ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
«Томский государственный архитектурно-строительный университет»

Г.Я. Мамонтов, И.А. Иконникова

АНАЛИЗ ДАННЫХ В MS EXCEL И OO CALC

Учебно-методическое пособие

Томск
Издательство ТГАСУ
2010

УДК 311:001.891.573
М 22

Мамонтов, Г.Я. Анализ данных в MS Excel и OO Calc [Текст] : учебно-методическое пособие / Г.Я. Мамонтов, И.А. Иконникова. – Томск : Изд-во Том. гос. архит.-строит. ун-та, 2010. – 60 с. – ISBN 978-5-93057-376-3.

В пособии представлена методика обработки результатов эксперимента в электронных таблицах MS Excel и OO Calc.

Учебно-методическое пособие по дисциплине «Статистическая обработка экспериментальных данных» предназначено для студентов специальностей 270102, 270105, 270201, 270205, 190601 заочного факультета.

Печатается по решению редакционно-издательского совета Томского государственного архитектурно-строительного университета.

Рецензенты: докт. физ.-мат. наук, проф. **Г.В. Кузнецов,** ТПУ;
докт. физ.-мат. наук, проф. **Н.В. Лаходынова,** ТГАСУ.

ISBN 978-5-93057-376-3

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Важной составляющей научных исследований является умение поставить эксперимент и грамотно обработать его результаты, то есть извлечь из проведенной работы ***максимум формализованной, численно выраженной информации.***

Изучение и применение научных методов обработки результатов эксперимента позволяет, во-первых, извлечь из них обоснованные заключения, во-вторых, ***сэкономить средства и время на постановку эксперимента.*** Практика показала, что стоимость теоретической обработки результатов эксперимента составляет малую часть стоимости эксперимента в целом, но значительно повышает ценность полученных выводов.

Каков уровень взаимодействия теоретических и экспериментальных методов в процессе познания окружающего мира?

Изучение любого явления начинается с систематического и целенаправленного его наблюдения. Полученная в результате информация нуждается, как правило, в теоретическом осмыслении и обобщении. Теоретические методы в виде математических моделей выявляют внутренние механизмы изучаемого явления и, следовательно, способы воздействия на него для изменения в желаемом направлении. Однако при построении математических моделей неизбежно введение каких-либо ограничений, допущений, гипотез. Поэтому возникает задача оценки достоверности построенной модели. И вновь на первый план выходит эксперимент как метод проверки ***адекватности*** теоретических моделей.

Таким образом, теоретические и экспериментальные исследования дополняют друг друга и являются неразрывными частями процесса познания окружающего нас мира.

Методы обработки экспериментальных данных составляют основу различных научных дисциплин, каждая из которых решает свойственные ей задачи. Так, технология измерений, их первичный анализ, оценка погрешностей измерения и тому по-

добное – вопросы метрологической науки. Применяемые здесь методы не предполагают прямого воздействия на изучаемый объект или явление. На противоположном полюсе находится теория планирования активного эксперимента, которая изучает процессы и явления при неполном знании их механизма, то есть с применением идеи «чёрного ящика». Здесь исследователь активно и целенаправленно воздействует на изучаемое явление для определения его движущих сил.

Необходимое связующее звено между этими дисциплинами – математическое моделирование экспериментальных данных на основе вероятностно-статистических методов. Перечислим основные задачи из этой области: подбор эмпирических формул и оценка их параметров, математическое моделирование наблюдаемых случайных показателей, исследование корреляционных зависимостей и некоторые другие.

Из указанного перечня в пособии будут представлены две практически важные задачи – математическое моделирование одного изолированного случайного показателя и двух статистически связанных случайных величин. Их рассмотрение состоит из трёх частей. Первая содержит постановку задачи и обзор необходимых для её решения теоретических сведений. Вторая – пример «ручного» решения задачи. И, наконец, в третьей части задача решается средствами наиболее актуальных табличных процессоров MS Excel и OO Calc.

