $\bar{x}$;σ; ИСТИНА). Тогда для любых значений $\bar{x}$ и σ будет справедливо равенство:

$$F(\bar{x} + 3\sigma) - F(\bar{x} - 3\sigma) = 0{,}99865 - 0{,}00135 = 0{,}9973.$$

Таким образом, мы показали, что случайная величина, распределённая по нормальному закону, с вероятностью 0.997 лежит в интервале $[\bar{x} - 3\sigma; \bar{x} + 3\sigma]$ или 99,7 % значений нормально распределённой случайной величины сосредоточены в интервале $[\bar{x} - 3\sigma; \bar{x} + 3\sigma]$ (рис. 24).

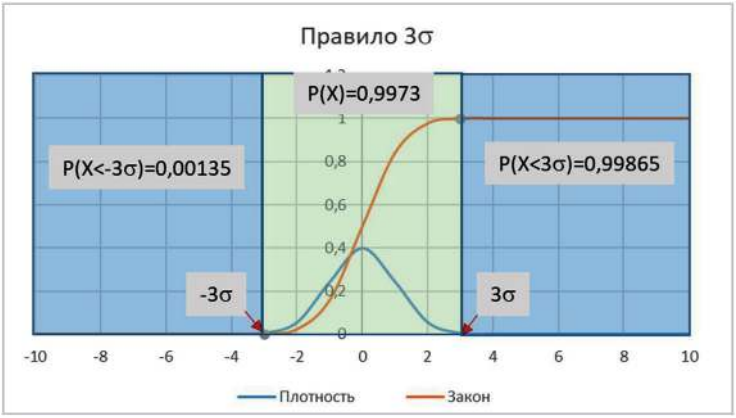


Рис. 24. Иллюстрация правила 3σ

Аналогичным образом для нормального распределения можно рассмотреть другие доверительные интервалы:

$$P(\bar{x} - \sigma < X < \bar{x} + \sigma) = 0{,}683;$$
$$P(\bar{x} - 2\sigma < X < \bar{x} + 2\sigma) = 0{,}954;$$
$$P(\bar{x} - 3\sigma < X < \bar{x} + 3\sigma) = 0{,}997.$$

Соответствующая иллюстрация найденных доверительных интервалов для нормального распределения при известных $\bar{x}$ и σ приведена на рис. 25.

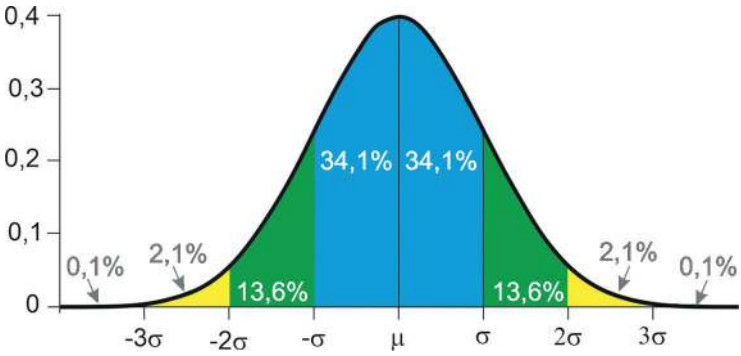


Рис. 25. Иллюстрация доверительных интервалов для нормального распределения при известных $\bar{x}$ и σ

8.2. Примеры доверительных интервалов

Далее мы часто будем предполагать, что генеральная совокупность имеет нормальный закон распределения.

Рассмотрим случай с известной дисперсией нормально распределённой генеральной совокупности. При этом необходимо построить доверительный интервал для важнейшей характеристики – среднего значения $\bar{x}$.

Рассмотрим случайную выборку объёма *n*, вычислим среднее значение по выборке и зададим уровень доверия β с соответствующим уровнем значимости *a*.

Доверительный интервал для среднего имеет вид ($\bar{x}$ – Δ; $\bar{x}$ +Δ). Если нам известно стандартное отклонение σ генеральной совокупности, тогда $\Delta = \frac{\sigma}{\sqrt{n}} z_\alpha$, где z_α – это квантиль нормального распределения уровня $1-\frac{\alpha}{2}$. Иными словами, $P(z_\alpha) = 1-\frac{\alpha}{2}$. В результате доверительный интервал для среднего с известной дисперсией имеет вид:

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_\alpha ; \bar{x} + \frac{\sigma}{\sqrt{n}} z_\alpha\right).$$

Для понимания и визуального восприятия доверительного интервала можно изобразить доверительную вероятность, которая равна площади $1-\alpha$ под графиком плотности нормального распределения (рис. 26). При этом, чтобы получить нужную величину, необходимо вычесть на хвостах распределения площади с каждой стороны, для чего и нужно найти квантиль нормального распределения $P(z_\alpha)=1-\frac{\alpha}{2}$. До недавнего времени квантили искали в таблицах нормального распределения. Мы воспользуемся функцией MS Excel НОРМ.СТ.ОБР (вероятность), которая возвращает обратное значение стандартного нормального распределения, т. е. квантиль – значение z_α для заданной вероятности.

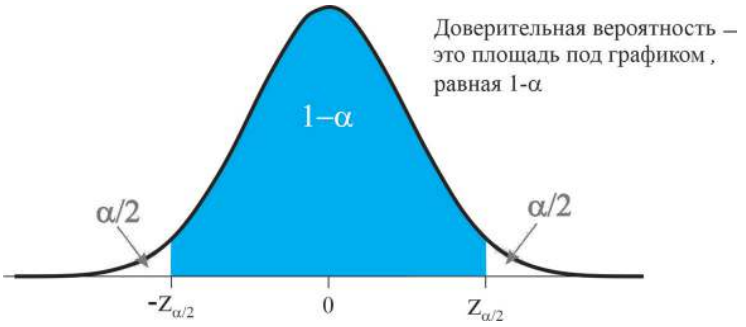
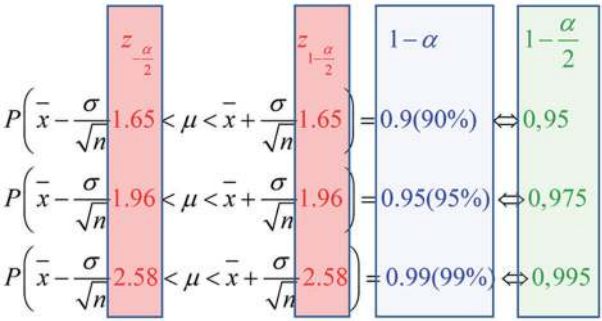


Рис. 26. Иллюстрация доверительного интервала для среднего с известной дисперсией

Разберём пример с уровнем доверия $\beta=0.9=1-0.1=1-\alpha$. Таким образом, нужно найти квантиль уровня $1-\frac{\alpha}{2}=0.95$. Используя функцию MS Excel, можно получить НОРМ.СТ.ОБР(0.95)=1.65. Ввиду симметричности распределения, НОРМ.СТ.ОБР(0.05) =–1.65. В результате доверительный интервал с уровнем доверия 0.9 для среднего с известной дисперсией будет выглядеть следующим образом:

$$\left(\bar{x}-\frac{\sigma}{\sqrt{n}}1.65;\bar{x}+\frac{\sigma}{\sqrt{n}}1.65\right).$$

Далее представлены доверительные интервалы, соответствующие площади, и уровни квантилей для β =0.9; 0.95; 0.99. Приведённые значения можно проверить самостоятельно в MS Excel.



Если выборка больше 30, но стандартное отклонение нам неизвестно, то вместо σ используется выборочное стандартное отклонение:

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}.$$

Тогда доверительный интервал для среднего при неизвестной дисперсии, но большой выборке $n>30$ будет иметь следующий вид:

$$\left(\bar{x}-\frac{s}{\sqrt{n}}z_{\alpha};\bar{x}+\frac{s}{\sqrt{n}}z_{\alpha}\right).$$

Приведённый случай, аналогичный прошлому вычислению квантилей, производится либо с использованием таблицы нормального распределения либо с помощью программных продуктов, например MS Excel и функции НОРМ.СТ.ОБР(вероятность).

Самый проблемный случай для любого исследователя – когда выборка маленькая, а про её параметры ничего неизвестно. Если дисперсия неизвестна и объём выборки небольшой ($n \leq 30$), вместо нормального распределения использует-

ся *t*-распределение. В этом случае доверительный интервал будет иметь следующий вид:

$$\left(\bar{x}-\frac{s}{\sqrt{n}}t_{\alpha}(n-1);\bar{x}+\frac{s}{\sqrt{n}}t_{\alpha}(n-1)\right).$$

Здесь $t_{\alpha}(n-1)$ – это квантиль распределения Стьюдента уровня $1-\frac{\alpha}{2}$ с $n-1$ степенью свободы. Это число можно найти в таблице *t*-распределения или воспользоваться функцией MS Excel СТЬЮДЕНТ.ОБР(вероятность; степени_свободы). Например, если нам необходимо найти по заданной выборке объёмом $n=10$ интервал уровня доверия $\beta=0.9$, то функция СТЬЮДЕНТ.ОБР$(1-\frac{\alpha}{2}; n-1=9) = 1.753$, а доверительный интервал примет вид:

$$\left(\bar{x}-1.753\frac{s}{\sqrt{n}};\bar{x}+1.753\frac{s}{\sqrt{n}}\right).$$

Как мы увидели в прошлом разделе, распределение Стьюдента стремится к нормальному распределению при $n \to \infty$, поэтому при больших выборках доверительные интервалы для среднего, посчитанные по любой из приведённых формул, будут почти совпадать.

8.3. Минимальный объём выборки

Благодаря тому, что мы знаем формулу для доверительного интервала, можно решить интересную задачу: найти минимальный необходимый объём выборки для того, чтобы с заданной точностью и уровнем доверия найти среднее значение.

Для того чтобы найти минимальный объём выборки в целях построения доверительного интервала для среднего значения с заданной точностью Δ и уровнем значимости α, можно вывести и применить формулу:

$$\Delta=\frac{\sigma}{\sqrt{n}}z_{\alpha} \Rightarrow n=\left(\frac{z_{\alpha}\sigma}{\Delta}\right)^2.$$

Контрольные вопросы и задания

1. Дайте определение доверительному интервалу.

2. Как найти минимальный объём выборки для построения доверительного интервала для известного среднего значения с заданной точностью и уровнем значимости?

3. Что такое правило 3σ?

4. Приведите примеры доверительных интервалов.

Список литературы

1. Букин В. С. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2014. 166 с.

2. Коган Р. И. Интервальные оценки в геологических исследованиях. М.: Недра, 1986. 160 с.

3. Курбацкий А. Н. Лекция 5. Доверительные интервалы. URL: https://www.mse.msu.ru/wp-content/uploads/2020/03/Лекция-5-доверительные-интервалы.pdf (дата обращения: 12.02.2023). Текст: электронный.

4. Математика и статистика (ч. 2). URL: http://www.math-info.hse.ru/2017-18/Математика_и_статистика,_часть_2 (дата обращения: 12.02.2023). Текст: электронный.

5. Ряды динамики. URL: https://www.chaliev.ru/statistics/ryady-dynamiki.php (дата обращения: 12.02.2023). Текст: электронный.

9. Ряды динамики. Анализ временных рядов

Ряд динамики – это числовые значения определённого статистического показателя в последовательные моменты или периоды времени (т. е. расположенные в хронологическом порядке).

Числовые значения того или иного статистического показателя, составляющего ряд динамики, называют ***уровнями*** ряда и обычно обозначают через y. Первый член ряда y_1 называют начальным (базисным) уровнем, а последний y_n – конечным. Моменты или периоды времени, к которым относятся уровни, обозначают через t.

В зависимости от характера изучаемого явления или процесса различают следующие виды динамических рядов: ***простой, производный, моментный и интервальный (периодический).***

Простой – ряд, составленный из абсолютных величин, характеризующих динамику одного явления. Простые ряды являются исходными данными для построения других рядов.

Производный – ряд, состоящий из средних или относительных величин.

Моментный ряд динамики – это такой ряд, уровни которого представлены рядом числовых значений, характеризующих состояние изучаемого явления или процесса на определённые моменты времени (например, на начало каждого рассматриваемого года, квартала, месяца). В этом случае каждый последующий уровень включает полностью или частично предыдущий показатель. При изучении моментного ряда определяют и исследуют разности уровней, которые характеризуют изменение/развитие изучаемого явления во времени.

В качестве примера моментного ряда можно привести динамику площадей основных видов деревьев в 1993–2015 гг. в разные моменты времени (табл. 7). Анализируя табл. 7, можно сказать, что сосны приросли сильнее всего в 2013 г., а площади, занятые елью, остались практически неизменными за 1998–2015 гг.

Интервальный (периодный) ряд динамики – это ряд числовых значений, уровни которого характеризуют размер изучаемого явления только за определённый (тот или иной, например, год, квартал, месяц и т. д.) период времени. Пример интервального ряда можно увидеть в табл. 8, где приводятся объёмы добычи газа в мире в 1970–2010 гг. за разные годы. Можно заметить, что объёмы неизменно росли за исключением 2009 г.

Таблица 7

Динамика площадей основных видов деревьев, тыс. га (Министерство природных ресурсов России)

Основные виды деревьев	***Год учёта***						
	1993	***1998***	***2003***	***2008***	***2013***	***2014***	***2015***
Хвойные							
Сосна	114326	116740	117473	116656,1	119906,1	119906,1	119259,7
Ель	75866,3	77658	77198,4	77363,9	77748,9	77706,9	77742,3
Лиственница	263348	265719	264287	175201,8	275388,2	275320,1	274827,1
Кедр	39797,9	41033,2	40852	38792,9	38893,8	38882,5	38859,9

Таблица 8

Добыча газа в мире в 1970–2010 гг.

Год	***1970***	***1980***	***1985***	***1990***	***1995***	***2000***	***2001***	***2002***	***2003***	***2004***	***2005***	***2006***	***2007***	***2008***	***2009***	***2010***
Млрд м^3	1021	1456	1676	2000	2141	2436	2493	2531	2617	2694	2778	2876	2945	3066	3045	3060

В представленных рядах изучаются абсолютные величины, но зачастую появляется необходимость изучить относительные или средние величины. ***Ряд динамики относительных величин*** – это такой ряд, уровни которого характеризуют изменение относительных размеров изучаемых явлений во времени. ***Ряд динамики средних величин*** – это такой ряд, уровни которого характеризуют изменение средних размеров изучаемых явлений во времени.

Рассчитывают показатели изменения уровней ряда динамики:

– абсолютное изменение (абсолютный прирост);

– относительное изменение (темп роста или индекс динамики);

– темп изменения (темп прироста).

Все приведённые показатели могут определяться базисным способом, когда уровень данного периода сравнивается с первым (базисным) периодом, либо цепным способом – когда сравниваются два уровня соседних периодов (рис. 27).

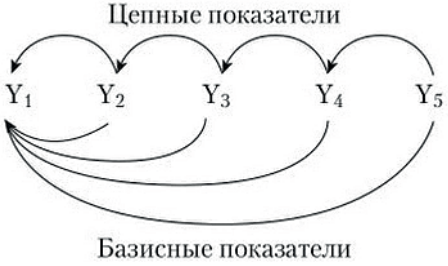


Рис. 27. Иллюстрация базисных и цепных показателей ряда динамики

Абсолютное изменение (абсолютный прирост) уровней рассчитывается как разность между двумя уровнями ряда по формулам:

Показатель	***Базисный***	***Цепной***
Абсолютный прирост	$\Delta y_i^{б}=y_i-y_1$	$\Delta y_i^{ц}=y_i-y_{i-1}$

где y_1 – значение первого уровня ряда;

y_i – уровень i-го периода.

Абсолютное изменение показывает, на сколько (в единицах показателей ряда) уровень одного (*i*-го) периода больше или меньше уровня какого-либо предшествующего периода, и, следовательно, может иметь знак «+» (при увеличении уровней) или «–» (при уменьшении уровней).

Между базисными и цепными абсолютными изменениями существует взаимосвязь: сумма цепных абсолютных изменений равна последнему базисному изменению, т. е.

$$\sum_{i=1}^{n} \Delta y_i^{ц} = \Delta y_n^{б}.$$

Относительное изменение (коэффициент роста или индекс динамики) уровней рассчитывается как отношение (деление) двух уровней ряда:

Показатель	***Базисный***	***Цепной***
Относительный прирост	$i_i^{б} = \frac{y_i}{y_1}$	$i_i^{ц} = \frac{y_i}{y_{i-1}}$

Относительное изменение показывает, во сколько раз уровень данного периода больше уровня какого-либо предшествующего периода (при $i_i > 1$) или какую его часть составляет (при $i_i < 1$). Относительное изменение может выражаться в виде коэффициентов, т. е. простого кратного отношения (если база сравнения принимается за единицу), и в процентах (если база сравнения принимается за 100 единиц) путём домножения относительного изменения на 100 %.

Между базисными и цепными относительными изменениями существует взаимосвязь: произведение цепных относительных изменений равно последнему базисному изменению, т. е.

$$\prod_{i=1}^{n} i_i^{ц} = i_n^{б}.$$

Темп изменения (темп прироста) уровней – относительный показатель, показывающий, на сколько процентов данный уровень больше (или меньше) другого, принимаемого

за базу сравнения. Он рассчитывается путём вычитания из относительного изменения 100 %:

$$T_i^ц = \frac{\Delta y_i^ц}{y_{i-1}} 100 = \frac{y_i - y_{i-1}}{y_{i-1}} 100 = \left(i_i^ц - 1\right) 100\,\%.$$

Темп роста – процентное отношение абсолютного изменения к тому уровню, по сравнению с которым рассчитано абсолютное изменение (базисный уровень):

$$T_i^б = \frac{\Delta y_i^б}{y_1} 100\,\%.$$

Каждый ряд динамики можно рассматривать как некую совокупность *n* меняющихся во времени показателей, которые можно обобщить в виде средних величин.

Обобщённой характеристикой ряда динамики может служить, прежде всего, средний уровень ряда. Способ расчёта среднего уровня зависит от того, моментный ряд или интервальный (периодный).

В случае интервального ряда его средний уровень определяется по формуле простой средней арифметической величины из уровней ряда:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Если имеется моментный ряд, содержащий *n* уровней (y_1, y_2, …,y_n) с равными промежутками между датами (моментами времени), то такой ряд легко преобразовать в ряд средних величин. При этом показатель (уровень) на начало каждого периода одновременно является показателем на конец предыдущего периода. Тогда средняя величина показателя для каждого периода (промежутка между датами) может быть рассчитана как полусумма значений *y* на начало и конец периода, т. е. как $\overline{y}_i = \frac{y_i + y_{i+1}}{2}$. Количество таких средних будет *n* – 1. Как указывалось ранее, для рядов средних величин средний уровень рассчитывается по средней арифметической. Следовательно, можно записать:

$$\bar{y} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \ldots + \frac{y_{n-2} + y_{n-1}}{2} + \frac{y_{n-1} + y_n}{2}}{n-1} = \frac{\frac{y_1 + y_n}{2} + \sum_{i=2}^{n-1} y_i}{n-1}.$$

Эта средняя известна в статистике как средняя хронологическая для моментных рядов. Такое название она получила от слова *cronos* (время, от лат.), т. к. рассчитывается из меняющихся во времени показателей.

В случае неравных промежутков между датами среднюю хронологическую для моментного ряда можно рассчитать как среднюю арифметическую из средних значений уровней на каждую пару моментов, взвешенных по величине расстояний (отрезков времени) между датами, т. е.

$$\bar{y} = \frac{\left(\frac{y_1 + y_2}{2}\right)t_1 + \left(\frac{y_2 + y_3}{2}\right)t_2 + \ldots + \left(\frac{y_{n-1} + y_n}{2}\right)t_{n-1}}{t_1 + t_2 + \ldots + t_{n-1}} = \frac{\sum_{i=1}^{n-1}(y_i + y_{i+1})t_i}{\sum_{i=1}^{n-1} t_i}.$$

В данном случае предполагается, что в промежутках между датами уровни принимали разные значения, и мы из двух известных (y_i и y_{i+1}) определяем средние, из которых затем уже рассчитываем общую среднюю для всего анализируемого периода.

Если же предполагается, что каждое значение y_i остаётся неизменным до следующего i+1-го момента, т. е. известна точная дата изменения уровней, расчёт можно осуществлять по формуле средней арифметической взвешенной:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i}.$$

Для удобства приведённые средние величины рядов динамики структурированы в табл. 9.