Эта методика наработана многолетней практикой преподавания «компьютерных» дисциплин на кафедре прикладной математики ТГАСУ. Действительно, вхождение в практику статистического анализа данных и все тонкости изучаемого подхода обеспечивают «ручные» расчёты. Повторное решение той же задачи автоматизированным путем позволяет, с одной стороны, убедиться в правильности найденного ранее решения, с другой – сформировать навык применения современных вычислительных пакетов в реальной производственной деятельности.

ВВЕДЕНИЕ

Содержание настоящего пособия определено программой изучаемой дисциплины: «…обобщение данных статистического наблюдения в виде рядов распределения. Проверка распределения изучаемой совокупности на близость к нормальному закону. Корреляционный и регрессионный анализ. Моделирование статистической зависимости случайных величин…».

В ходе освоения дисциплины студентами выполняется контрольная работа, которая оформляется в отдельной тетради. Титульный лист должен содержать следующую информацию: фамилию студента, номер зачётной книжки и номер варианта, который определяется следующим образом.

Пусть M – число, состоящее из трех последних цифр номера зачётной книжки. Тогда ваш вариант будет N : $N = M$, при $M \leq 30$; $N = M - 30$, при $30 < M \leq 60$; $N = M - 60$, при $60 < M \leq 90$; $N = M - 90$, при $90 < M \leq 120$.

Например, номеру 461-036 ($M = 36$) соответствует 6-й вариант; номеру 462-024 ($M = 24$) соответствует 24-й вариант; номеру 462-068 ($M = 68$) соответствует 8-й вариант.

В работе следует привести текст задания. Изложение «ручного» варианта решения должно быть подробным и обеспечивать защиту работы без обращения к другим источникам.

При выполнении задания в MS Excel и OO Calc необходимо указать версию программы и версию операционной системы. К контрольной работе прилагается дискета (CD, DVD) с расчетами, подписанная так же, как тетрадь. Каждое задание контрольной работы оформляется в отдельном файле, названном Вашей фамилией и номером задания. На том же носителе должна размещаться резервная копия файлов.

Приведённый в пособии список литературы дает широкие возможности для выбора подходящего источника при изучении дисциплины по статистической обработке данных эксперимента.

Учебники и методические указания [1–4], приведённые в начале списка литературы, обеспечивают теоретическую поддержку при выполнении первого задания. Прежде всего, это хорошо известный и переизданный десятки раз учебник В.Е. Гмурмана, который поможет «освежить» необходимые сведения из теории вероятностей и математической статистики.

Специальная литература по обработке экспериментальных статистических данных [5–7] ориентирована на широкий круг читателей от рядового студента и вплоть до практикующего инженера-исследователя. Следует обратить внимание на неразрывную связь между научным подходом к постановке эксперимента и применением выверенных методов для обработки его результатов. Важно отметить, что каждый метод исследования имеет чёткие границы применения. К разряду базовых учебников относится издание [7]. Оно отличается простым и доступным изложением материала, необходимого для выполнения второй контрольной работы.

Наконец, источники [8–11] информационно обеспечивают вхождение в современные вычислительные среды, пригодные для построения и анализа математических моделей в инженерной практике. Конечно, на первых порах можно обойтись и хорошо знакомыми MS Excel или его свободным аналогом OO Calculation. Оба продукта располагают необходимым набором средств автоматизированной обработки таблично организованных экспериментальных данных. Однако после наработки базовых навыков и при решении серьёзных практических задач удобней использовать более специализированные вычислительные среды, такие как Statistica, MathCAD и некоторые другие.

1. УКАЗАНИЯ К ВЫПОЛНЕНИЮ ЗАДАНИЯ № 1

Необходимость в изучении вероятностно-статистических методов диктуется ***случайной*** природой экспериментальных данных. Действительно, почти все инженерные показатели обнаруживают характерный для случайных величин разброс значений даже в неизменных условиях испытания.

От привычной для нас детерминированной величины случайная отличается тем, что каждому её значению соответствует определённая вероятность реализации. Таким образом, описать случайную величину – это значит указать все её возможные значения и для каждого из них определить соответствующую вероятность. ***Такое описание называется законом распределения случайной величины.***

Первое задание состоит в подборе закона распределения для случайной величины, значения которой наблюдаются в серии независимых экспериментов.

Такая задача, во-первых, важна сама по себе, так как закон распределения даёт полную информацию об изучаемой случайной величине в виде множества её возможных значений и отвечающих им вероятностей.

Во-вторых, выявление вида распределения случайной величины – составная часть при решении ряда практически важных задач. Например, применение эффективных методов для описания совместного изменения двух статистически связанных переменных предполагает проверку каждой из этих переменных на соответствие нормальному закону распределения.

Настоящий раздел помогает освоить методику выявления закона распределения случайной величины. Такое знание весьма желательно, так как позволяет ***оценивать вероятности возникновения той или иной ситуации.*** Предполагается использование выборочного метода, согласно которому ***генеральная совокупность*** моделируется набором эксперимен-

тальных данных – **выборкой**. Действительно, оперативное обследование генеральной совокупности, численность которой бывает огромной, требует непомерных затрат (людских, материальных, финансовых). Поэтому для изучения свойств генеральной совокупности обследуют её часть – выборку, извлечённую ***случайным образом*** из генеральной совокупности. Случайный отбор предполагает равные возможности для элементов генеральной совокупности быть включёнными в выборку. Тогда (в силу закона больших чисел) выборка адекватно отображает структуру и свойства генеральной совокупности.

В практическом варианте под выборкой $x_1, x_2,..., x_n$ понимают фактически полученные в данном конкретном эксперименте значения исследуемой случайной величины. Число элементов n в выборке называется ее ***объемом.***

Заключение о свойствах генеральной совокупности, полученное на основании исследования выборки, называется ***статистическим заключением.***

Таким образом, выборочный метод предполагает статистическую обработку полученных в эксперименте выборочных данных с целью теоретического описания свойств объекта наблюдения.

Основные этапы выполнения работы № 1

1. Группировка выборочных данных.

Производить анализ достаточно большой выборки неудобно, поэтому на начальном этапе её преобразуют к компактному виду. Наиболее эффективным способом обобщения и сжатия выборочных данных является построение ряда распределения, или вариационного ряда.

Эта процедура предполагает упорядочение и объединение в группы близких по величине выборочных данных. Каждой

группе соответствует свой диапазон значений (***интервал группировки)*** и своя численность ***(частота).***

Графическое представление вариационного ряда в виде гистограммы позволяет провести быстрый визуальный анализ важных характеристик распределения: наибольшего и наименьшего значений, зон концентрации данных и т. д.

Гистограмма – графическое изображение интервалов группировки и соответствующих им частот – представляет эмпирическую функцию плотности распределения, по её виду выдвигают гипотезу о виде распределения исследуемой случайной величины.

2. Числовые характеристики случайной величины.

В первую очередь это выборочное среднее и выборочная дисперсия, которые необходимы для определения параметров закона распределения для большинства практически важных случайных величин.

Отметим, что числовые характеристики случайной величины могут быть вычислены как на основе исходной, то есть полученной в эксперименте выборки, так и на основе той же выборки, но после процедуры группировки. Сравнение этих двух величин между собой даёт представление о погрешности вследствие группировки выборочных данных – информации, весьма полезной для исследователя.

3. Построение теоретических аналогов для эмпирического закона распределения.

Гипотеза о виде распределения изучаемой случайной величины принимается или отвергается по результатам сравнения наблюдаемого эмпирического закона распределения и его теоретического аналога. При этом в качестве параметров теоретического закона распределения следует использовать выборочные числовые характеристики, найденные выше, на втором этапе. Тем самым достигается «привязка» теоретического закона распределения к области

реальных, то есть зафиксированных в эксперименте значений случайной величины.

4. Формулировка заключения по результатам проверки гипотезы о виде распределения изучаемой случайной величины.

Согласованность выборочного и теоретического распределений можно проверить, например, с помощью χ^2-критерия Пирсона. Фактически речь идёт о сопоставлении выборочной и теоретической функций плотности распределения в области экспериментально наблюдаемых значений случайной величины.