Таблица 9

Виды рядов динамики и соответствующие средние величины

Виды ряда динамики	*Название средней величины*	*Формула средней величины*
Равномерный интервальный	Арифметическая простая	$\bar{y} = \frac{\sum y}{n}$
Равномерный моментальный	Хронологическая простая	$\bar{y} = \frac{\frac{y_1}{2} + y_2 + y_3 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1} = \frac{\frac{y_1 + y_n}{2} + \sum_{i=2}^{n-1} y_i}{n-1}$
Неравномерный интервальный	Арифметическая взвешенная	$\bar{y} = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i}$
Неравномерный моментальный	Хронологическая взвешенная	$\bar{y} = \frac{\sum_{i=1}^{n-1} (y_i + y_{i+1}) t_i}{2\sum_{i=1}^{n-1} t_i}$

Кроме среднего уровня в рядах динамики рассчитываются и другие средние показатели – ***среднее изменение уровней ряда*** (базисным и цепным способами), ***средний темп изменения***.

Базисное среднее абсолютное изменение представляет собой частное от деления последнего базисного абсолютного изменения на количество изменений. Иными словами,

$$\overline{\Delta y}^{б} = \frac{\Delta y_n^{б}}{n-1}.$$

Цепное среднее абсолютное изменение уровней ряда – это частное от деления суммы всех цепных абсолютных изменений на количество изменений:

$$\overline{\Delta y}^{ц} = \frac{\sum_{i=1}^{n-1} \Delta y_i^{ц}}{n-1}.$$

Из известной нам формулы следует, что базисное и цепное среднее изменение должны быть равными.

По знаку средних абсолютных изменений судят о характере изменения явления в среднем: рост, спад или стабильность.

Наряду со средним абсолютным изменением рассчитывается и среднее относительное тоже базисным и цепным способами.

Базисное среднее относительное изменение определяется по формуле:

$$\overline{i}^{б} = \sqrt[n-1]{i_n^{б}} = \sqrt[n-1]{\frac{y_n}{y_1}}.$$

Цепное среднее относительное изменение определяется по формуле:

$$\overline{i}^{ц} = \sqrt[n-1]{\prod_{i=1}^{n} i_i^{ц}}.$$

Известно, что $\prod_{i=1}^{n} i_i^{ц} = i_n^{б}$. Соответственно, базисное и цепное среднее относительное изменения должны быть одинако-

выми и сравнением их с критериальным значением 1 делается вывод о характере изменения явления в среднем: рост, спад или стабильность.

Вычитанием 1 из базисного или цепного среднего относительного изменения образуется соответствующий средний темп изменения, по знаку которого также можно судить о характере изменения изучаемого явления, отражённого данным рядом динамики.

Для иллюстрации разобранных характеристик временного ряда рассмотрим официальную статистику численности населения РФ, которая размещена на сайте правительства https://rosstat.gov.ru/storage/mediabank/demo11.xls. Скачав файл, возьмём короткий период с 2011 по 2020 гг. Для такого временного ряда вычислим абсолютные и относительные изменения, средний уровень ряда и среднее изменение. На рис. 28 и в файле Ex4_task.xlsx (https://disk.yandex.ru/i/JLEpsRptSLTJBQ) приводятся упомянутые параметры временного ряда. В конце столбцов «Абсолютный и относительный цепной прирост» представлены сумма (ячейка D12) и произведение (ячейка F12) всех вышестоящих значений. Это подтверждение верности формул $\sum_{i=1}^{n} \Delta y_i^{ц} = \Delta y_n^{б}$ и $\prod_{i=1}^{n} i_i^{ц} = i_n^{б}$.

Ряд динамики можно представить в виде гистограммы, полигона, графика, как и любой из его показателей. На рис. 28 синей кривой представлены численность населения РФ и абсолютный прирост. Можно заметить, что в 2011–2018 гг. наблюдался рост населения, а наибольший прирост произошёл в 2014 г. С 2018 г. началось незначительное снижение. Аналогичные выводы можно сделать, исследуя абсолютный прирост, – пока он положительный, наблюдается рост, в последние два года прирост переходит в отрицательную зону (рис. 29, оранжевая кривая).

	A	B	C	D	E	F	G	H
1	Годы	Население РФ, млн.чел	$\Delta y_i^{Б}$	$\Delta y_i^{Ц}$	$i_i^{Б}$	$i_i^{Ц}$	$T_i^{Б}$	$T_i^{Ц}$
2	2011	142,9						
3	2012	143,0	0,1	0,1	1,0007	1,0007	0,06998	0,06998
4	2013	143,3	0,4	0,3	1,0028	1,0021	0,27992	0,20979
5	2014	143,7	0,8	0,4	1,0056	1,00279	0,55983	0,27913
6	2015	146,3	3,4	2,6	1,02379	1,01809	2,37929	1,80932
7	2016	146,5	3,6	0,2	1,02519	1,00137	2,51924	0,13671
8	2017	146,8	3,9	0,3	1,02729	1,00205	2,72918	0,20478
9	2018	146,9	4,0	0,1	1,02799	1,00068	2,79916	0,06812
10	2019	146,8	3,9	-0,1	1,02729	0,99932	2,72918	-0,0681
11	2020	146,7	3,8	-0,1	1,02659	0,99932	2,6592	-0,0681
12				3,8		1,02659		
13								
14	Среднее абсолютное изменение				0,4			
15	(в среднем растёт на...)							
16	Среднее относительное измерение				1,00292			
17	(в среднем растёт в ... раз)							

Рис. 28. Расчёты показателей динамического ряда

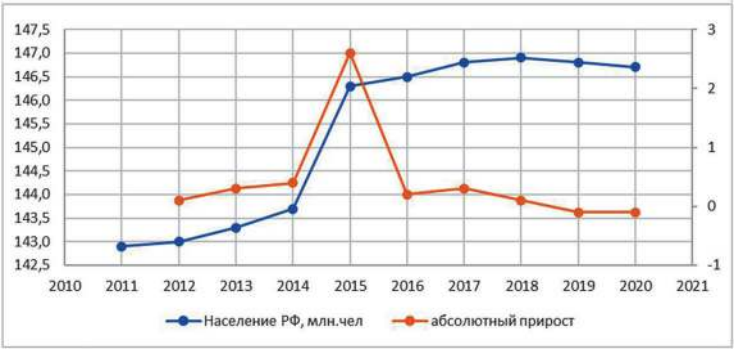


Рис. 29. Иллюстрация ряда динамики «Численность населения РФ»

9.1. Методы выявления основной тенденции (тренда) в рядах динамики

На первом этапе исследования временных рядов необходимо сделать полный анализ динамического ряда и выявить в нём основные его составляющие:

1) тренд – это основная тенденция изменения уровней, которую необходимо определить;

2) циклические колебания – это изменения уровня от момента времени, к которым можно относить часовые колебания в течение дня, дневные от дня недели или месяца и т. д.;

3) случайные колебания уровней – это случайные погрешности измерений уровня или его колебания от внешних не учитываемых воздействий.

Одна из основных задач изучения рядов динамики, как и многих других вариационных рядов, – выявить основную тенденцию (закономерность) в изменении уровней ряда, именуемую трендом. Исследование тренда динамического ряда включает два основных этапа.

1. Ряд динамики проверяется на наличие тренда.

2. Производятся выравнивание временного ряда и непосредственное выделение тренда с экстраполяцией полученных показателей – результатов.

Проверка на наличие тренда в ряду динамики может быть осуществлена по нескольким критериям.

1. ***Метод средних*** – ряд разбивается на несколько интервалов (обычно на два), для каждого из которых определяется средняя величина. Выдвигается гипотеза о существенном различии средних. Если эта гипотеза принимается, то признаётся наличие тренда.

2. Фазочастотный критерий знаков первой разности (***критерий Валлиса и Мура***). Суть его заключается в следующем: наличие тренда в динамическом ряду утверждается в том случае, если этот ряд не содержит либо содержит в приемлемом количестве фазы – изменение знака разности первого порядка (абсолютного цепного прироста).

3. ***Критерий Кокса и Стюарта.*** Весь анализируемый ряд динамики разбивают на три равные по числу уровней группы (в том случае, когда число уровней ряда не делится на три, недостающие уровни надо добавить) и сравнивают между собой уровни первой и последней групп.

4. ***Метод серий.*** Он применяется для рядов динамики с числом уровней не менее 10.

После обнаружения основной тенденции ряда определяют её характер. Для этого используют один из трёх методов:

1) простейший метод сглаживания уровней ряда – ***метод укрупнения интервалов,*** когда интервалы объединят в более крупные (например, месячные – в квартальные). Для новых интервалов рассчитывают новые уровни, усредняя старые уровни по формуле средней арифметической (для интервального ряда) или хронологической (для моментного ряда). Усреднение уровней позволяет сгладить их колебания и получить тенденцию в чистом виде;

2) ***метод скользящей средней,*** когда каждый уровень заменяют на усреднённую величину. Её получают, усредняя данный уровень и несколько уровней, расположенных симметрично справа и слева. Расчёт методом скользящей средней производится в следующем порядке:

– определяются укрупнённые периоды;

– подсчитывается среднее значение нескольких укрупнённых членов ряда (трёхлетних, пятилетних), начиная с первого, затем со второго и т. д.

Таким образом, вычисленная средняя величина как бы скользит по ряду динамики, передвигаясь на один срок;

3) ***метод аналитического выравнивания*** является наиболее совершенным методом обработки временных рядов, имеющим цель устранить случайные колебания и выявить основную тенденцию развития явления. Данный метод позволяет получить конкретное аналитическое выражение, формулу, описывающую характер изменения интересующего нас показателя во времени.

Суть аналитического выравнивания заключается в замене эмпирических (фактических, исходных) уровней y_i теоретиче-

скими, которые рассчитаны по определённому уравнению, принятому за математическую модель тренда, где теоретические уровни рассматриваются как функция времени $\overset{\circ}{y}_t = f(t)$ – систематические составляющие. При этом $y_i = \overset{\circ}{y}_i + \varepsilon_i = f(t) + \varepsilon_i$, где ε_i – случайная величина.

Задача аналитического выравнивания сводится к следующему:

– определение на основе фактических данных формы (вида) гипотетической функции *f(t)*, способной наиболее адекватно отразить тенденцию развития исследуемого показателя;

– нахождение по эмпирическим данным параметров указанной функции (уравнения);

– расчёт по найденному уравнению теоретических (выравненных) уровней $\overset{\circ}{y}_t = f(t)$.

Целью аналитического выравнивания динамического ряда является определение аналитической зависимости для его описания. Для этого по виду динамического ряда выбирают вид функции и потом находят её параметры/коэффициенты, после чего анализируют отклонение уровней, рассчитанных по уравнению, и фактических уровней ряда. Обычно используются три вида этой функции:

1) линейная функция (прямая линия), когда наблюдаются стабильные цепные абсолютные приросты уровней;

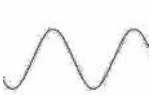
2) параболическая функция (полином второго порядка), когда сами приросты изменяются, но величина этого изменения стабильна;

3) экспоненциальная (показательная) функция, когда наблюдаются стабильные цепные коэффициенты роста уровней.

Помимо трёх упомянутых функций менее часто применяют и другие виды функций для аналитического выравнивания: показательные, степенные, гиперболические и др. Обобщение этих функций приведено в табл. 10.

Таблица 10

Наиболее часто применяющиеся виды функций для аналитического выравнивания

Название функции	***График функции***	***Формула***
Линейная		$\tilde{y}_i = \beta_0 + \beta_1 t$
Квадратичная (парабола)	 или 	$\tilde{y}_i = \beta_0 + \beta_1 t - \beta_2 t^2$
Полиномиальная степени *m*		$\tilde{y}_i = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_n t^m$
Гиперболическая		$\tilde{y}_i = \beta_0 + \frac{\beta_1}{t}$
Показательная (экспоненциальная) Степенная	 	$\tilde{y}_i = \beta_0 e^{\beta_1}$ $\tilde{y}_i = \beta_0 l^{\beta_1}$
Ряд Фурье		$\tilde{y}_i = a_0 + \sum_{k-1}^{m} (a_k \cos kt + b_k \sin kt)$

Параметры искомых уравнений/функций *(a_0, a_1, a_2, ...)* при аналитическом выравнивании могут быть определены по-разному, но наиболее распространённым методом является метод наименьших квадратов (МНК). При этом методе учитываются все эмпирические уровни и должна обеспечиваться минимальная сумма квадратов отклонений эмпирических значений уровней y_i от теоретических уровней:

$$\Phi = \sum \left(\overset{\circ}{y_i} - y_i \right)^2 \longrightarrow \min.$$

Уравнение тренда проверяют на адекватность фактическим данным по критерию Фишера.

Аналитическое выравнивание представляет собой математическую модель развития явления, соответственно, как и любая другая математическая модель, оно содержит ряд условностей. Они связаны с тем, что уровни ряда динамики рассматриваются как функции от времени.

В действительности развитие исследуемого процесса обусловлено не временем, которое изменяется от начального момента, а тем, что какие-то внешние факторы влияли на развитие процесса. Влияние времени выступает как внешнее выражение этих факторов, их суммарное воздействие. Выявить основную тенденцию с помощью аналитического выравнивания можно тогда, когда выяснены количественные и качественные изменения анализируемого параметра от определённого комплекса внешних факторов.

Стоит упомянуть о ещё одном понятии, которое необходимо знать при анализе временных рядов, – ***цикличность колебаний временного ряда.*** Если уровни ряда регулярно с определённым интервалом отклоняются от его основной тенденции, а при отсутствии этой тенденции – от среднего уровня, то это говорит о наличии циклических колебаний. Так, например, сезонно колеблется температура в течение года.

Изучение циклических колебаний осуществляют двумя способами – с помощью ***индексов сезонности*** и ***гармонического анализа.***

Порядок расчёта индексов сезонности определяется наличием или отсутствием в ряду основной тенденции и временными рамками ряда. Индекс сезонности показывает, во сколько раз уровень за данный интервал больше уровня основной тенденции, а при её отсутствии – больше среднего уровня.

Обычно индексы сезонности выражают в процентах и затем используют при построении сезонной волны. Сезонная волна – это график в виде ломаной кривой, описывающей зависимость величины индекса сезонности от номера месяца

(квартала). Она наглядно отображает характер сезонных колебаний уровней.

Другой метод моделирования сезонных и циклических колебаний основан на применении одномерных рядов Фурье и называется ***гармонический анализ.*** В свою очередь, ряды Фурье являются одной из разновидностей спектрального анализа.

С помощью спектрального анализа в структуре временного ряда определяется пик отклонений от тренда, что позволяет рассчитать длительность периодической компоненты ряда.

Для того чтобы к временному ряду можно было применять методы спектрального анализа, его необходимо привести к стационарному виду.

Суть спектрального анализа заключается в том, что случайный стационарный процесс представляется как сумма гармонических колебаний различных частот, называемых гармониками.

Спектром называется функция, которая описывает распределение амплитуд случайного стационарного процесса по различным частотам.

Сезонная компонента временного ряда может быть разложена в ряд Фурье.

Сезонные колебания, разложенные рядом Фурье, представляют собой сумму нескольких синусоидальных и косинусоидальных гармоник с различными периодами:

$$f(x) = a_0 + \sum_{n=1}^{k}\left(a_n \cdot \cos\left(\frac{\pi \cdot t}{q} + \alpha\right) + a_n \cdot \sin\left(\frac{\pi \cdot t}{q} + \alpha\right)\right).$$

Найденные индексы сезонности и уравнение сезонных отклонений могут быть использованы для прогнозирования сезонных уровней.

Далее на конкретном примере попробуем разобрать некоторые из приведённых понятий и методов.

9.2. Методы выявления основной тенденции (тренда) в температурном климатическом ряде динамики

Рассмотрим пример по обработке погоды. Архив погоды можно скачать на сайте «Специализированные массивы для климатических исследований» (http://aisori-m.meteo.ru/waisori/index1.xhtml). Здесь собраны климатические данные с начала XX в. Для получения доступа к данным необходимо зарегистрироваться. После регистрации необходимо ввести логин и пароль в окне авторизации (рис. 30).

После ввода логина и пароля появится окно с объявлениями о последних обновлениях данных, где нужно нажать на кнопку «Переход к выбору данных» (рис. 31). Загрузится «Страница выбора станций и источников данных» (рис. 32).

В окне «Раздел базы данных» выберем «Сутки», а в поле «Источник данных» – «TTTR – температура и осадки». Ниже справа в окне «Список станций» необходимо выбрать нужную метеостанцию. Выберем «Чита, Россия», кликнув по ней левой кнопкой мыши, после чего она подсветится серым цветом. Далее нужно нажать кнопку «Выбрать» и после появления станции в окне «Список выбранных станций» необходимо нажать кнопку «Дальше» (рис. 32).

В следующем окне появятся информация о разделе базы данных, источнике данных и настройки экспорта данных. В окне «Параметры для выбора» можно выбрать показатели для экспорта (рис. 33). Кликнем левой кнопкой мыши по параметру «Средняя температура воздуха», после чего она подсветится серым, далее нажмём кнопку «Выбрать». Выбранный параметр появится в окне «Параметры запроса», где по умолчанию уже находятся номер станции и дата.

Извлечение данных начнётся после нажатия кнопки «Получить результат» и появится окно «Ожидайте завершение запроса» (рис. 34). В случае успешного завершения активизируется кнопка «Получить результат». Нажав эту кнопку, система перенаправит на страницу «Результат» (рис. 34).