1.1. Варианты первого задания

Варианты первого задания, приведенные в табл. 1.1, можно также найти в источнике [4].

Таблица 1.1

1	2	3	4	5
7,705	10,515	6,693	11,604	–2,597
8,947	7,971	8,038	9,018	–1,955
9,14	10,315	7,424	11,013	–3,031
9,923	8,484	9,322	10,573	–0,228
5,703	7,245	11,056	10,283	–2,185
9,166	5,751	6,311	8,944	0,066
9,551	7,026	9,082	7,605	–1,834
4,600	8,334	9,629	8,762	–3,189
6,946	6,753	10,799	8,057	–2,909
8,882	9,924	8,858	7,980	–3,225
7,921	6,137	9,168	9,444	–2,979
5,533	5,822	9,861	12,557	–0,655

Продолжение табл. 1.1

1	2	3	4	5
9,097	7,754	10,855	7,322	0,010
5,770	9,422	10,565	12,886	0,396
5,704	6,117	6,795	12,006	–2,678
4,532	5,217	10,234	12,915	–1,054
4,998	9,002	6,875	8,754	–2,542
6,064	8,193	10,282	12,501	–1,096
4,255	10,468	7,846	8,447	–0,054
7,482	10,602	8,072	8,876	–3,860
6,389	10,995	10,394	11,322	–0,737
5,425	10,624	6,903	7,274	–3,810
6,044	10,919	7,313	12,281	–0,319
6,415	10,804	11,983	8,756	–2,717
4,065	8,406	8,433	7,491	–3,666
6,876	7,822	9,958	10,334	–3,219
4,428	5,757	6,979	12,200	–1,142
7,757	8,176	7,245	9,008	–1,805
4,821	10,325	10,92	11,517	–2,456
9,619	7,061	11,008	9,370	0,055
9,795	5,894	11,508	7,860	–1,005
9,757	6,373	10,947	8,737	–3,668
7,633	5,493	8,801	11,454	–2,622
6,774	6,955	10,404	11,106	–2,952
7,159	6,841	11,931	10,308	–2,581
4,135	5,448	9,941	11,259	–2,781
8,614	6,520	6,354	8,536	0,745
4,657	5,940	11,882	8,999	–1,032
6,936	6,675	6,152	7,883	–2,881
6,574	5,969	10,219	8,870	–2,202
7,726	5,227	9,985	10,693	–3,806
4,585	8,042	11,992	12,396	0,194

Продолжение табл. 1.1

1	2	3	4	5
9,710	6,064	7,185	10,602	–3,404
6,931	7,514	7,992	11,271	–3,777
6,388	8,414	10,139	11,909	0,912
4,312	7,880	6,432	7,964	–1,477
6,690	9,276	7,495	8,645	–3,576
7,004	8,150	7,588	10,152	–1,256
4,419	9,371	7,106	12,839	–2,328
6,937	5,586	10,386	11,819	–2,870
5,736	8,034	11,031	7,941	–1,964
9,343	6,219	10,447	9,399	0,763
6,415	8,815	7,893	11,313	–2,065
5,721	5,292	6,618	10,903	–1,930
9,59	5,800	9,614	8,920	–2,549
4,372	7,140	7,317	11,275	–1,936
6,574	10,832	11,677	8,139	0,440
9,003	8,460	6,808	11,915	–0,765
7,460	5,111	6,873	8,440	0,954
7,415	7,305	8,100	7,561	–0,959
9,985	6,103	6,128	8,269	–2,913
8,001	9,766	9,711	9,899	–3,233
9,142	10,183	8,723	11,937	0,585
9,244	10,796	11,791	10,908	–2,775
5,412	10,407	6,491	12,195	–2,173
7,849	10,188	10,981	8,855	–1,544
9,801	8,388	10,277	10,995	0,456
8,728	5,772	8,698	8,997	–3,582
9,593	5,195	7,819	9,331	–0,888
9,344	10,456	8,330	12,837	0,849
7,388	6,761	7,502	11,502	–3,829
5,299	9,962	7,084	12,433	–2,607
6,735	6,935	11,473	11,127	0,725