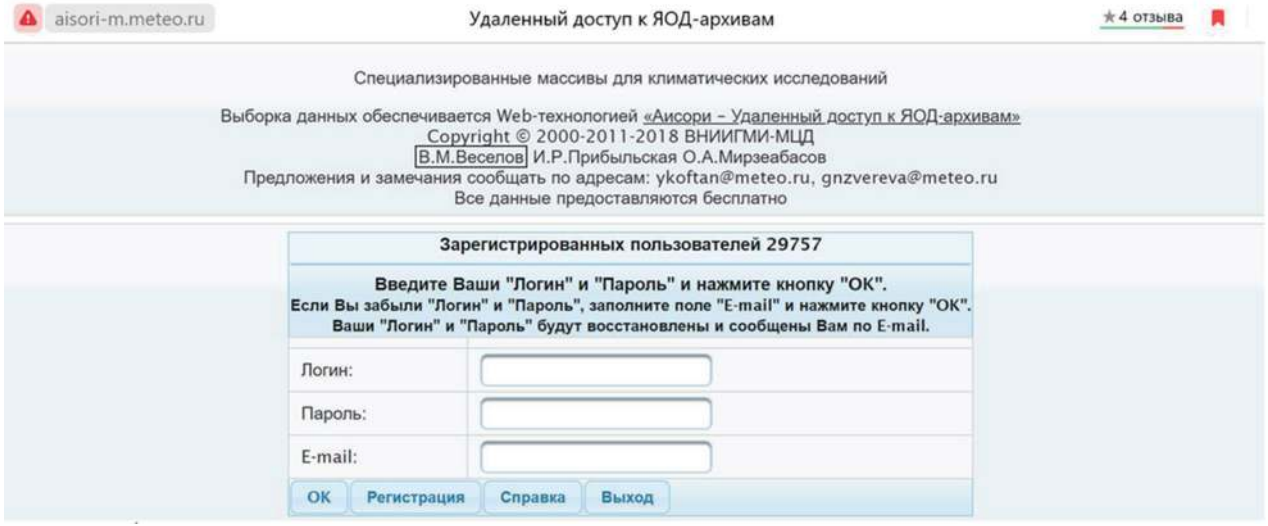


Рис. 30. Окно авторизации на сайт «Специализированные массивы для климатических исследований»

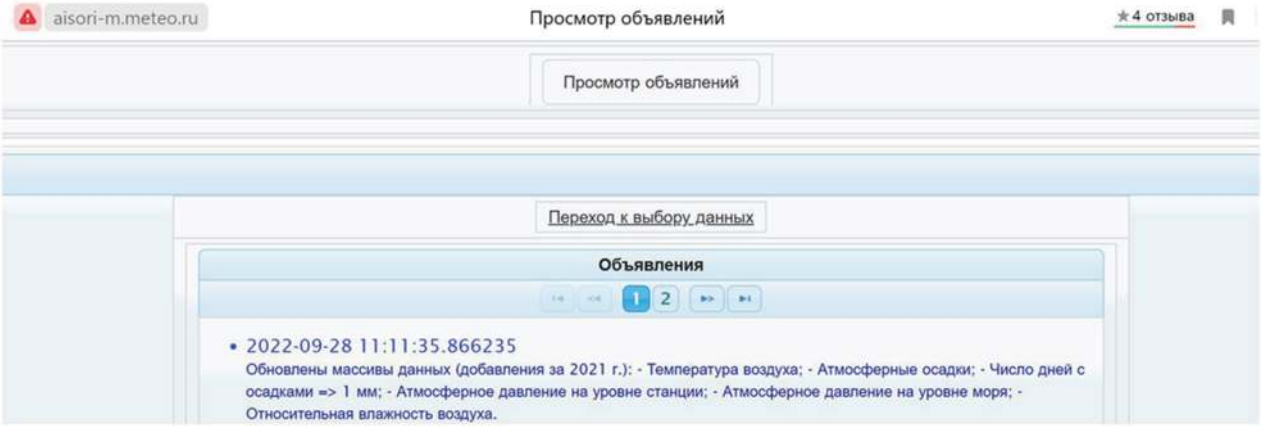


Рис. 31. Стартовая страница сайта «Специализированные массивы для климатических исследований»

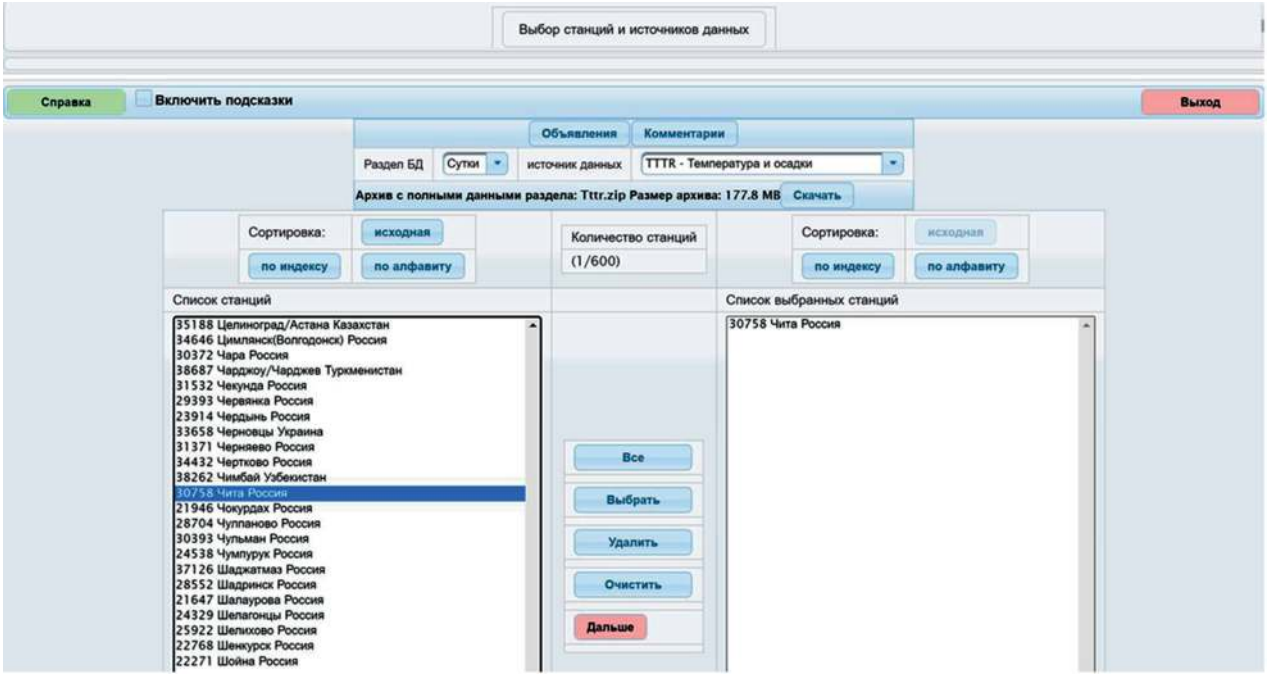


Рис. 32. Страница выбора станций и источников данных сайта «Специализированные массивы для климатических исследований»

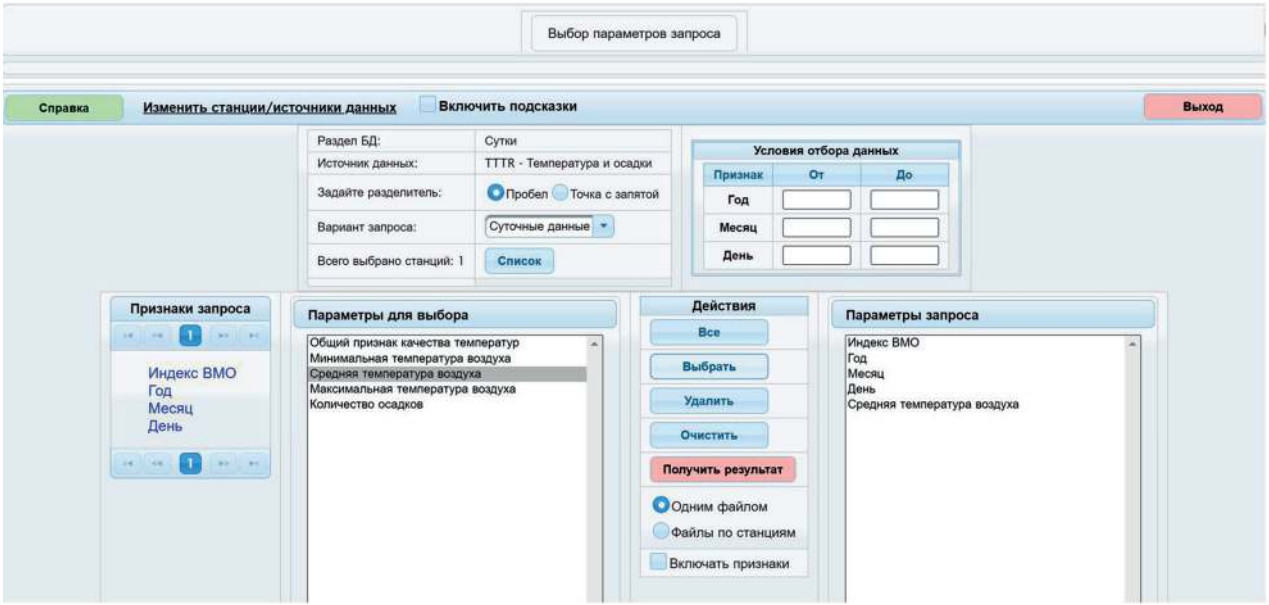


Рис. 33. Страница выбора параметров запроса на сайте «Специализированные массивы для климатических исследований»

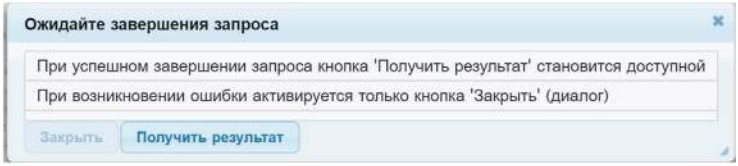


Рис. 34. Окно «Ожидайте завершение запроса»

На странице «Результат» появятся информация о данных и окно с фрагментом выбранных данных (рис. 35). Чтобы экспортировать данные, необходимо нажать кнопку «Загрузить», после чего начнётся загрузка архива, в котором будут находиться три текстовых файла.

В файле fld___.txt перечислены параметры\столбцы, выбранные и извлечённые из базы данных. В нашем случае это пять столбцов: Индекс ВМО (название станции), Год, Месяц, День, Средняя температура воздуха. В statlist___.txt приводятся название и номер метеостанции, а в файле wr___.txt – извлечённые данные.

В нашем случае в файле с данными отсутствуют средние температуры до 15 мая 1890 г. Кроме того, если использовать стандартные функции MS Excel, то нужно оперировать датами не ранее 1 января 1900 г. Таким образом, используя извлечённые данные, можно сформировать ряд MS Excel Ex5_task.xlsx (https://disk.yandex.ru/i/6FRhnkkcZ2XgqA), для которого будет проводиться статистический анализ (рис. 36).

Прежде всего, построим график среднедневной температуры за имеющийся период, т. е. с 1 января 1900 г. по 31 декабря 2021 г. Для этого воспользуемся функцией MS Excel ДАТА(год;месяц;день), аргументы которой у нас имеются в столбцах B, C, D. В результате можно сформировать два столбца для построения графика и качественного анализа данных (рис. 37, столбцы G,H). Выделив эти столбцы, построим точечную диаграмму (рис. 38). Видно, что присутствуют сезонные колебания температуры, а также в 1921 г. наблюдается разрыв данных, который необходимо учитывать при дальнейшем анализе.

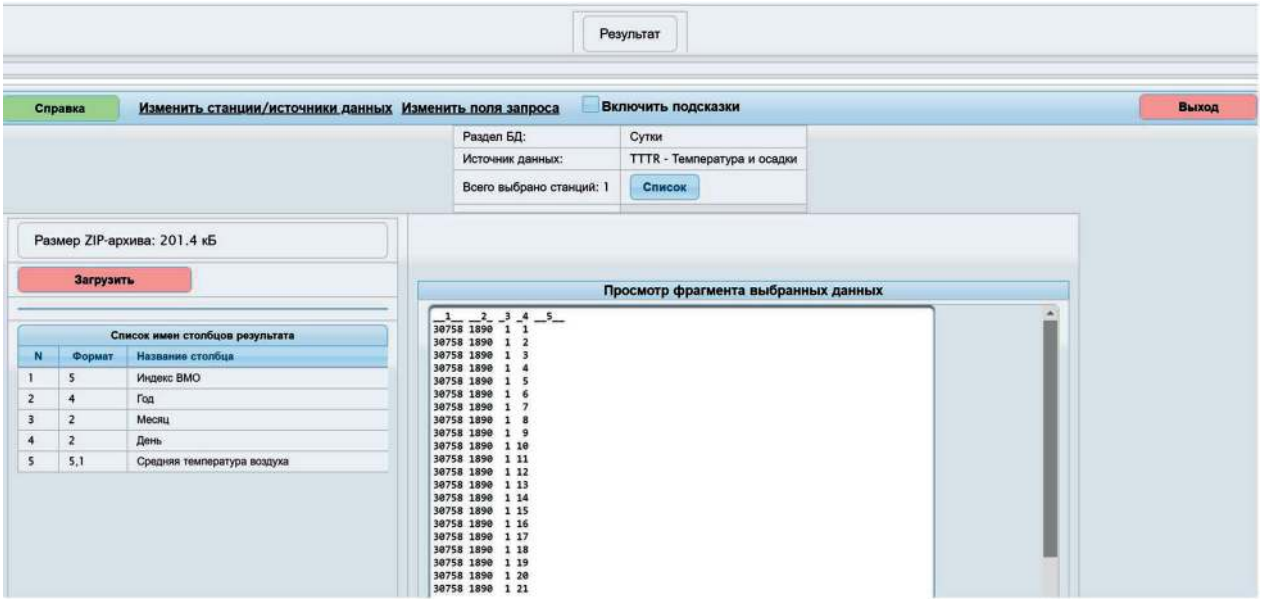


Рис. 35. Окно «Результат» на сайте «Специализированные массивы для климатических исследований»

	A	B	C	D	E
1	Индекс ВМО	Год	Месяц	День	Средняя температура воздуха
2	30758	1900	1	1	-23,7
3	30758	1900	1	2	-21,9
4	30758	1900	1	3	-28,4
5	30758	1900	1	4	-34,8
6	30758	1900	1	5	-34,9
7	30758	1900	1	6	-33,4
8	30758	1900	1	7	-26,8
9	30758	1900	1	8	-24,2

Рис. 36. Динамический ряд, сформированный для климатических температурных данных для метеостанции «Чита»

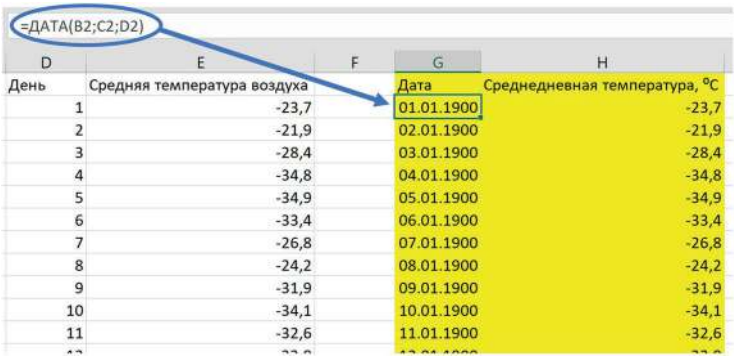


Рис. 37. Динамический ряд, сформированный для климатических температурных данных для метеостанции «Чита»

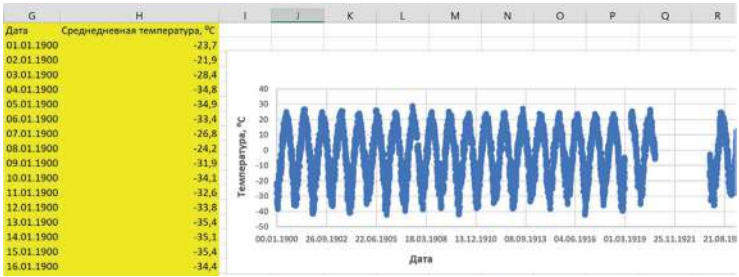
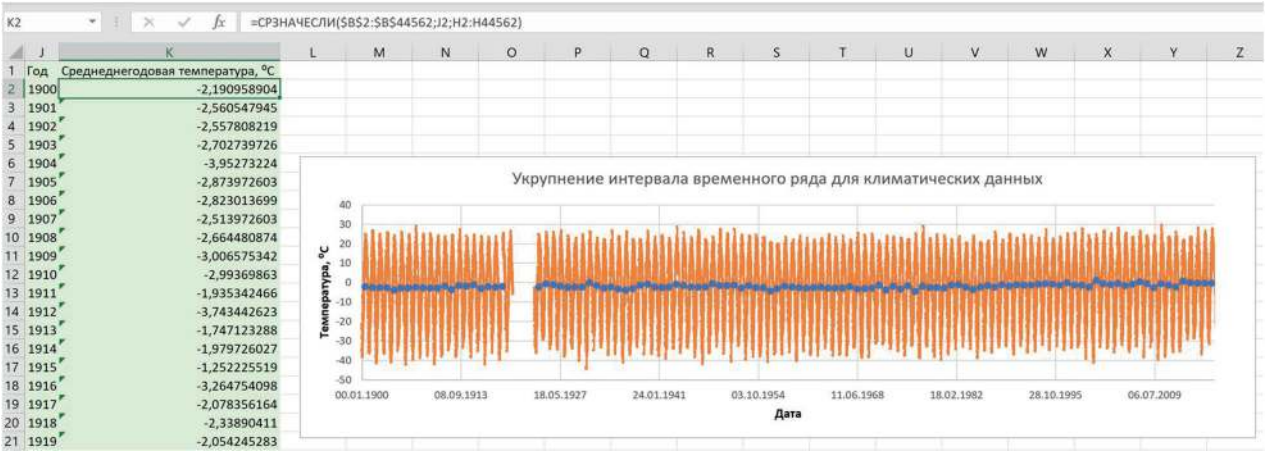


Рис. 38. Динамический ряд, сформированный для климатических данных метеостанции «Чита», и соответствующий график температуры

Как мы упоминали ранее, простейший метод сглаживания уровней ряда – ***метод укрупнения интервалов***, когда интервалы объединят в более крупные. В нашем случае можно объединить среднедневные температуры в среднегодовые. Для новых интервалов рассчитывают новые уровни, усредняя старые уровни по формуле средней арифметической. Усреднение уровней позволяет сгладить их колебания и получить тенденцию в чистом виде. Проделаем это для нашего ряда, введя столбец J, где будут представлены все года исследуемого периода.

Для сглаживания нужно воспользоваться функцией СРЗНАЧЕСЛИ($B$2:$B$44562;J26;$H$2:$H$44562), где в качестве первого аргумента вводится «диапазон» значений, для которых будем применять второй аргумент – «критерий». В нашем случае мы будем исследовать столбец B «год». Отберутся и рассчитается среднее для значений третьего аргумента функции «диапазона усреднения». В нашем случае мы усредняем значения столбца H «Среднедневная температура, °C» определённого «критерием» года. Определив среднее для 1900 г., можно распространить эту формулу для остальных годов.

Для построения среднегодовой и среднедневной температуры на одном планшете нужно определить дату среднегодовой температуры, например 01.07.XXXX. Построив среднедневную и среднегодовую температуру (рис. 39), увидим, что среднегодовая температура (рис. 39, синяя кривая) проходит примерно посередине сезонных колебаний среднедневной температуры (рис. 39, оранжевая кривая).



	J	K
1	Год	Среднегодовая температура, °С
2	1900	-2,190958904
3	1901	-2,560547945
4	1902	-2,557808219
5	1903	-2,702739726
6	1904	-3,95273224
7	1905	-2,873972603
8	1906	-2,823013699
9	1907	-2,513972603
10	1908	-2,664480874
11	1909	-3,006575342
12	1910	-2,99369863
13	1911	-1,935342466
14	1912	-3,743442623
15	1913	-1,747123288
16	1914	-1,979726027
17	1915	-1,252225519
18	1916	-3,264754098
19	1917	-2,078356164
20	1918	-2,33890411
21	1919	-2,054245283

Рис. 39. Динамический ряд, сформированный для температурных данных метеостанции «Чита», соответствующий среднедневным и сглаженный укрупнением интервалов рядов среднегодовых значений

Построим отдельно среднегодовую температуру, в зависимости от года (рис. 40, синяя кривая). Можно заметить очевидный тренд на потепление. Особенно он становится заметен, начиная с 1980 г. Можно продолжить укрупнение интервалов и рассчитать среднюю температуру за декаду. Если будем рассчитывать среднюю с 1901 по 1910 г. с использованием функции СРЗНАЧ, то привязкой будем считать 1905 г. Для последующих декад нужно использовать такой же принцип. Полученные среднедекадные значения температуры (рис. 38, оранжевая кривая) можно нанести на уже имеющуюся кривую среднегодовых значений. Хорошо видно, что сглаживаются колебания температуры, а тренд на повышение становится ещё более очевидным. Можно оценить, что с 1980 г. температура поднялась с –2.8 до –0.4 ºС.

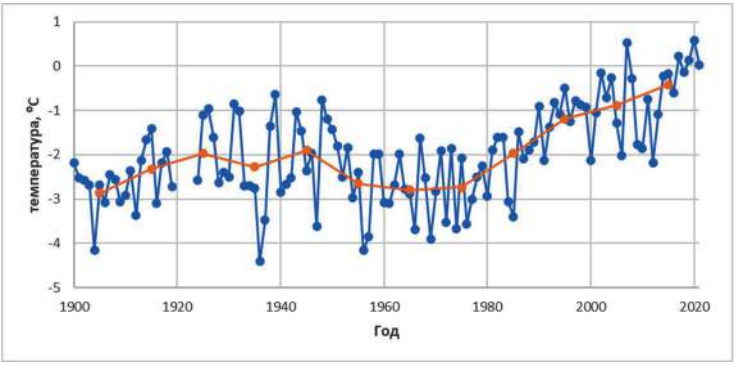


Рис. 40. Динамический ряд, сформированный для температурных данных метеостанции «Чита», соответствующий среднегодовым и сглаженный укрупнением интервалов рядов средних значений за декаду

На этих же данных можно продемонстрировать действие ***метода скользящей средней,*** когда каждый уровень заменяют на усреднённую величину. Определим укрупнённые периоды так же, как и в последнем случае, – 10 лет. Подсчитаем среднее значение температуры в 1900–1910 гг., используя функцию СРЗНАЧ(K2:K12), при этом рассчитанное значение поместим напротив 1905 г. (рис. 41). Далее нам нужно просто распространить эту формулу, кликнув по правому нижнему краю

активной ячейки с формулой, либо растянуть её, потянув за тот же край до 2016 г. Получится, что для каждого последующего года окно усреднения также будет сдвигать\скользить. Таким образом, вычисленная средняя величина как бы скользит по ряду динамики, передвигаясь на один год. Усреднённый ряд – столбец L (рис. 41, красная кривая) – можно нанести на один планшет со среднегодовой (рис. 39, синяя кривая) и усреднённой укрупнением интервалов (рис. 41, оранжевая кривая). Метод скользящей средней в данном случае практически повторяет кривую, полученную при укрупнении интервалов. Тем не менее, при использовании метода скользящей средней наблюдаются некоторые тонкости, которые повторяют колебания среднегодовых температур. Очевидно, что чем больше будет интервал усреднения, тем слаженнее будет кривая.

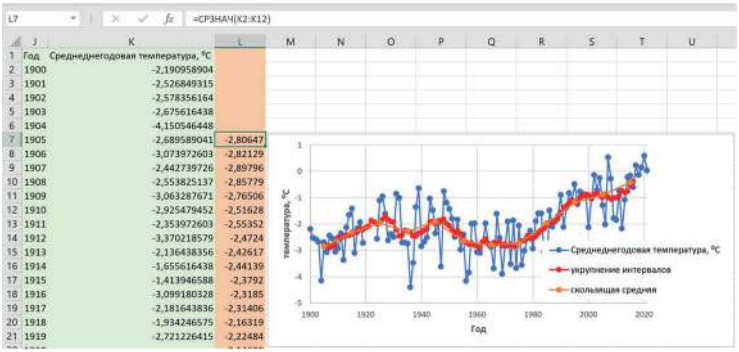


Рис. 41. Среднегодовые, среднедекадные и сглаженные методом скользящего среднего температурные данные метеостанции «Чита»

На следующем этапе мы можем рассмотреть аналитическое линейное сглаживание – линейную регрессию.

9.3. Аналитическое сглаживание. Линейная регрессия

Аналитическое линейное сглаживание – это аппроксимация линейной зависимостью нашего динамического ряда (Шеин, Потапов, 2018). Таким образом, мы можем сформули-

ровать задачу в общем виде. Даны n пар значений аргумента и функции (x_i, y_i), $i = 1, \ldots, n$. В случае динамического ряда x_i – это временные интервалы или моменты. Кроме того, известен общий вид искомой функции, которая связывает исследуемые переменные величины x и y:

$$y = f(x, A, B, C, \ldots).$$

Здесь $A, B, C, \ldots$ – неизвестные постоянные параметры, которые нужно определить, используя имеющиеся данные. Определив эти параметры, мы найдём зависимость между x и y, которая называется аппроксимирующей функцией или аппроксимантом или аналитическим сглаживанием. В случае линейного приближения $y = f(x, A, B, C, \ldots) = Ax + B$ такое аналитическое сглаживание называется ***линейной регрессией***.

Ещё раз заметим следующее: предполагается, что исходные данные известны нам с некоторой погрешностью, поэтому и искомые параметры $A, B, C, \ldots$ ищутся с некоторым приближением. Иными словами, метод аналитического сглаживания даёт нам не истинную функцию $y = f(x)$, а некоторое её приближение. Следовательно, задача аналитического сглаживания сводится к методике нахождения параметров $A, B, C, \ldots$, которые обеспечивают «наилучшее» приближение сглаживающей\аппроксимирующей функции к истинной зависимости. Критерий «наилучшего» приближения базируется на минимизации отклонений значений построенной функции $y = f(x, A, B, C, \ldots)$ в узлах x_i от соответствующих величин y_i.

Существует множество оценок качества аппроксимации и различных критериев «наилучшего» приближения. Одним из самых очевидных условий является среднее арифметическое абсолютных значений отклонений $|f(x, A, B, C, \ldots) - y_i|$, а одной из самых распространённых оценок – среднеквадратическое отклонение:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[f(x_i, A, B, C, \ldots) - y_i\right]^2}.$$

Одним из наиболее распространённых подходов для поиска функций аналитического сглаживания является ***метод наименьших квадратов***.

Не трудно представить, что уменьшение отклонения в одной из точек может привести к увеличению для другой. Практически никогда при аналитическом приближении не удаётся уменьшить до нуля все абсолютные значения отклонений. Соответственно, важно однозначно сформулировать критерий, по которому будет определяться «наилучшее» приближение искомой функции к экспериментальным данным.

Если после выбора класса аппроксимирующей функции $f(x, A, B, C,\ldots)$, которая содержит определённое количество неизвестных параметров, записать следующую сумму квадратов разностей (приближений, отклонений)

$$\Phi = \sum_{i=1}^{n} \left[f\left(x_i, A, B, C, \ldots\right) - y_i \right]^2,$$

то искомыми параметрами будут числа $A, B, C \ldots$, которые обеспечивают минимум суммы Q. Изменяя величины $A, B, C \ldots$, можно добиться уменьшения ***суммы квадратов*** до её ***минимального значения.*** Отсюда и название – ***метод наименьших квадратов***.

На этапе поиска минимизирующих значений параметров $A, B, C, \ldots$, сумма Ф рассматривается как функция от переменных $A, B, C, \ldots,$. Исходные данные (x_i, y_i), $i = 1, \ldots, n$ в этом случае являются постоянными числами.

Известно, что необходимым условием экстремума (в данном случае – минимума) функции нескольких переменных является равенство нулю всех первых частных производных. Следовательно, нужно продифференцировать сумму Q по каждому из параметров $A, B, C, \ldots$ и приравнять эти производные к нулю:

$$\frac{\partial \Phi}{\partial A} = 0; \frac{\partial \Phi}{\partial B} = 0; \frac{\partial \Phi}{\partial C} = 0; \ldots$$

При этом мы получаем систему уравнений для нахождения параметров $A, B, C, \ldots$ Важно, что количество уравнений

равно количеству неизвестных. В некоторых случаях, например если выбран полиномиальный класс функций аналитического сглаживания, получим СЛАУ, решение которой не вызывает принципиальных трудностей. Однако часто вид функции таков, что получаются системы нелинейных уравнений, которые решаются гораздо сложнее.

Вернёмся к функциям, которые представляются в виде алгебраических полиномов с действительными коэффициентами. В этом случае задача нахождения минимизирующих значений параметров *A, B, C,...* становится достаточно простой. Выбрав такой класс аналитических функций, можно записать:

$$f(x, A, B, C, ...) = P_m(x) = \sum_{j=0}^{m} a_j x^j,$$

где $m \le n$, а n – число элементов исходных данных (x_i, y_i), $i = 1, \ldots, n$.

Теперь задача сводится к вычислению таких коэффициентов полинома a_j, которые минимизируют сумму квадратов отклонений полинома в заданных узлах x_i от известных значений y_i, $i = 1, \ldots, n$.

$$\Phi = \sum_{i=1}^{n} \left[f(x_i, A, B, C, ...) - y_i \right]^2 =$$

$$= \sum_{i=1}^{n} \left[P_m(x_i) - y_i \right]^2 = \sum_{i=1}^{n} \left[\sum_{j=0}^{m} a_j x^j - y_i \right]^2.$$

Полином степени m содержит $m + 1$ коэффициент. Чтобы найти их, как уже говорилось ранее, представим a_j, $j = 0, 1, \ldots, m$ переменными и найдём частные производные суммы квадратов отклонений Ф. Для нахождения наименьшей суммы квадратов необходимо приравнять к нулю найденные производные. Получится система из $m+1$ уравнений

$$\frac{\partial \Phi}{\partial a_k} = 0; \; k = 0, 1, ..., m.$$

Эта система, в случае полиномиального класса функций аналитического сглаживания, является линейной относительно неизвестных коэффициентов a_j, минимизирующих значе-

ние выражения Ф, которое, по сути, тоже является полиномом. В явном виде эта система выглядит следующим образом:

$$\sum_{i=1}^{n}\left[y_i-\sum_{j=0}^{m}a_j x_i^j\right]x_i^k=0;\ k=0,1,\ldots,m.$$

Если преобразовать полученную СЛАУ к стандартному виду и выделить столбец свободных членов, то полученную систему можно решить с помощью любого подходящего метода решения систем линейных алгебраических уравнений.

Частным случаем аналитического выравнивания полиномами является уравнение ***линейной регрессии***, или случай линейной функции (полином степени $m=1$). Тогда

$$f(x,A,B,C,\ldots)=P_1(x)=\sum_{j=0}^{1}a_j x^j=a_0+a_1x;$$

$$\Phi=\sum_{i=1}^{n}\left[y_i-\left(a_0+a_1x_i\right)\right]^2\longrightarrow\min.$$

Чтобы отыскать минимум функции $\Phi(a_0,a_1)$, необходимо найти частные производные от функции по неизвестным a_0, a_1 и приравнять производные нулю:

$$\begin{aligned}&\frac{\partial\Phi}{\partial a_1}=-2\sum_{i=1}^{n}\left(y_i-\left(a_1x_i+a_0\right)\right)\cdot x_i=0;\\&\frac{\partial\Phi}{\partial a_0}=-2\sum_{i=1}^{n}\left(y_i-\left(a_1x_i+a_0\right)\right)=0.\end{aligned}\quad\Rightarrow\quad\begin{aligned}&-\sum_{i=1}^{n}y_ix_i+a_1\sum_{i=1}^{n}x_i^2+a_0\sum_{i=1}^{n}x_i=0;\\&-\sum_{i=1}^{n}y_i+a_1\sum_{i=1}^{n}x_i+na_0=0.\end{aligned}$$

Тогда

$$a_1=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2};$$

$$a_0=\frac{\left(\sum_{i=1}^{n}x_i^2\right)\left(\sum_{i=1}^{n}y_i\right)-\left(\sum_{i=0}^{n}x_i\right)\left(\sum_{i=0}^{n}x_iy_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}.$$

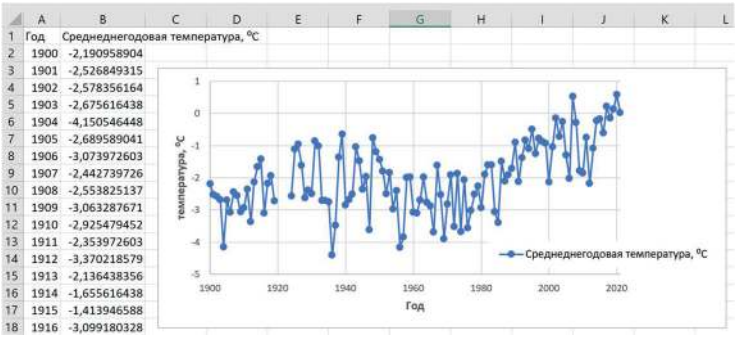
Используя выписанные коэффициенты, уже можно пользоваться линейным аналитическим выравниванием, но часто уравнение линейной регрессии записывают через статистические характеристики:

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}); x = \bar{x} + r\frac{\sigma_x}{\sigma_y}(y - \bar{y}).$$

Здесь x, y – средние значения, σ_x, σ_y – среднеквадратические отклонения, $r = \frac{K_{xy}}{\sigma_x \sigma_y}$ – коэффициент корреляции или критерий близости исходных данных к линейному тренду, $K_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ – корреляционный момент или ковариация – отражает взаимосвязь между случайными величинами x и y. Подробнее разберём указанные понятия в последующих разделах.

Приведённые формулы линейной регрессии, выраженные через статистические характеристики, эквивалентны и легко выводятся одна из другой.

Применим выведенные формулы на практическом примере – среднегодовые температуры, полученные в примере прошлого пункта, скопировав их в отдельный файл Ex6_task.xlsx (https://disk.yandex.ru/i/cVDFujMrdOiMKA, рис. 42).



	A	B
1	Год	Среднегодовая температура, °C
2	1900	-2,190958904
3	1901	-2,526849315
4	1902	-2,578356164
5	1903	-2,675616438
6	1904	-4,150546448
7	1905	-2,689589041
8	1906	-3,073972603
9	1907	-2,442739726
10	1908	-2,553825137
11	1909	-3,063287671
12	1910	-2,925479452
13	1911	-2,353972603
14	1912	-3,370218579
15	1913	-2,136438356
16	1914	-1,655616438
17	1915	-1,413946588
18	1916	-3,099180328

Рис. 42. Среднегодовые температурные данные метеостанции «Чита»

Рассчитаем статистические характеристики (СРЗНАЧ, СТАНДОТКЛОН.Г, КОРРЕЛ) для столбцов года и среднегодовой температуры и подставим в формулу:

$$y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = -1{,}98 + 0{,}522\frac{1{,}081}{35{,}06}(x - 1961{,}82) =$$

$$= 0{,}0161x - 33{,}548.$$

Применяя данное уравнение для каждого года, получим аналитическое сглаживание к временному ряду в виде линейной регрессии (рис. 43, синие маркеры).

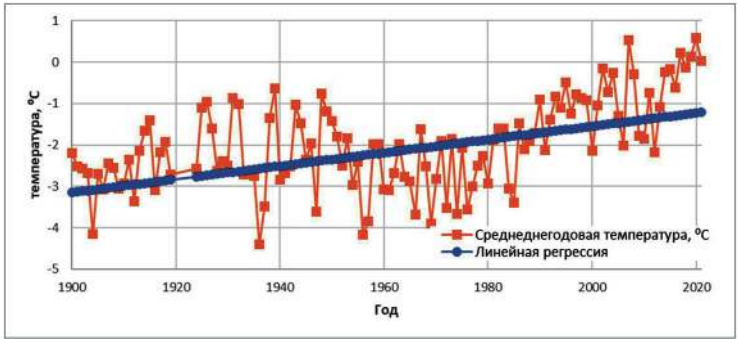


Рис. 43. Среднегодовые температурные данные метеостанции «Чита» и аналитическое сглаживание в виде линейной регрессии

Нахождение уравнения линейной регрессии можно упростить, если воспользоваться возможностями MS Excel. Для этого нужно выделить область диаграммы с построенными на ней среднегодовыми температурами, кликнув по ней левой кнопкой мыши. После выделения рядом с областью диаграммы в правом верхнем углу появятся три значка. Нужно кликом левой кнопки мыши выбрать верхний [+], который позволяет добавлять/изменять/удалять элементы на диаграмму. В появившемся списке нужно поставить галочку напротив пункта «Линия тренда». Если построено несколько графиков/рядов данных, и заранее не был выделен тот, для которого мы хотим построить линейную регрессию – найти тренд, то появится диалоговое окно, где нужно будет выбрать нужный ряд. После завершения описанных операций на графике появится линия тренда – по умолчанию линейная регрессия (рис. 44, пунктир-

ная линия). Для изменения параметров линии тренда необходимо выделить её, кликнуть по ней правой кнопкой мыши и в появившемся окне выбрать пункт «Формат линии тренда». Справа в окне Excel появится область с параметрами линии тренда. Здесь можно убедиться, что выбрана линейная регрессия 

, а также на диаграмму можно нанести уравнение линейного приближения, если активизировать соответствующий пункт 

. В результате на графике появятся линия тренда и уравнение (рис. 44). Можно проверить полученные коэффициенты линейной регрессии через статистические характеристики и с помощью линии тренда MS Excel. В обоих случаях a_0=33,548 и a_1=0,0161.

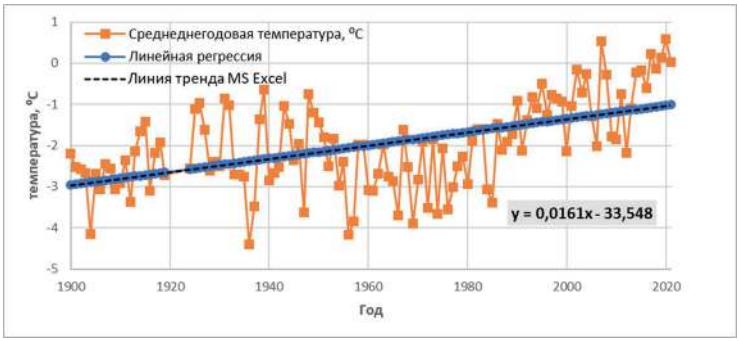


Рис. 44. Среднегодовые температурные данные метеостанции «Чита» и аналитическое сглаживание в виде линейной регрессии, построенное двумя способами

Полученную линейную зависимость, в частности, можно использовать для восстановления недостающих фрагментов данных или прогнозирования изучаемого параметра. Например, по найденному уравнению линейной регрессии для температурного ряда данных метеостанции «Чита» y=0,0161x – 33,548 можно рассчитать прогноз температуры на 2050 г. Для этого нужно подставить соответствующий год в уравнение y=0,0161 × 2050 – 33,548=–0,543 (рис. 45, красный ромб). Следует учитывать, что полученное уравнение линейной регрессии будет изменяться в зависимости от рассматриваемого временного отрезка. Очевидно, что если мы рассмотрим

метеоданные с 1980 г., где изменился тренд, а потепление стало более стремительным, то предсказание изменится. Соответственно, чаще всего рассчитывают несколько значений, и рассматривают отрезок между максимальным и минимальным прогнозами.

Существует более простой способ поиска недостающих данных или прогнозирования значений по линейному тренду, без явного поиска уравнения линейной регрессии. Для этого нужно воспользоваться функцией ТЕНДЕНЦИЯ(известные значения_y; известные значения_x; новые значения_x). В нашем случае «известные значения_y» – это столбец B с известными температурами, «известные значения_x» – это рассмотренные годы, а «новые значения_x» – 2050 прогнозный год. В результате получим температуру, равную рассчитанной по найденному уравнению в прошлом примере. Если значения отличаются, то нужно проверить и увеличить количество знаков после запятой для коэффициентов в уравнении линейной регрессии.

Рис. 45. Среднегодовые температурные данные метеостанции «Чита», аналитическое сглаживание в виде линейной регрессии и прогноз среднегодовой температуры на 2050 г.

В следующем примере проиллюстрируем восстановление истинных параметров зависимости с помощью линейной регрессии Ex7_task.xlsx (https://disk.yandex.ru/d/XQAnOyxohL-1bug, рис. 46). Для этого смоделируем ряд, где в качестве временных отсчётов возьмём года из прошлых примеров 1900–

2021 гг. (рис. 46, столбец А), а значения параметра рассчитаем по заданной линейно зависимости, например $y = 0{,}01x - 10$ (рис. 46, столбец В). Далее наложим случайный шум с помощью функции СЛУЧМЕЖДУ: к каждому элементу столбца В прибавим величину, не превышающую 5 % от среднего вычисленного ряда В2+СРЗНАЧ(В2:В119)*(СЛУЧМЕЖДУ(-$E$2;$E$2)/100). Таким образом получим временной ряд, осложнённый случайным шумом (рис. 46, синие маркеры). Теперь мы можем применить к этому ряду аналитическое выравнивание в виде линейной регрессии любым из разобранных способов: применяя формулу или используя линию тренда MS Excel. В результате получим линейное уравнение $y = 0{,}011x - 11{,}94$. Видно, что восстановленные коэффициенты a_0 и a_1 незначительно отличаются от истинных, как и сама кривая (рис. 46, оранжевая и чёрная линии). Изменяя шум в большую сторону, прямая логично будет восстанавливаться с большей ошибкой, и наоборот. Таким образом, мы показали, что линейная регрессия при достаточном количестве качественных измерений с хорошей точностью восстанавливает истинную закономерность.

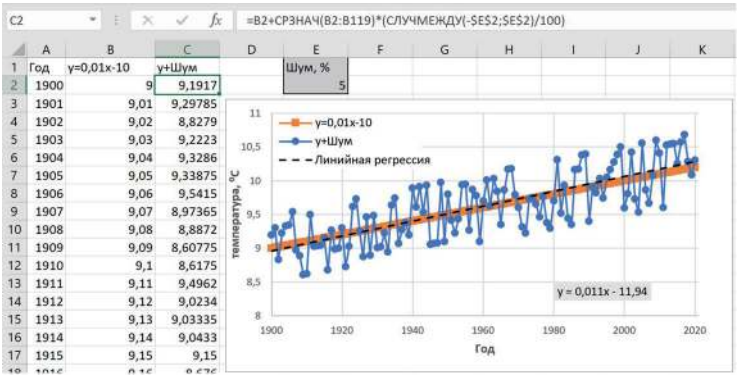


Рис. 46. Восстановление истинных параметров линейной зависимости

9.4. Нелинейная параболическая регрессия

Зависимости между свойствами могут быть не только линейными, но и более сложными – нелинейными и многофакторными. Вид аппроксимирующей функции *f(x)* должен быть задан либо на основании теоретических соображений, либо путём эмпирического подбора. В любом случае, наиболее эффективным методом для поиска коэффициентов является уже разобранный ***метод наименьших квадратов.***

Ещё одним частным случаем аналитического выравнивания полиномами является уравнение ***параболической зависимости,*** или случай квадратической функции (полином степени m=2). Тогда

$$f(x, A, B, C, ...) = P_2(x) = \sum_{j=0}^{2} a_j x^j = a_0 + a_1 x + a_2 x^2;$$

$$\Phi = \sum_{i=1}^{n} \left[y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 \right) \right]^2 \longrightarrow \min.$$

Чтобы отыскать минимум функции $\Phi(a_0, a_1, a_2)$, необходимо найти частные производные от функции по неизвестным a_0, a_1, a_2 и приравнять производные нулю:

$$\frac{\partial \Phi}{\partial a_2} = -2 \sum_{i=1}^{n} \left(y_i - \left(a_2 x_i^2 + a_1 x_i + a_0 \right) \right) \cdot x_i^2 = 0;$$

$$\frac{\partial \Phi}{\partial a_1} = -\sum_{i=1}^{n} \left(y_i - \left(a_2 x_i^2 + a_1 x_i + a_0 \right) \right) \cdot x_i = 0;$$

$$\frac{\partial \Phi}{\partial a_0} = -\sum_{i=1}^{n} \left(y_i - \left(a_2 x_i^2 + a_1 x_i + a_0 \right) \right) = 0.$$

После раскрытия скобок и преобразования получим систему трёх уравнений с тремя неизвестными:

$$a_2 \sum_{i=1}^{n} x_i^4 + a_1 \sum_{i=1}^{n} x_i^3 + a_0 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i;$$

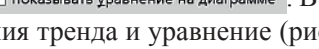
$$a_2 \sum_{i=1}^{n} x_i^3 + a_1 \sum_{i=1}^{n} x_i^2 + a_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i;$$

$$a_2\sum_{i=1}^{n}x_i^2+a_1\sum_{i=1}^{n}x_i^1+a_0\sum_{i=1}^{n}1=\sum_{i=1}^{n}y_i.$$

По аналогии с линейной регрессией проиллюстрируем восстановление истинных параметров зависимости с помощью аналитического выравнивания в виде полинома второго порядка Ex8_task.xlsx (https://disk.yandex.ru/d/XQAnOyxohL1bug). Очевидно, что все приведённые формулы будут справедливы не только для временных рядов, но и для любого вариационного ряда. Соответственно, x зададим от 2 до 8 с арифметическим шагом 0,1. Значения y рассчитаем по заданной квадратичной зависимости, например $y=0.5x^2-2x+2$ (рис. 47, столбец B). Далее наложим случайный шум с помощью функции СЛУЧМЕЖДУ: к каждому элементу столбца B прибавим величину, не превышающую заданному шуму в ячейке E2 процентов от среднего значения y: B2+СРЗНАЧ(B2:B119)*(СЛУЧМЕЖДУ(-$E$2;$E$2)/100).

Теперь у нас есть смоделированный ряд (столбцы A и C), для которого можно применить аналитическое выравнивание и восстановить параметры a_0, a_1, a_2. Для этого можно воспользоваться выведенной системой линейных уравнений, где элементами матрицы будут $\sum_{i=1}^{n}x_i^4, \sum_{i=1}^{n}x_i^3 \ldots$ Решить эту систему можно любым способом и найти вектор коэффициентов a_0, a_1, a_2.

Другой способ – это использование линии тренда в MS Excel. Нужно добавить линию тренда для соответствующих данных, как это описано в прошлом разделе, перейти в область с параметрами линии тренда и выбрать пункт «Полиномиальная» 

. Для отображения полученного нелинейного приближения активизируем соответствующий пункт 

. В результате на графике появятся линия тренда и уравнение (рис. 47). Можно проверить полученные коэффициенты полинома и сравнить с заданными изначально.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	*x*	*y=0,5x²-2x+2*	*y+Шум*		*Шум, %*		*x²*	*x³*	*x⁴*	*xy*	*x²y*		*Au=b*		*u=A⁻¹b*
2	2	0	-0,484		10		4	8	16	-0,968	-1,936				
3	2,1	0,005	0,126												
4	2,2	0,02	0,2015												
5	2,3	0,045	-0,2575										*u =*	*a2*	
6	2,4	0,08	-0,404											*a1*	
7	2,5	0,125	-0,117											*a0*	
8	2,6	0,18	-0,3645												
9	2,7	0,245	0,487												
10	2,8	0,32	0,441												
11	2,9	0,405	0,889												
12	3	0,5	0,863												
13	3,1	0,605	0,121												
14	3,2	0,72	1,325												
15	3,3	0,845	1,0265												
16	3,4	0,98	1,4035												
17	3,5	1,125	1,0645												
18	3,6	1,28	1,4615												
19	3,7	1,445	0,9005												
20	3,8	1,62	1,741				14,44	54,872	208,5136	6,156	23,14004				

Рис. 47. Восстановление истинных параметров нелинейной параболической зависимости

9.5. Оценка адекватности (надёжности) тренда

Одна из основных задач изучения рядов динамики, как и многих других вариационных рядов, – выявить основную тенденцию (закономерность) в изменении уровней ряда, именуемую трендом. Однако мало найти зависимость, нужно оценить адекватность (надёжность) найденного аналитического приближения.

Для найденного уравнения тренда необходимо провести оценку его надёжности (адекватности), что осуществляется с помощью критерия Фишера, сравнивая его расчётное значение F_p с теоретическим (табличным) значением F_T. При этом расчётный критерий Фишера определяется по формуле:

$$F_p = \frac{(n-k)\sum_{i=1}^{n}\left(y_i^t - \overline{y^t}\right)^2}{(k-1)\sum_{i=1}^{n}\left(y_i^t - y_i\right)^2},$$

где k – число неизвестных параметров (членов) выбранного уравнения тренда;

n – объём выборки.

Теоретический критерий Фишера ищется из таблицы для разных уровней значимости, но сейчас более простой способ – использование MS Excel и функции для поиска квантилей F-распределения F.ОБР (вероятность, v_1, v_2).

Сравнение расчётного F_p и теоретического F_T значений критерия Фишера ведётся при заданном уровне значимости с учётом степеней свободы $v_1 = k - 1$, $v_2 = n - k$.

При условии $F_p > F_T$ считается, что выбранная математическая модель ряда динамики адекватно отражает обнаруженный в нём тренд.

Продемонстрируем применение критерия Фишера, используя прошлый пример, где мы восстанавливали параметры истинной параболической зависимости. При этом рассмотрим для зашумлённых значений линейный тренд и критерий адекватности для него, а затем параболическое приближение, оценим её адекватность. Будем использовать уже имеющиеся на-

работки Ex8_task.xlsx (https://disk.yandex.ru/d/XQAnOyxohL-1bug), но вычисления продолжим на «Лист2», где в качестве x зададим значения от -10 до 12 с арифметическим шагом 1. Значения y рассчитаем по заданной квадратичной зависимости, например $y=0.5x^2-2x+2$ (рис. 48, столбец B). Далее прибавим к y случайный шум с помощью функции СЛУЧМЕЖДУ, как в прошлом примере.

Для полученного ряда можно найти линейный и параболический тренд: $y=-1{,}0146x+23{,}325$ и $y=0{,}5084x^2-2{,}0315x+1{,}4625$. Используя эти уравнения, можно рассчитать $y_i^t, \overline{y^t}$, а потом расчётное значение критерия адекватности F_p. В представленном примере $n=23$, $k=2$ в случае линейного тренда $k=3$ в случае квадратического приближения. Теоретическое значение критерия можно вычислить, используя функцию F.ОБР (вероятность, v_1, v_2), где $v_1=k-1$, $v_2=n-k$, а вероятность – это уровень адекватности нашей регрессии, например 0,95 означает, что с вероятностью 95 % найденный тренд адекватен. В результате мы получим, что для линейного тренда $F_p=3{,}586<F_T=4{,}414$, что говорит о неправильно найденном тренде. Для параболического приближения $F_p=1\,257$, $F_E=3{,}593$, что говорит об адекватности тренда.

Следует сказать о том, что часто можно получить адекватность линейного тренда для очевидно нелинейной зависимости. Соответственно, в любом случае решение остаётся за интерпретатором, геофизиком, геохимиком, аналитиком – за человеком.

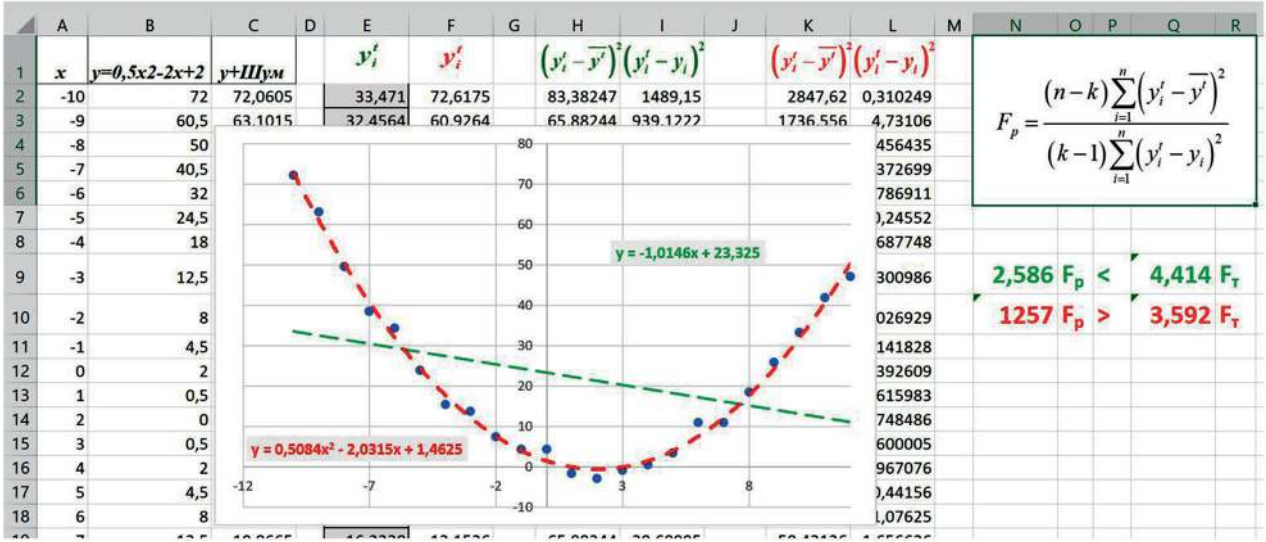


Рис. 48. Оценка адекватности линейного и параболического тренда

9.6. Анализ сезонных колебаний

В рядах динамики, уровни которых являются месячными или квартальными показателями, наряду со случайными колебаниями часто наблюдаются ***сезонные колебания,*** под которыми понимаются периодически повторяющиеся из года в год повышение и снижение уровней в отдельные месяцы или кварталы. В качестве простейшей иллюстрации рядов с сезонными колебаниями могут служить данные температуры воздуха на метеостанциях (рис. 38–39).

Измерение сезонных колебаний в статистике производится путём вычисления индексов сезонности, которые чаще всего исчисляются следующим образом:

– определяются абсолютные значения уровней ряда;

– исчисляется средне-месячный/квартальный… уровень ряда методом вычисления простой средней арифметической величины;

– определяются индексы сезонности путём сопоставления абсолютных уровней ряда со средним уровнем.

Существуют две различные модели сезонности: ***аддитивная*** и ***мультипликативная***.

В ***аддитивной модели*** сезонность выражается в виде абсолютной величины (тонны, градусы, сопротивления …), которая добавляется или вычитается из среднего значения ряда, чтобы выделить показатель сезонности. Уровень такого ряда можно представить следующим образом:

$$y_i - \bar{y} = (\bar{y}_s - \bar{y}) + (y_i - \bar{y}_s) \Rightarrow y_i = \bar{y} + S + \varepsilon.$$

Общая колеблемость уровней динамического ряда раскладывается на две составляющие: $S = (\bar{y}_s - \bar{y})$ – влияние сезонности; $\varepsilon = (y_i - \bar{y}_s)$ – влияние случайности.

В ***мультипликативной модели*** сезонность выражена как процент от среднего уровня (например, 120 %), который должен быть учтён при прогнозировании путём умножения на него среднего значения ряда. При ***мультипликативной модели*** уровень динамического ряда можно представить как произведение его составляющих:

$$y_i = \bar{y} \cdot \frac{\bar{y}_S}{\bar{y}} \cdot \frac{y_i}{y_S},$$

где $K_s = \frac{\bar{y}_S}{\bar{y}}$ – коэффициент сезонности, а $K_e = \frac{y_i}{y_S}$ отражает влияние случайного фактора. Чем больше коэффициент сезонности, тем больше амплитуда колебаний уровней ряда относительно его среднего уровня, тем существеннее влияние сезонности. Чем меньше влияние случайной составляющей, тем в большей мере рассматриваемая модель адекватно описывает исходный временной ряд.

Прогнозирование динамического ряда с сезонными колебаниями при отсутствии в нём тенденции сводится к прогнозированию среднего уровня с последующей корректировкой его на сезонную компоненту.

Контрольные вопросы и задания

1. Дайте определение ряда динамики. Какие бывают ряды динамики?
2. Приведите примеры методов сглаживания уровней ряда динамики.
3. Что такое тренд, циклические колебания, случайные колебания уровней?
4. Какие методы аналитического выравнивания вам известны?
5. В чём отличие метода скользящей средней от метода округления интервалов?
6. Приведите методы линейной регрессии.
7. Какие бывают регрессии помимо линейной?
8. Как оценить адекватность найденного тренда?
9. Как выявить сезонные колебания и как от них избавиться?

Список литературы

1. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.

2. Статистические характеристики процессов. URL: https://www.moodle.kstu.ru/pluginfile.php/383238/mod_resource/content/1/ЦМХТП_Т2_Статистические%20характеристики%20процессов_ЛР4.pdf (дата обращения: 17.02.2023). Текст: электронный.

3. Численность населения. URL: https://www.rosstat.gov.ru/storage/mediabank/demo11.xls (дата обращения: 25.02.2023). Текст: электронный.

4. Meteo. URL: http://www.aisori-m.meteo.ru/waisori/index1.xhtml (дата обращения: 04.02.2023). Текст: электронный.

10. Двумерный статистический анализ и его применение

Важной задачей обработки геофизических данных является изучение зависимостей между изучаемыми признаками (например, между различными физическими свойствами горных пород), между показаниями различных методов (например, между глубиной залегания сейсмического горизонта и аномальными значениями силы тяжести в гравиразведке) и т. д. Другой распространённой задачей обработки геофизических данных является их аппроксимация некоторой зависимостью, в частности полиномом заданной степени. Изучение и построение указанных зависимостей, некоторые из которых мы уже разобрали для временных рядов, предусматривают оценку тесноты связи и формы проявления этой связи. Тогда задачи обработки решаются на основе корреляционно-регрессионного анализа. Такой анализ проводится в рамках двумерной статистической модели.

Пусть имеется система из n однородных геологических или геофизических объектов, у каждого из которых измерены характеристики двух свойств $x_1, x_2, ..., x_n$ и $y_1, y_2,..., y_n$.

В основе двухмерной модели лежат три гипотезы:

1) значения $x_1, x_2, ..., x_n$ и $y_1, y_2,..., y_n$ носят случайный характер;

2) значения каждого из свойств не зависят между собой (но могут существовать зависимости между свойствами x и y);

3) совокупность измеренных свойств является однородной.

Система значений $x_1, x_2, ..., x_n$ и $y_1, y_2,..., y_n$ называется системой двух случайных величин, двумерной случайной величиной или случайным вектором.

Двумерные случайные величины принято изображать на графике, где по оси абсцисс откладывают характеристику одного свойства ($x_1, x_2, ..., x_n$), а по оси ординат – другого ($y_1, y_2,..., y_n$). Каждое измерение на таком графике изображают точкой, а объектов изучения – облаком точек. Расположение точек на

графике позволяет сделать предварительные выводы о характере зависимости между изучаемыми свойствами.

Если точки расположены вдоль линии, то между характеристиками свойств имеется функциональная зависимость (рис. 49*а, б, г, д, ж, з*). Она может быть линейной (рис. 49*а, б, г, д*) и нелинейной (рис. 49*ж, з*). Если же точки расположены беспорядочно, то зависимости между характеристиками свойств нет (рис. 49*е*).

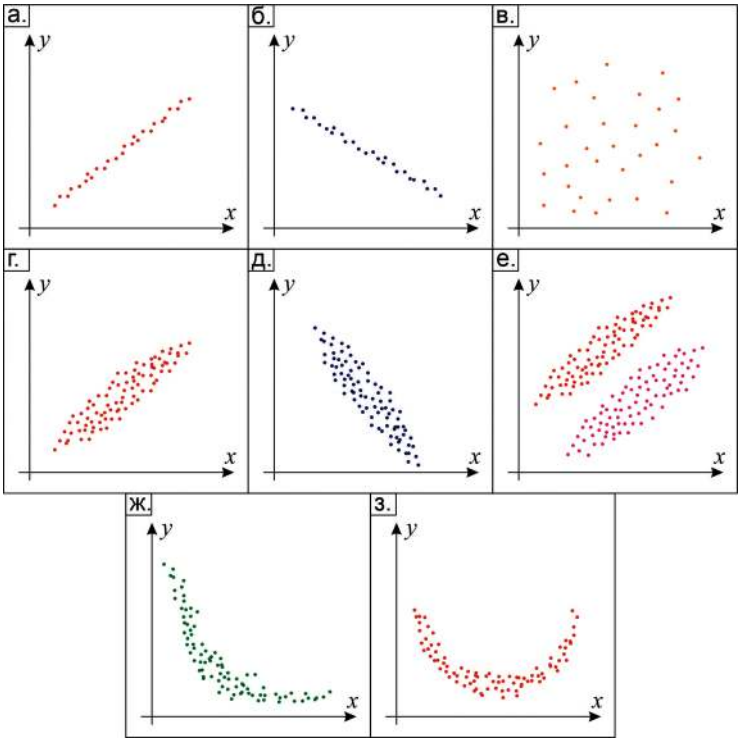


Рис. 49. Изображение двумерных случайных величин для различных случаев зависимостей и с разной корреляцией

Чаще всего точки располагаются в виде облака, группирующегося вдоль какой-то линии (рис. 49*г, д*). В этом случае наблюдается нестрогая статистическая зависимость между свойствами. Она также может быть линейной и нелинейной.

Функциональные и статистические зависимости могут быть положительными, когда с возрастанием характеристики одного свойства увеличивается и другая (рис. 49*а, г*), но могут быть и отрицательными, когда характеристика одного свойства растёт, а другого – убывает (рис. 49*б, д*).

Иногда точки могут образовать два и более изолированных или частично перекрывающихся облака (рис. 49*е*), что свидетельствует о двух и более однородных совокупностях, которые следует изучать раздельно.

Зависимость, при которой изменение одной величины вызывает изменение распределения другой, называется ***статистической (стохастической).*** При статистической зависимости различают ***корреляцию***, когда устанавливают существование взаимосвязи между двумя (или более) случайными величинами и оценивают силу (тесноту) этой связи, и ***регрессию,*** когда выясняют характер (форму/тренд) зависимости между величинами $x_1, x_2, ..., x_n$ и $y_1, y_2,..., y_n$ и возможность оценки $x_1, x_2, ..., x_n$ по $y_1, y_2,..., y_n$ или наоборот.

Теснота взаимосвязи может быть оценена качественно по ширине облака точек – чем меньше его ширина, тем больше теснота и сильнее зависимость. Например, на рис. 49*а, б* зависимость сильнее, чем в случаях, приведённых на рис. 49*г, д*.

Оценка тесноты связи производится путём расчёта ***коэффициента корреляции,*** определяющего и степень взаимосвязи между случайными величинами. Существует несколько способов вычисления коэффициентов корреляции, каждый из которых характеризует ту или иную характеристику искомой зависимости.

Коэффициент корреляции знаков (Фехнера) измеряет силу и направление связи между двумя переменными. Он является простейшим показателем тесноты связи, который основан на сравнении поведения отклонений индивидуальных значений каждого признака от своей средней величины – $(x_i - \bar{x})$ и $(y_i - \bar{y})$.

Во внимание принимаются не величины отклонений $(x_i - \bar{x})$ и $(y_i - \bar{y})$, а их знаки («+» или «–»). Определив эти знаки в каждом ряду, рассматривают все пары знаков и подсчитыва-

ют число их совпадений (C) и несовпадений (H). Тогда коэффициент Фехнера рассчитывается как отношение разности чисел пар совпадений и несовпадений знаков к их сумме, т. е. к общему числу наблюдений:

$$-1 < K_{\phi} = \frac{\sum C - \sum H}{\sum C + \sum H} < 1.$$

Чаще остальных используется ***линейный коэффициент корреляции (Пирсона)***, который измеряет силу и направление связи между двумя переменными. Для его расчёта используют ***корреляционный момент*** (или ***ковариацию***)

$$K_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right).$$

Ковариация отражает взаимосвязь между случайными величинами x и y. Для удобства её преобразуют в безразмерную величину – ***линейный коэффициент корреляции (Пирсона)***, который измеряет силу и направление связи между двумя переменными.

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y},$$

где $\overline{x}$, $\overline{y}$ – средние значения;

σ_x, σ_y – среднеквадратические отклонения величин x и y соответственно.

Отметим основные свойства коэффициента корреляции.

1. Коэффициент корреляции изменяется на отрезке от –1 до +1:

$$-1 \le r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} \le 1.$$

2. Если между переменными существует сильная положительная связь, то $r_{xy} \approx 1$ (рис. 50*а*).

3. Если между переменными существует сильная отрицательная связь, то $r_{xy} \approx -1$ (рис. 50*б*).

4. Если между переменными нет линейной связи или она очень слабая, то $r_{xy} \approx 0$ (рис. 50*в*).

5. С возрастанием абсолютной величины r_{xy} линейная корреляционная зависимость становится всё более тесной и при $|r_{xy}| = 1$ переходит в функциональную.

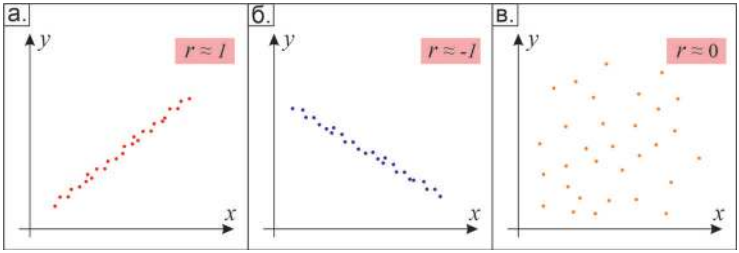


Рис. 50. Изображение двумерных случайных величин с различной корреляционной связью

В реальных условиях коэффициент корреляции не бывает равным единице (или минус единице) и характеризует степень статистической связи между свойствами x и y. Чем ближе по абсолютной величине r к единице, тем сильнее связь между свойствами. Она может быть положительной ($r > 1$, рис. 50*а*) и отрицательной ($r < 1$, рис. 50*б*). Ещё раз отметим, что ***коэффициент корреляции является мерой линейной зависимости между двумя величинами***. Для оценки нелинейных зависимостей он непригоден.

Существует ***эмпирическое правило (шкала Чэддока)*** для оценки тесноты связи:

$\|r\|$	***Теснота связи***
менее 0,1	отсутствует **линейная связь**
0,1–0,3	слабая
0,3–0,5	умеренная
0,5–0,7	заметная
более 0,7	сильная (тесная)

После оценки тесноты связи логично найти уравнение зависимости. Это можно сделать различными способами, но один из наиболее эффективных и часто используемых – это уже известный нам регрессионный анализ, на основании которого наблюдаемое облако точек аппроксимируется уравнением регрессии. Под регрессией понимают сглаживание экспериментальной зависимости по методу наименьших квадратов. Согласно этому методу, сумма квадратов отклонений экспери-

ментальных данных от сглаживающей функции обращается в минимум, т. е.

$$\Phi = \sum_{i=1}^{n} \left[f\left(x_i, a_0, a_1, a_2, \ldots\right) - y_i \right]^2 \to \min.$$

Для оценки формы зависимости необходимо задать конкретный вид функции *f(x)*, функционал продифференцировать по a_0, a_1, a_2…, а производную приравнять к нулю:

$$\frac{\partial \Phi}{\partial a_k} = 0;\ k = 0,1,\ldots,m.$$

Решив полученную систему нормальных уравнений, можно найти коэффициенты a_0, a_1, a_2… и уравнение регрессии.

Разберём несколько примеров, чтобы установить связи между двумя величинами/переменными и рассчитаем различными способами степень этой зависимости – коэффициент корреляции.

В файле Ex9_task.xlsx (https://disk.yandex.ru/d/XQAnOyxohL1bug) и на рис. 51 представлены значения содержания общего и магнетитового железа в руде и соответствующее изображение. Видно, что общее содержание железа находится в линейной положительной зависимости с содержанием магнетита – точки расположены вдоль линии и одна величина возрастает с ростом другой. Для оценки тесноты линейной зависимости рассчитаем коэффициент корреляции.

Используем формулу для коэффициента корреляции $r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}$, $K_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Найдём отклонения значений рассматриваемых параметров от своей средней величины $(x_i - \bar{x})$ и $(y_i - \bar{y})$, перемножим соответствующие элементы и найдём среднюю произведений (рис. 52). Таким образом, мы найдём ковариацию 179,6. Теперь мы можем это значение разделить на произведение среднеквадратических отклонений каждого параметра σ_x, σ_y. В результате получим коэффициент корреляции 0,9817. Используя ***эмпирическое правило (шкала Чэддока)*** для оценки тесноты связи, можно с уверенностью сказать, что зависимость является сильной.

№ пробы	Содержание железа, % (общего)	Содержание железа, % (магнетит)
1	52	45,7
2	49,4	45,4
3	34,5	28,4
4	41,5	36,6
5	36,5	22,1
6	22,7	10,9
7	42,3	27,5
8	20	10,3
9	23,9	17,3
10	23,8	16
11	33,2	23,8
12	61,8	55,8
13	63,7	57,3
14	22,1	15,2

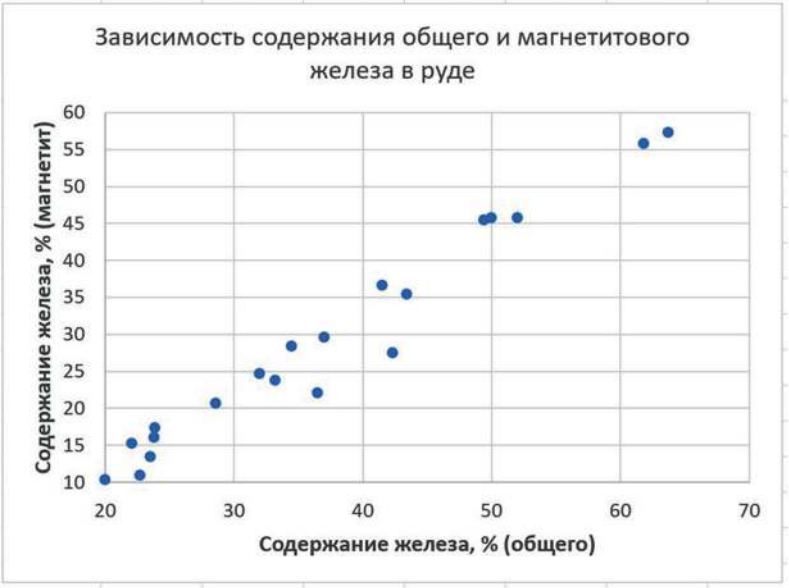
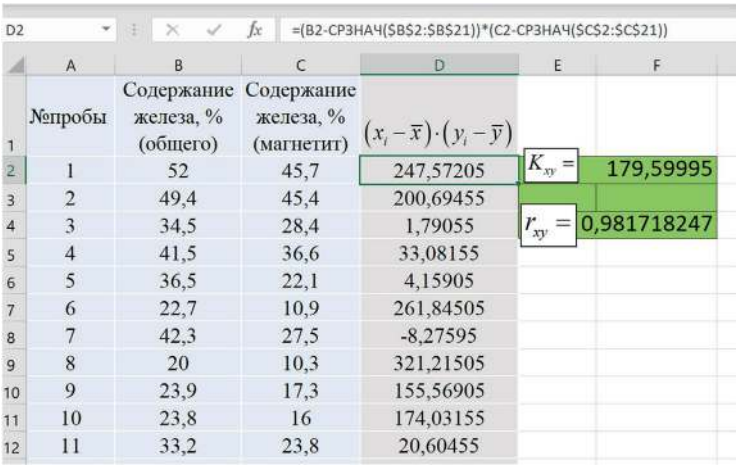


Рис. 51. Таблица значений и изображение зависимости содержания общего и магнетитового железа в руде



D2 =(B2-СРЗНАЧ($B$2:$B$21))*(C2-СРЗНАЧ($C$2:$C$21))

	A	B	C	D	E	F
1	№пробы	Содержание железа, % (общего)	Содержание железа, % (магнетит)	$(x_i-\bar{x})\cdot(y_i-\bar{y})$		
2	1	52	45,7	247,57205	$K_{xy}=$	179,59995
3	2	49,4	45,4	200,69455		
4	3	34,5	28,4	1,79055	$r_{xy}=$	0,981718247
5	4	41,5	36,6	33,08155		
6	5	36,5	22,1	4,15905		
7	6	22,7	10,9	261,84505		
8	7	42,3	27,5	-8,27595		
9	8	20	10,3	321,21505		
10	9	23,9	17,3	155,56905		
11	10	23,8	16	174,03155		
12	11	33,2	23,8	20,60455		

Рис. 52. Фрагмент расчёта корреляции между содержанием общего и магнетитового железа в руде

Другой способ нахождения корреляционного момента (или ковариацию) и корреляции – это использование встроенных функций MS Excel. Для ковариации нужно использовать функцию КОВАРИАЦИЯ.Г(B2:B21;C2:C21), для корреляции – КОРРЕЛ(B2:B21;C2:C21). Можно проверить себя и убедиться, что вычисленные двумя способами значения совпадают и равны 179,6 и 0,9817 соответственно.

После оценки тесноты связи логично найти уравнение зависимости, например, используя линию тренда в MS Excel. Для этого нужно выделить область диаграммы с построенными на ней зависимостью *x* от *y* (рис. 51), кликнув на неё левой кнопкой мыши. После выделения рядом в правом верхнем углу нужно кликом левой кнопки мыши выбрать значок  , который позволяет добавлять/изменять/удалять элементы на диаграмму. В появившемся списке нужно поставить галочку напротив пункта «Линия тренда», после чего появится линия тренда – по умолчанию линейная регрессия (рис. 53, пунктирная линия).

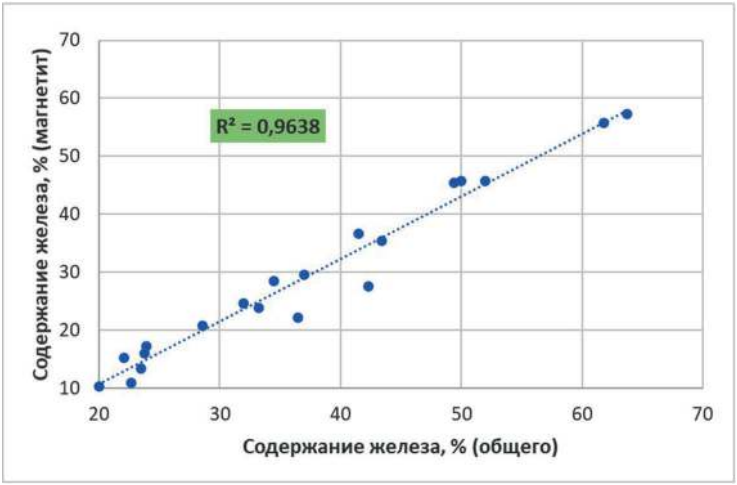
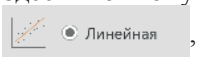
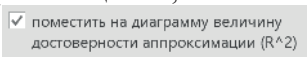


Рис. 53. Изображение зависимости содержания общего и магнетитового железа в руде, линейная регрессия и квадрат корреляции (достоверность аппроксимации)

Для изменения параметров линии тренда необходимо выделить её, кликнуть по ней правой кнопкой мыши и в появившемся окне выбрать пункт «Формат линии тренда». Справа в окне Excel появится область с параметрами линии тренда. Здесь можно убедиться, что выбрана линейная регрессия 

, а также на диаграмму можно нанести величину достоверности аппроксимации R^2, если активизировать соответствующий пункт 

в настройках. В нашем случае достоверность аппроксимации – это квадрат коэффициента корреляции, в чём можно убедиться: $0{,}9817=\sqrt{0{,}963}$. Таким образом, найденные коэффициенты корреляции тремя способами совпадают Ex9_task.xlsx (https://disk.yandex.ru/d/XQAnOyxohL1bug).

Приведем ещё один пример, позволяющий продемонстрировать взаимосвязь двух параметров. Здесь будем использовать метеоданные, но уже с другого открытого источника – с сайта https://rp5.ru/. Набрав соответствующий адрес, попадём на главную страницу, где в поисковой строке напечатаем

интересующий нас населённый пункт – Чита Забайкальского края. Если в населённом пункте несколько метеостанций, то нужно выбрать из списка, в нашем случае – «метеостанция» или «аэропорт». Появятся прогноз погоды для данного населённого пункта и ссылки на архив погоды (рис. 54, красные овалы), которые нам понадобятся для анализа. Перейдём в архив погоды на метеостанции.

В появившемся окне «Архив погоды в Чите» на вкладке «Смотреть архив погоды» (рис. 55) можно выбрать период выборки и увидеть таблицу с имеющимися в архиве метеорологическими параметрами. Пояснение к шифрам в заголовках таблицы появится при наведении на них курсора мыши.

Для работы с данными нам понадобится таблица в пригодном для анализа формате. Для этого перейдём на вкладку «Скачать архив погоды», где выберем настройки для экспорта данных из архива (рис. 56): диапазон дат с 1 января по 31 декабря 2021 г., будем скачивать все дни в формате XLS (Excel). Для скачивания нужно сформировать архив, нажав кнопку «Выбрать в файл GZ(архив)», после чего активизируется кнопка «Скачать» (рис. 56), кликнув которую скачается архив погоды.

Открыв загруженный архив в MS Excel (лист «Архив Погоды rp5 Чита»), можно увидеть, что структура таблицы идентична представленной на вкладке «Смотреть архив погоды» на странице «Архив погоды в Чите» сайта https://rp5.ru/ (рис. 57). Сверху таблицы указана информация о метеостанции и источнике данных, а замеры приведены с периодичностью 3 ч в файле Ex10_task.xlsx (https://disk.yandex.ru/i/KVn-NyYB1tM04oA).

Будем рассматривать только температуру воздуха. Для этого создадим на отдельном листе колонку с датами, перенеся с начальных данных функцией ЦЕЛОЕ(‘Архив Погоды rp5 Чита’!A…), а столбик с температурой перенесём копированием (рис. 58, Ex10_task.xlsx, лист "Correlation").

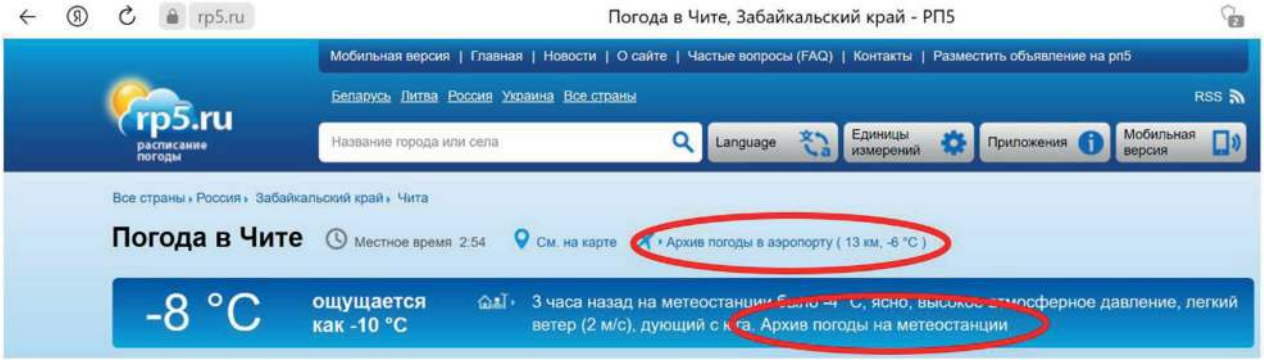


Рис. 54. Страница погоды в Чите на сайте https://rp5.ru

Рис. 55. Вкладка «Смотреть архив погоды» на странице «Архив погоды в Чите» на сайте https://rp5.ru/

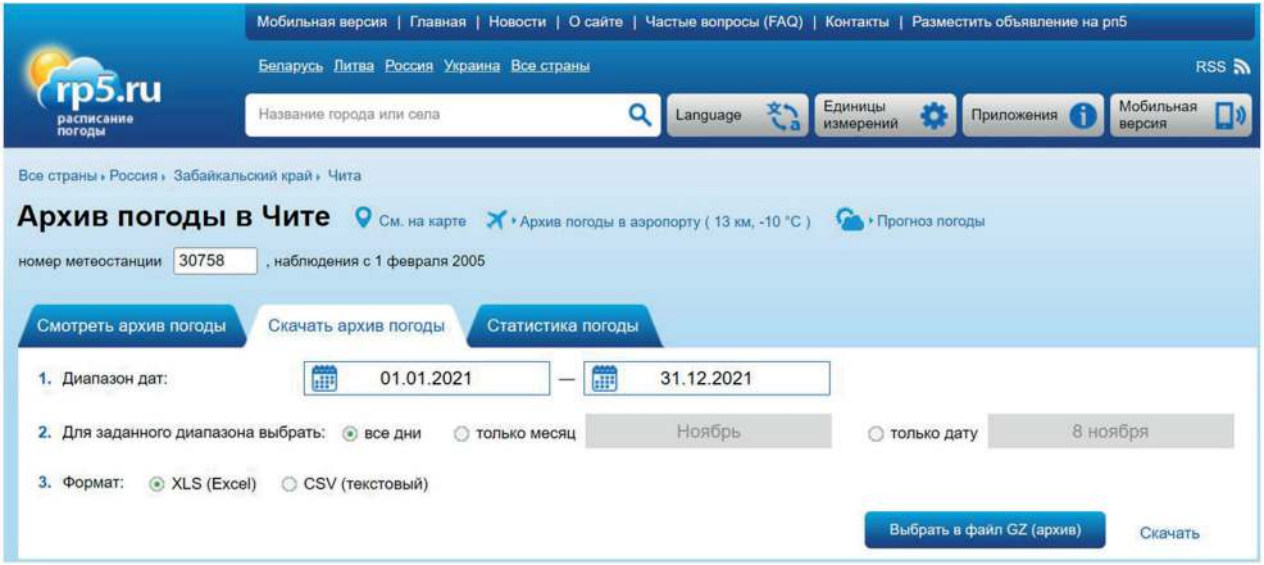


Рис. 56. Вкладка «Скачать архив погоды» на странице «Архив погоды в Чите» на сайте https://rp5.ru/

	A	B	C	D	E	F	G	H	I	J	K	L
1	# Метеостанция Чита, Россия, WMO_ID=30758, выборка с 01.01.2021 по 31.12.2021, все дни											
2	# Кодировка: UTF-8											
3	# Информация предоставлена сайтом "Расписание Погоды", rp5.ru											
4	# Пожалуйста, при использовании данных, любезно указывайте названный сайт.											
5	# Обозначения метеопараметров см. по адресу http://rp5.ru/archive.php?wmo_id=30758&lang=ru											
6	#											
7	Местное время в Чите	T	Po	P	Pa	U	DD	Ff	ff10	ff3	N	W
8	31.12.2021 21:00	-19,9	705,4	771,5	0,3	81	Ветер, дук	1			100%.	Снег
9	31.12.2021 18:00	-20,2	705,1	771,4	0,2	79	Ветер, дук	1			100%.	Снег
10	31.12.2021 15:00	-19,5	704,9	770,9	-1,3	68	Ветер, дук	1			100%.	Мгла.
11	31.12.2021 12:00	-25,0	706,2	773,8	-0,7	74	Ветер, дук	1			70 – 80%.	Мгла.
12	31.12.2021 09:00	-28,7	706,9	775,6	-0,4	73	Штиль, бе:	0			60%.	Мгла.
13	31.12.2021 06:00	-29,6	707,3	776,5	-0,6	73	Штиль, бе:	0			40%.	Мгла.
14	31.12.2021 03:00	-29,1	707,9	776,9	-0,6	74	Ветер, дук	1			60%.	Мгла.
15	31.12.2021 00:00	-27,9	708,5	777,2	-0,2	75	Ветер, дук	1			60%.	Мгла.
16	30.12.2021 21:00	-25,8	708,7	776,9	0,0	76	Ветер, дук	2			60%.	Мгла.
17	30.12.2021 18:00	-22,9	708,7	776,1	-0,1	77	Ветер, дук	2			70 – 80%.	Мгла.
18	30.12.2021 15:00	-22,8	708,8	776,2	-1,1	68	Ветер, дук	1			Облаков н	Мгла.
19	30.12.2021 12:00	-29,6	709,9	779,2	0,4	72	Штиль, бе:	0			Облаков н	Мгла.
20	30.12.2021 09:00	-32,8	709,5	779,8	0,2	69	Штиль, бе:	0			Облаков н	Мгла.
21	30.12.2021 06:00	-32,7	709,3	779,5	0,4	71	Штиль, бе:	0			Облаков н	Мгла.
22	30.12.2021 03:00	-29,4	708,9	778,1	0,9	75	Ветер, дук	1			40%.	
23	30.12.2021 00:00	-26,3	708,0	776,2	0,9	79	Ветер, дук	1			60%.	

Рис. 57. Фрагмент архива погоды на метеостанции «Чита» за 2021 г., загруженный с сайта https://rp5.ru/

	A	B	C	D	E	F	G	H	I	J	K	L
1	Чита						Улан-Удэ					
2	Дата	T			среднедневная, T		Дата	T			среднедневная, T	
3	31.12.2021	-19,9		01.01.2021	-28,85		31.12.2021	-15,2		01.01.2021	-22,3857	
4	31.12.2021	-20,2		02.01.2021	-28,95		31.12.2021	-12,0		02.01.2021	-23,6375	
5	31.12.2021	-19,5		03.01.2021	-23		31.12.2021	-9,2		03.01.2021	-18,9625	
6	31.12.2021	-25,0		04.01.2021	-20,2875		31.12.2021	-12,3		04.01.2021	-18,4875	
7	31.12.2021	-28,7		05.01.2021	-23,15		31.12.2021	-17,5		05.01.2021	-25,7625	
8	31.12.2021	-29,6		06.01.2021	-27,4125		31.12.2021	-19,1		06.01.2021	-24,8125	
9	31.12.2021	-29,1		07.01.2021	-27,975		31.12.2021	-20,0		07.01.2021	-20,875	
10	31.12.2021	-27,9		08.01.2021	-17,1125		31.12.2021	-19,0		08.01.2021	-12,675	
11	30.12.2021	-25,8		09.01.2021	-13,7125		30.12.2021	-20,3		09.01.2021	-12,9125	
12	30.12.2021	-22,9		10.01.2021	-15,2625		30.12.2021	-20,7		10.01.2021	-18,8375	
13	30.12.2021	-22,8		11.01.2021	-15,075		30.12.2021	-19,2		11.01.2021	-17,7875	
14	30.12.2021	-29,6		12.01.2021	-19,075		30.12.2021	-17,3		12.01.2021	-17,5875	
15	30.12.2021	-32,8		13.01.2021	-21,225		30.12.2021	-23,0		13.01.2021	-24,775	
16	30.12.2021	-32,7		14.01.2021	-29,9625		30.12.2021	-25,9		14.01.2021	-25,7375	
17	30.12.2021	-29,4		15.01.2021	-21,4		30.12.2021	-24,8		15.01.2021	-20,7875	
18	30.12.2021	-26,3		16.01.2021	-17,525		30.12.2021	-23,0		16.01.2021	-12,9875	

Рис. 58. Фрагмент температурных данных на метеостанциях «Чита» и «Улан-Удэ», загруженных с сайта https://rp5.ru/ и приведённых к виду для корреляционного анализа

В столбике D сформируем даты в порядке возрастания для расчёта среднедневной температуры. В колонке E рассчитаем среднедневную температуру, используя функцию СРЗНАЧЕСЛИ($A$3:$A$2921;D3;$B$3:$B$2921), как мы уже разбирали в примере Ex5_task.xlsx с динамическими рядами.

Точно такие же операции, начиная со скачивания данных и заканчивая построением среднедневных температур, проделаем для метеостанции «Улан-Удэ». В результате получим данные на каждый день 2021 г. для двух городов (рис. 59, Ex10_task.xlsx, Лист "Correlation"). Теперь мы можем рассмотреть корреляционную зависимость температуры в Чите (ось X) и Улан-Удэ (ось X).

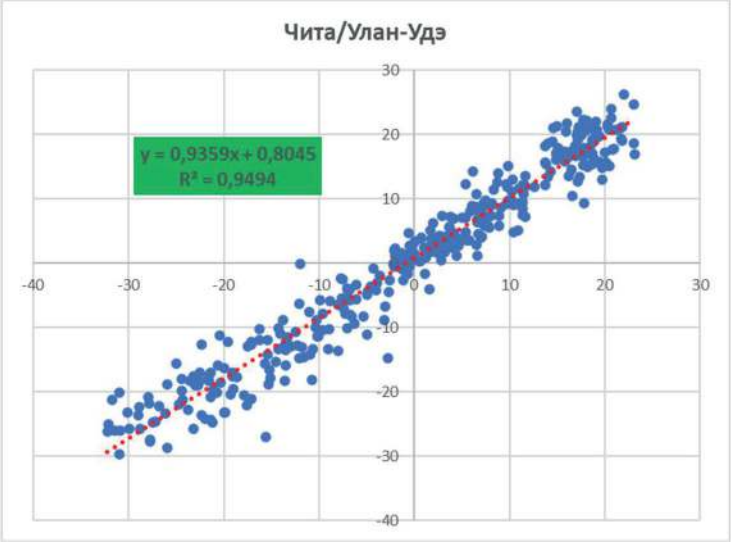


Рис. 59. Корреляции температуры воздуха на метеостанциях «Чита» и «Улан-Удэ» в 2021 г.

Изображение зависимости двух рассматриваемых параметров, уравнение линейной регрессии и достоверность аппроксимации приведены на рис. 59. Коэффициент корреляции предсказуемо близок к единице (0,9744) из-за близости этих

городов, что говорит о сильной связи (схожести) температуры в Чите и Улан-Удэ.

Другой случай, если мы возьмём удалённые на значительное расстояние друг от друга населённые пункты, например Читу и Новосибирск. Скачаем метеоданные за 2021 г. для Новосибирска и найдём среднедневную температуру (рис. 60, Ex10_task.xlsx, Лист "Correlation"). Приведённые данные, как и в прошлом случае, изобразим на плоскости в виде облака точек: ось X – Чита, ось Y – Новосибирск (рис. 61, Ex10_task.xlsx, Лист "Correlation"), здесь же выведем уравнение линейной регрессии и достоверность аппроксимации.

	A	B	C	D	E	M	N	O	P	Q	R
1		Чита					Новосибирск				
2	Дата	T			среднедневня, T	Дата	T			среднедневня, T	
3	31.12.2021	-19,9		01.01.2021	-28,85	31.12.2021	-18,3		01.01.2021	-28,825	
4	31.12.2021	-20,2		02.01.2021	-28,95	31.12.2021	-17,9		02.01.2021	-33,275	
5	31.12.2021	-19,5		03.01.2021	-23	31.12.2021	-16,8		03.01.2021	-32,15	

Рис. 60. Фрагмент температурных данных на метеостанциях «Чита» и «Новосибирск», загруженных с сайта https://rp5.ru/ и приведённых к виду для корреляционного анализа

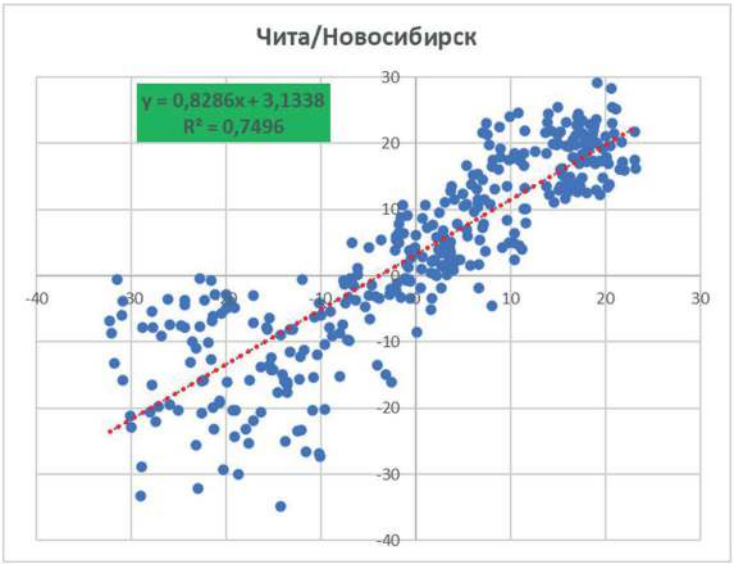


Рис. 61. Корреляции температуры воздуха на метеостанциях «Чита» и «Новосибирск» в 2021 г.

Корреляция (0,866) температуры воздуха в Чите и Новосибирске является не такой тесной, как в прошлом случае, но всё ещё остаётся сильной исходя из шкалы Чэддока. Если рассмотреть отдельно отрицательные и положительные температуры, то можно даже на качественном уровне говорить об отсутствии корреляции в зимние месяцы. Это говорит о значительном различии погоды в Чите и Новосибирске в зимние месяцы.

Мы разобрали простые примеры, позволяющие понять использование двумерного статистического анализа на практике. Логичным продолжением двумерного случая является система множества случайных величин.

Контрольные вопросы и задания

1. Дайте определение двумерной случайной величины.
2. Что такое коэффициент корреляции и чем он отличается от коэффициента корреляции знаков (Фехнера)?
3. Перечислите основные свойства коэффициента корреляции.
4. В чём заключается эмпирическое правило (шкала Чэддока) для оценки тесноты связи двух случайных величин?

Список литературы

1. Дьяконов В. В., Жорж Н. В. Компьютерные методы обработки геологической информации: учеб. пособие. М.: РУДН, 2008. 266 с.
2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.
3. Никитин А. А., Петров А. В. Теоретические основы обработки геофизической информации: учеб. пособие. М.: РГГУ, 2008. 112 с.
4. ООО «Расписание Погоды». URL: https://www.rp5.ru (дата обращения: 03.02.2023). Текст: электронный.
5. Статистические характеристики процессов. URL: https://www.moodle.kstu.ru/pluginfile.php/383238/mod_resource/content/1/ЦМХТП_Т2_Статистические%20характеристики%20процессов_ЛР4.pdf (дата обращения: 17.02.2023). Текст: электронный.
6. Численность населения. URL: https://www.rosstat.gov.ru/storage/mediabank/demo11.xls (дата обращения: 25.02.2023). Текст: электронный.
7. Шеин А. Н., Потапов В. В. Математическое моделирование в разведочной геофизике: учеб.-метод. пособие. Чита: ЗабГУ, 2017. 125 с.
8. Meteo. URL: http://www.aisori-m.meteo.ru/waisori/index1.xhtml (дата обращения: 04.02.2023). Текст: электронный.

11. Система множества случайных величин и её статистические характеристики

Дальнейшим развитием двумерной статистической модели служит многомерная статистическая модель, которая состоит из совокупности множества сопряжённых случайных величин (называемых многомерными случайными векторами) и выражается матрицей свойств размером *k x n*:

$$\begin{vmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{vmatrix},$$

где n – число наблюдений;

k – число свойств.

В основе многомерной статистической модели лежит гипотеза о том, что измеренные значения являются **независимыми случайными** величинами (векторами), т. е. строки матрицы можно располагать в любом порядке. Однако между столбцами матрицы связь может присутствовать. В ряде задач некоторые из измерений могут быть неслучайными величинами, например заранее заданными пространственными или временными координатами.

Для изображения множества случайных величин используется многомерное признаковое пространство, имеющее k осей. Каждое отдельное измерение (строка матрицы) изображается в таком пространстве точкой, а их совокупность, т. е. матрица, – облаком точек.

Многомерная статистическая модель имеет различные статистические характеристики, наиболее употребительными из которых являются средние значения случайных величин $\bar{x}_1$, $\bar{x}_2$, …, $\bar{x}_k$, дисперсии $\sigma^2_{x_1}$, $\sigma^2_{x_2}$, …, $\sigma^2_{x_k}$ и среднеквадратичные отклонения σ_{x_1}, σ_{x_2}, …, σ_{x_k}. Кроме того, часто используются ***матрицы ковариации*** и ***коэффициентов корреляции*** случайных величин. Матрица ковариации имеет симметричный вид:

$$\begin{vmatrix} \sigma_{x_1}^2 & K_{12} & \dots & K_{1k} \\ K_{21} & \sigma_{x_2}^2 & \dots & K_{2k} \\ \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & \sigma_{x_k}^2 \end{vmatrix},$$

где $K_{ij} = \frac{1}{n}\sum_{t=1}^{n}\left(x_{ti} - \overline{x}_i\right)\left(x_{tj} - \overline{x}_j\right).$

Матрица коэффициентов корреляции между свойствами (их называют парными коэффициентами корреляции) также имеет симметричный вид:

$$\begin{vmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{vmatrix},$$

где $r_{ij} = \frac{K_{ij}}{\sigma_{x_i}\sigma_{x_j}}$.

В матрице коэффициентов корреляции по диагонали находятся единицы, а в остальных полях – собственно коэффициенты корреляции. Данные матрицы коэффициентов корреляции могут быть представлены в виде графа связей (рис. 62). Для построения графа использованы результаты силикатного анализа горных пород. Чем больше коэффициент корреляции между компонентами, тем толще соединяющая их линия.

Во многих случаях возникает необходимость изучить зависимость одной случайной величины от множества других случайных величин. Многофакторная зависимость обычно выражается уравнением множественной линейной регрессии:

$$y = a_1x_1 + a_2x_2 + \dots + a_kx_k + b,$$

где $x_1, x_2, \dots, x_k$ – свойства;

$a_1, a_2, \dots, a_k, b$ – постоянные коэффициенты.

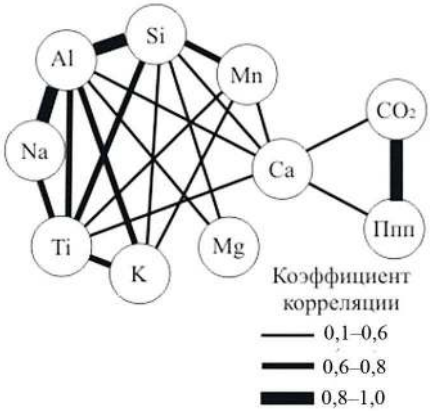


Рис. 62. Данные матрицы коэффициентов корреляции, представленные в виде графа связей

Как и в двумерном случае, выражение множественного уравнения регрессии можно записать через статистические характеристики:

$$y = \overline{y} + A_1 \frac{\sigma_y}{\sigma_{x_1}}\left(x_1 - \overline{x_1}\right) + A_2 \frac{\sigma_y}{\sigma_{x_2}}\left(x_2 - \overline{x_2}\right) + \ldots + A_k \frac{\sigma_y}{\sigma_{x_k}}\left(x_k - \overline{x_k}\right).$$

Значения A_1, A_2, ..., A_k находят путём решения системы линейных уравнений, составленной из коэффициентов корреляции:

$$\begin{pmatrix} 1 & r_{12} & \ldots & r_{1k} \\ r_{21} & 1 & \ldots & r_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & 1 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \\ A_k \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \\ \\ r_{ky} \end{pmatrix}.$$

Сравнение фактических y и расчетных $y_{рас}$ значений даёт отклонения δ. Рассчитав дисперсию отклонений и дисперсию исходных данных, можно найти коэффициент множественной корреляции R, который характеризует степень зависимости свойства y от множества других случайных величин $x_1, x_2, \ldots, x_k$:

$$R = \sqrt{1 - \frac{\sigma_\delta^2}{\sigma_y^2}}.$$

Рассмотрим трёхмерный случай множественного уравнения регрессии:

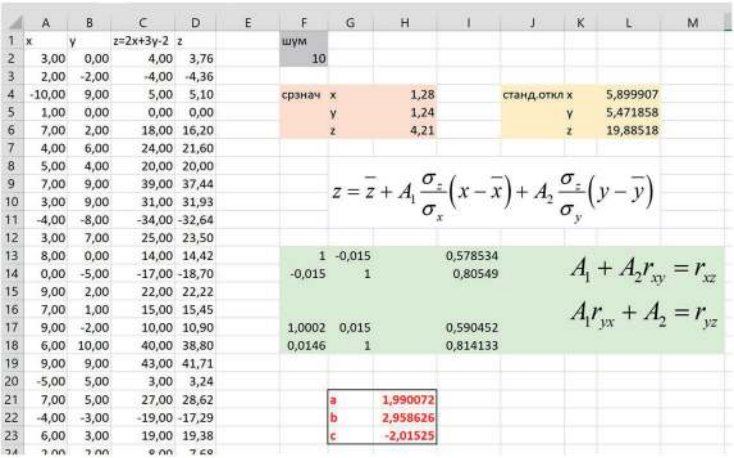
$$z = \bar{z} + A_1 \frac{\sigma_z}{\sigma_x}\left(x - \bar{x}\right) + A_2 \frac{\sigma_z}{\sigma_y}\left(y - \bar{y}\right).$$

Значения A_1, A_2, в этом случае находят путём решения системы линейных уравнений, составленной из коэффициентов корреляции:

$$A_1 + A_2 r_{xy} = r_{xz},$$
$$A_1 r_{yx} + A_2 = r_{yz}.$$

Для анализа смоделируем трёхмерные данные. Для этого зададим значения x и y в столбце A и B таблицы Excel (рис. 63, Ex11_task.xlsx, https://disk.yandex.ru/i/EUwZu0uYo4AOGQ), используя функцию СЛУЧМЕЖДУ(-10;10). В третьем столбце зададим зависимость первого порядка $z = 2x + 3y - 2$. По отработанной схеме прибавим к значениям z случайный шум определённой величины (рис. 63, ячейка F2), используя функцию СЛУЧМЕЖДУ. В результате получим три параметра (рис. 63, столбцы A, B, D) для применения уравнения линейной регрессии.

Для вычисления коэффициентов трёхмерной зависимости первого порядка (уравнение плоскости) нужно найти средние значения, среднеквадратические отклонения и составить уравнение из коэффициентов корреляции (рис. 63). После чего можно вычислить значения коэффициентов *a, b, c*. В приведённом примере (рис. 63, Ex11_task.xlsx) $a = 1{,}99$, $b = 2{,}96$, $c = -2{,}02$. Таким образом, использование уравнения многомерной линейной регрессии с хорошей точностью восстанавливает истинную зависимость ($a = 2$, $b = 3$, $c = -2$).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	x	y	z=2x+3y-2	z		шум							
2	3,00	0,00	4,00	3,76		10							
3	2,00	-2,00	-4,00	-4,36									
4	-10,00	9,00	5,00	5,10		срзнач	x	1,28		станд.откл	x	5,899907	
5	1,00	0,00	0,00	0,00			y	1,24			y	5,471858	
6	7,00	2,00	18,00	16,20			z	4,21			z	19,88518	
7	4,00	6,00	24,00	21,60									
8	5,00	4,00	20,00	20,00									
9	7,00	9,00	39,00	37,44									
10	3,00	9,00	31,00	31,93									
11	-4,00	-8,00	-34,00	-32,64									
12	3,00	7,00	25,00	23,50									
13	8,00	0,00	14,00	14,42			1	-0,015	0,578534				
14	0,00	-5,00	-17,00	-18,70			-0,015	1	0,80549				
15	9,00	2,00	22,00	22,22									
16	7,00	1,00	15,00	15,45									
17	9,00	-2,00	10,00	10,90			1,0002	0,015	0,590452				
18	6,00	10,00	40,00	38,80			0,0146	1	0,814133				
19	9,00	9,00	43,00	41,71									
20	-5,00	5,00	3,00	3,24									
21	7,00	5,00	27,00	28,62			a	1,990072					
22	-4,00	-3,00	-19,00	-17,29			b	2,958626					
23	6,00	3,00	19,00	19,38			c	-2,01525					

$$z = \bar{z} + A_1 \frac{\sigma_z}{\sigma_x}\left(x - \bar{x}\right) + A_2 \frac{\sigma_z}{\sigma_y}\left(y - \bar{y}\right)$$

$$A_1 + A_2 r_{xy} = r_{xz}$$

$$A_1 r_{yx} + A_2 = r_{yz}$$

Рис. 63. Фрагмент файла Ex11_task.xlsx с демонстрацией применения уравнения множественной линейной регрессии

В качестве одного из примеров применения множественной линейной регрессии в трёхмерном случае для геофизических данных можно привести выделение региональной составляющей потенциальных полей.

Контрольные вопросы и задания

1. Как определить многомерную статистическую модель?
2. Что такое матрицы ковариации и коэффициентов корреляции?
3. Как можно изобразить многомерную статистическую модель?
4. Охарактеризуйте множественное уравнение регрессии.

Список литературы

1. Дьяконов В. В., Жорж Н. В. Компьютерные методы обработки геологической информации: учеб. пособие. М.: РУДН, 2008. 266 с.
2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.
3. Никитин А. А., Петров А. В. Теоретические основы обработки геофизической информации: учеб. пособие. М.: РГГУ, 2008. 112 с.
4. Шеин А. Н., Потапов В. В. Математическое моделирование в разведочной геофизике : учеб.-метод. пособие. Чита: ЗабГУ, 2017. 125 с.

12. Основы дисперсионного анализа

Основой дисперсионного анализа являются изучение и сравнение дисперсий наблюденных данных и их линейных комбинаций, при этом измеренные значения геофизических полей полагаются случайными величинами.

Задача дисперсионного анализа – оценка влияния того или иного фактора y или комбинации таких факторов на распределение измеряемых величин $x_1, x_2,...,x_n$ проявляющееся в изменении математического ожидания при неизменной дисперсии.

Одно из основных требований, которому должны удовлетворять результаты геофизических измерений, состоит в том, чтобы они не содержали систематических ошибок. Факторами, вызывающими систематические ошибки, являются погрешности аппаратуры, методики измерений, оператора, а также неучтённые изменения физических условий наблюдения.

Когда решается задача, например, о систематической ошибке гравиметра как фактора, влияющего на показания силы тяжести, измерение на двух приборах (гравиметрах) может оказаться недостаточным.

Для получения обоснованного вывода следует провести измерения при нескольких состояниях фактора, т. е. на нескольких приборах, и убедиться в отсутствии или наличии систематической ошибки в полученных результатах. Совершенно аналогично мы можем рассматривать аномалию того или иного геофизического поля как фактор, для надёжного установления которого необходимо выполнить измерения поля на нескольких профилях, т. е. при различных состояниях фактора. При комплексной интерпретации геофизических полей различными состояниями фактора (комплексного параметра) будут данные измерений различных геофизических полей.

В общем случае дисперсионный анализ применяют для того, чтобы установить, оказывает ли существенное влияние некоторый фактор y, имеющий ряд состояний $y_1, ..., y_p$, на изучаемую величину. Попарное сравнение средних значений при этом исключается, поскольку с возрастанием числа средних

увеличивается наибольшее различие между ними: среднее новой выборки может оказаться больше наибольшего или меньше наименьшего из средних, полученных до нового эксперимента.

Идея дисперсионного анализа состоит в сравнении факторной дисперсии, обусловленной воздействием фактора, и остаточной дисперсии, вызванной случайными причинами. Если различие между этими дисперсиями значимо, фактор y оказывает существенное влияние на систему величин $x_1, x_2, \ldots, x_n$.

В этом случае средние при каждом состоянии фактора (групповые средние) также будут иметь значимое различие.

Однофакторный дисперсионный анализ

Пусть на нормально распределённую случайную величину x (геофизическое поле) воздействует фактор y (систематическая ошибка прибора, аномалия и т. д.), имеющий p состояний (p – число приборов, профилей и т. д.), т. е. имеем

$(x_{11}, x_{12}, \ldots, x_{1n})$ при состоянии у $(\bar{x}_1)$;

……

$(x_{p1}, x_{p2}, \ldots, x_{pn})$ при состоянии у $(\bar{x}_p)$.

Ещё раз подчеркнём, что в качестве состояний фактора $y_1, \ldots, y_p$ можно рассматривать измерения поля разными приборами, на разных профилях или измерения p полей на одном и том же профиле.

Пользуясь этими статистическими данными, необходимо проверить нулевую гипотезу, согласно которой групповые средние распределений равны между собой. Если проверяемая гипотеза верна, при сопоставлении средних значений по каждому состоянию фактора не должно быть значимого расхождения между ними, а если такое расхождение обнаружено, нулевую гипотезу следует отбросить.

Обозначим $\bar{x}_1$ – среднее значение из n измерений, выполненных по первому состоянию фактора (т. е. первым гравиметром или по первому профилю и т. д.), $\bar{x}_2$– среднее из измерений по второму состоянию фактора и т. д., следовательно,

$$\overline{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{i1}; \ldots; \overline{x_p} = \frac{1}{n}\sum_{i=1}^{n} x_{ip},$$

$\sigma_1, \sigma_2, \ldots, \sigma_p$ – дисперсия по состояниям фактора.

Обозначим через $\underline{x}$ общее среднее всех измерений, т. е.

$$\overline{x} = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij} \text{ или } \overline{x} = \frac{1}{p}\sum_{j=1}^{p}\overline{x_j}.$$

Тогда величина

$$S = \sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}\right)^2$$

является общей суммой квадратов отклонений измеряемых значений от общей средней $\overline{x}$, а

$$S_{\Phi} = n\sum_{j=1}^{p}\left(\overline{x_j} - \overline{x}\right)^2$$

называется ***факторной суммой*** квадратов отклонений групповых средних (средних по состояниям фактора) от общей средней. Величина S_{Φ} характеризует разброс между различными состояниями фактора y. Наконец, величина

$$S_{ост} = \sum_{i=1}^{n}\left(x_{i1} - \overline{x_1}\right)^2 + \ldots + \sum_{i=1}^{n}\left(x_{ip} - \overline{x_p}\right)^2$$

называется ***остаточной суммой*** квадратов отклонений наблюденных значений внутри групп (или состояний фактора), характеризующей остаточное рассеяние случайных погрешностей результатов измерений. Очевидно, что

$$S_{\text{ост}} = S - S_{\Phi}.$$

Для сумм S, $S_{\text{ф}}$ и $S_{\text{ост}}$ получаем соответствующие выражения общей, факторной и остаточной дисперсий:

$$D = \frac{S}{pn-1}; \; D_{\Phi} = \frac{S_{\Phi}}{p-1}; \; D_{ост} = \frac{S_{ост}}{p(n-1)}.$$

Чтобы проверить нулевую гипотезу о равенстве средних для различных состояний фактора, достаточно применить ***критерий Фишера***, сравнивая факторную и остаточную дисперсию, т. е.

$$F = \frac{D_{\Phi}}{D_{ост}} = \frac{S_{\Phi}p(n-1)}{(p-1)S_{ост}}.$$

Если $F < F\gamma$, где $F\gamma$ – γ-квантиль распределения Фишера с $(p - 1)$ и $p(n - 1)$ степенями свободы (числителя и знаменателя), гипотеза о равенстве средних выполняется, т. е. $\bar{x}_1 = \ldots = x_p$. При этом отсутствует систематическая ошибка, вызванная влиянием гравиметра как фактора, если $y_1, \ldots, y_p$ различные приборы, или можно говорить об отсутствии аномалии при наблюдениях по p профилям съёмки и т. д. Если $F > F\gamma$, имеет место влияние фактора на измеряемую величину.

Контрольные вопросы и вопросы

1. Охарактеризуйте задачу дисперсионного анализа.
2. Что можно рассматривать в качестве состояния фактора при геофизических измерениях?
3. Какую нулевую гипотезу следует проверить при однофакторном дисперсионном анализе?
4. Что характеризуют и как рассчитываются общая, факторная и остаточная суммы?
5. Как рассчитываются общая, факторная и остаточная дисперсии?
6. Для чего применяют критерий Фишера?

Список литературы

1. Дьяконов В. В., Жорж Н. В. Компьютерные методы обработки геологической информации: учеб. пособие. М.: РУДН, 2008. 266 с.
2. Иваненкова А. П. Статистическая обработка геофизической информации: учеб. пособие. Чита: ЧитГУ, 2003. 150 с.
3. Никитин А. А., Петров А. В. Теоретические основы обработки геофизической информации: учеб. пособие. М.: РГГУ, 2008. 112 с.
4. Шеин А. Н., Потапов В. В. Математическое моделирование в разведочной геофизике: учеб.-метод. пособие. Чита: ЗабГУ, 2017. 125 с.

Практические задания

1. Сгруппировать данные, приведённые в примере Ex1_task.xlsx на листе 1 (https://disk.yandex.ru/i/x9RFf9er29VRkA):

а) разбить ряд на 8 интервалов;

б) определить количество интервалов, используя Формулу Стёрджесса.

2. Построить гистограмму, полигон, огиву, кумулляту для сгруппированных данных из задания 1.

3. Рассчитать статистические характеристики для ряда из задания 1 тремя способами: используя формулы, функции MS Excel и описательную статистику модуля «Анализ данных».

4. Построить нормальное распределение для разных средних и дисперсий, наглядно показав, что среднее является характеристикой положения распределения, а дисперсия характеризует степень отклонения случайных величин данной совокупности от среднего.

5. Построить логнормальное распределение для разных средних и дисперсий, наглядно показав, что среднее является характеристикой положения распределения, а дисперсия характеризует степень отклонения случайных величин данной совокупности от среднего.

6. Проверить утверждение, согласно которому логнормальное и нормальное распределения совпадают при малых дисперсиях.

7. Дана выборка 9, 5, 7, 7, 4, 10, дисперсия $\sigma^2 = 1$. Постройте 99 % доверительный интервал $(\bar{x} - \Delta; \bar{x} + \Delta)$, $\Delta = \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$. Для поиска квантилей $z_{\frac{\alpha}{2}}$ использовать таблицу нормального распределения или функцию в MS Excel НОРМ.СТ.ОБР. Ответ: (5,9; 8,05).

Пусть для выборки $n=25$ назначено среднее $\bar{x}=130$. Из предыдущих исследований известно стандартное отклонение $\sigma = 12$. Постройте 98 % доверительный интервал для среднего значения $(\bar{x} - \Delta; \bar{x} + \Delta)$, $\Delta = \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$. Для поиска квантилей $z_{\frac{\alpha}{2}}$ ис-

пользовать таблицу нормального распределения или функцию в MS Excel НОРМ.СТ.ОБР. Ответ: (124,41; 135,59).

8. Пусть объём выборки $n=16$, выборочное среднее $\bar{x}=5$, выборочная дисперсия $s^2=4$. Необходимо построить 99 % доверительный интервал $(\bar{x}-\Delta;\ \bar{x}+\Delta)$, $\Delta=\frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}$. Для поиска квантилей $t_{\frac{\alpha}{2}}$ использовать таблицу t-распределения или функцию в MS Excel СТЬЮДЕНТ.ОБР. Ответ: (3,355; 6,645).

9. Распределение Стьюдента стремится к нормальному распределению при $n \to \infty$, поэтому при больших выборках доверительные интервалы для среднего, посчитанные по любой из приведённых в разделе 8 формул, будут почти совпадать. Смоделировать выборку или использовать один из примеров, проверить данное утверждение.

10. Используя сайты с архивом метеоданных (источники 13, 14 библиографического списка), скачать метеоданные одной из метеостанций РФ за период не менее 70 лет. Применить для них метод укрупнения интервалов, используя следующие интервалы: месяц, квартал, год. Охарактеризовать данные, обработанные с разными интервалами.

11. Используя данные из задания 11, применить метод скользящей средней с разными окнами: месяц, квартал, год. Охарактеризовать данные, обработанные с разными интервалами. Сравнить данные с полученными в задании 11.

12. Для среднегодовых данных из задания 11 найти линию тренда двумя способами: использовать уравнение линейной регрессии и линию тренда MS Excel. Показать, что полученные коэффициенты линейного уравнения совпадают.

13. Для среднегодовых данных из задания 11 рассчитать прогнозные значения температуры по полному набору данных. Второй прогноз сделать, принимая во внимание данные с 1980 г. Сравнить полученные прогнозы. Какой из прогнозов более консервативный и почему?

14. Сформировать модельные данные в виде линейной зависимости $y=ax+b$ с наложенным шумом (использовать пример Ex7_task.xlsx, https://disk.yandex.ru/d/XQAnOyxohL1bug). Показать, что коэффициенты заданного линейного уравнения

восстанавливаются с хорошей точностью при использовании линейной регрессии. Чем больше шум, тем хуже восстановление?

15. Сформировать модельные данные в виде линейной зависимости $y=ax^2+bx+c$ с наложенным шумом (использовать пример Ex8_task.xlsx, https://disk.yandex.ru/i/mSwQ1TvzmuF3Fw). Показать, что коэффициенты заданного квадратичного уравнения восстанавливаются с хорошей точностью при использовании регрессии. Чем меньше шум, тем лучше восстановление?

16. Сформировать модельные данные в виде линейной зависимости $y=ax+b$ с наложенным шумом (использовать пример Ex7_task.xlsx, https://disk.yandex.ru/d/XQAnOyxohL1bug). Найти тренд для смоделированных шумных данных и проверить его адекватность, используя критерий Фишера. Изменяя шум, найти неадекватное по критерию Фишера линейное приближение.

17. Известны содержания общего и магнетитового железа в руде, которые приведены в таблице. Требуется рассчитать Коэффициент корреляции знаков (Фехнера) и линейный коэффициент корреляции (Пирсона) между этими величинами. Используя эмпирическое правило (шкала Чэддока), оценить тесноту связи.

Содержания общего и магнетитового железа в руде

№ пробы	*Содержание железа, % (общего)*	*Содержание железа, % (магнетит)*
1	52	45,7
2	49,4	45,4
3	34,5	28,4
4	41,5	36,6
5	36,5	22,1
6	22,7	10,9
7	42,3	27,5
8	20	10,3
9	23,9	17,3

Окончание таблицы

№ пробы	***Содержание железа, % (общего)***	***Содержание железа, % (магнетит)***
10	23,8	16
11	33,2	23,8
12	61,8	55,8
13	63,7	57,3
14	22,1	15,2
15	50	45,7
16	43,4	35,4
17	37	29,6
18	28,6	20,7
19	23,5	13,4
20	32	24,7

18. Используя сайты с архивом метеоданных (источник 13 библиографического списка), скачать метеоданные с метеостанций, пользуясь примером Ex10_task.xlsx, лист "Correlation" (https://disk.yandex.ru/i/KVnNyYB1tM04oA). Провести корреляционный анализ и сделать выводы:

а) температурных данных на разных метеостанциях;

б) осадков на разных метеостанциях;

в) осадков и температуры.

19. Сформировать модельные данные в виде трёхмерной линейной зависимости $z=ax + yb + c$ с наложенным шумом (использовать пример Ex11_task.xlsx, https://disk.yandex.ru/i/EUwZu0uYo4AOGQ). Показать, что коэффициенты заданного уравнения восстанавливаются с хорошей точностью при использовании множественной линейной регрессии. Чем меньше шум, тем лучше восстановление?

Заключение

Мы познакомились с основными статистическими понятиями и разобрали на практических примерах некоторые методы, алгоритмы и подходы для обработки геоданных, которые наиболее часто могут встретиться интерпретатору или аналитику.

Полученные знания пригодятся современным исследователям ввиду технологического рывка и непрерывного роста объёма геоинформации, что обусловлено, в первую очередь, переходом на цифровую регистрацию физических полей.

Очевидно, что применение современных методов обработки стало невозможным без использования вычислительной техники и современных прикладных программ. Тем не менее, главным действующим лицом при обработке полевого материала ещё долгое время будет оставаться человек, а итоговый результат зависит не от выбранной программы или метода обработки информации, а большей частью от опыта и знаний геофизика-интерпретатора. Соответственно, интерпретатору эти знания необходимы в основном не для написания программного обеспечения, а для понимания сути уже созданных программ, в которых используется статистическая обработка геоданных. В работе мы предложили обучающемуся получить первые знания и опыт в статистической обработке данных.

В учебном пособии разбираются возможности программы MS Excel при статистической обработке информации, которая довольно популярна при анализе данных, а все примеры можно скачать одним архивом по ссылке https://disk.yandex.ru/d/cQFFU4ZpcJOf7A или обратиться к авторам. Такое изложение материала позволяет студентам связать теорию с практическим получением результатов.

Учебное издание предназначено для студентов геолого-геофизических специальностей, но будет полезно и тем, кто занимается статистической обработкой данных.

Библиографический список

1. Букин, В. С. Статистическая обработка геофизической информации: учеб. пособие / В. С. Букин. – Чита: ЗабГУ, 2014. – 166 с.

2. Дьяконов, В. В. Компьютерные методы обработки геологической информации: учеб. пособие / В. В. Дьяконов, Н. В. Жорж. – Москва: РУДН, 2008. – 266 с.

3. Иваненкова, А. П. Статистическая обработка геофизической информации: учеб. пособие / А. П. Иваненкова. – Чита: ЧитГУ, 2003. – 150 с.

4. Калинин, А. Г. Обработка статистических данных / А. Г. Калинин. – Чита: ЗИП СибУПК, 2010. – 106 с.

5. Коган, Р. И. Интервальные оценки в геологических исследованиях / Р. И. Коган. – Москва: Недра, 1986. – 160 с.

6. Курбацкий, А. Н. Лекция 5. Доверительные интервалы / А. Н. Курбацкий. – URL: https://www.mse.msu.ru/wp-content/uploads/2020/03/Лекция-5-доверительные-интервалы.pdf (дата обращения: 12.02.2023). – Текст: электронный.

7. Математика и статистика (ч. 2). – URL: http://www.math-info.hse.ru/2017-18/Математика_и_статистика,_часть_2 (дата обращения: 12.02.2023). – Текст: электронный.

8. Никитин, А. А. Теоретические основы обработки геофизической информации: учеб. пособие / А. А. Никитин, А. В. Петров. – Москва: РГГУ, 2008. – 112 с.

9. ООО «Расписание Погоды». – URL: https://www.rp5.ru (дата обращения: 03.02.2023). – Текст: электронный.

10. Розенцвайг, А. К. Статистика. Сводка и группировка данных статистического наблюдения: учеб.-метод. пособие / А. К. Розенцвайг, А. Г. Исавнин. – Набережные Челны: Изд-во Набережночелнинского института КФУ, 2019. – 29 с.

11. Ряды динамики. – URL: https://www.chaliev.ru/statistics/ryady-dynamiki.php (дата обращения: 12.02.2023). – Текст: электронный.

12. Савинский, И. Д. Таблицы вероятностей подсечения эллиптических объектов прямоугольной сетью наблюдений / И. Д. Савинский. – Москва: Недра, 1964. – 86 с.

13. Статистические характеристики процессов. – URL: https://www.moodle.kstu.ru/pluginfile.php/383238/mod_resource/content/1/ЦМХТП_Т2_Статистические%20характеристики%20процессов_ЛР4.pdf (дата обращения: 17.02.2023). – Текст: электронный.

14. Численность населения. – URL: https://www.rosstat.gov.ru/storage/mediabank/demo11.xls (дата обращения: 25.02.2023). – Текст: электронный.

15. Шауцукова, Л. З. Информатика. 10–11-е классы / Л. З. Шауцукова. – Москва: Просвещение, 2000.

16. Шеин, А. Н. Математическое моделирование в разведочной геофизике: учеб.-метод. пособие / А. Н. Шеин, В. В. Потапов. – Чита: ЗабГУ, 2017. – 125 с.

17. Meteo. – URL: http://www.aisori-m.meteo.ru/waisori/index1.xhtml (дата обращения: 04.02.2023). – Текст: электронный.

Учебное издание

Шеин Александр Николаевич
Юдицких Евгений Юрьевич
Потапов Владимир Владимирович

Статистические методы обработки геофизической информации

Редактор Е. В. Голованова
Вёрстка Г. А. Зенковой

Подписано в печать: 19.09.2023.
Формат 60×84/16. Бумага ксерографическая.
Гарнитура Times New Roman. Способ печати цифровой.
Усл. печ. л. 8,8. Уч.-изд. л. 6,1. Заказ № 23053.
Тираж 100 экз. (1-й з-д 1–35 экз.).

ФГБОУ ВО «Забайкальский государственный университет»
672039, г. Чита, ул. Александро-Заводская, 30