Продолжение табл. 1.1

1	2	3	4	5
7,289	6,285	11,444	7,303	–2,262
9,735	6,765	10,679	11,489	–2,871
6,113	9,075	9,127	11,385	–1,340
6,782	5,996	8,672	11,300	–1,553
9,322	9,395	8,915	8,412	–3,121
5,453	10,330	7,081	9,983	–0,501
4,164	6,250	10,593	10,795	0,023
6	7	8	9	10
0,132	–0,271	–2,225	4,662	9,295
–0,825	0,551	–2,919	3,622	9,012
0,055	–1,475	–1,652	5,523	7,162
–1,009	–0,172	–0,608	7,088	8,417
–1,914	1,444	–0,96	6,561	4,108
–1,574	1,022	–3,875	2,188	9,454
–0,068	2,647	–3,506	2,741	3,305
1,238	–0,260	–1,954	5,068	9,176
–0,474	–1,360	–3,995	2,007	8,372
–1,734	2,480	–1,860	5,209	7,525
0,238	–0,396	–3,324	3,015	6,802
0,504	0,149	–3,700	2,451	5,003
0,901	0,914	–0,405	7,392	7,516
–0,475	0,144	–1,486	5,771	4,910
1,522	–0,11	–2,042	4,938	9,497
–0,424	0,954	–1,730	5,406	6,532
–2,896	–1,087	–3,889	2,167	6,268
1,773	2,613	–1,232	6,152	9,494
1,563	–0,286	–2,458	4,312	4,042
–0,480	–0,434	–2,901	3,649	7,936
–2,021	1,005	–1,704	5,444	5,721
–1,788	1,618	–3,376	2,936	4,902
1,833	–1,945	–3,113	3,330	8,750

Продолжение табл. 1.1

6	7	8	9	10
1,145	0,724	–3,917	2,125	7,444
1,430	–0,736	–1,474	5,789	3,017
–1,516	1,301	–3,132	3,302	6,567
–1,577	2,241	–1,647	5,530	5,070
–0,360	–1,830	–3,908	2,138	8,940
1,257	0,654	–0,467	7,299	8,638
–3,000	0,003	–0,480	7,28	8,557
1,300	–0,168	–3,694	2,459	3,289
–2,005	1,967	–1,99	5,016	9,738
1,622	2,580	–3,213	3,181	6,133
1,817	–0,543	–1,898	5,153	5,160
–1,300	1,612	–2,594	4,109	9,702
0,808	–1,157	–2,451	4,323	5,388
–0,129	2,603	–0,385	7,423	6,209
–2,042	2,726	–2,703	3,946	5,728
0,734	–0,687	–1,316	6,026	9,664
0,871	2,346	–2,059	4,912	5,482
–1,117	2,713	–3,539	2,691	8,395
0,199	0,152	–3,682	2,477	5,945
0,156	–0,030	–0,852	6,722	3,071
1,918	0,683	–1,446	5,831	6,398
0,674	2,166	–1,959	5,061	5,528
1,107	1,380	–2,956	3,566	8,073
1,343	–1,323	–3,754	2,368	5,658
0,059	0,886	–1,953	5,070	3,981
–1,732	–1,411	–0,633	7,050	7,605
0,679	0,154	–3,215	3,177	3,521
0,535	1,483	–2,375	4,438	7,434
0,196	2,501	–3,387	2,92	9,818
–1,567	–1,351	–3,448	2,829	7,384
–0,134	1,730	–0,651	7,023	6,379

Продолжение табл. 1.1

6	7	8	9	10
–1,233	0,870	–1,392	5,912	6,739
0,207	1,922	–1,264	6,103	9,284
–1,610	–1,326	–3,532	2,702	7,318
–0,723	–1,487	–3,478	2,782	7,360
–2,080	0,598	–3,993	2,011	9,923
1,257	–0,644	–1,857	5,215	9,048
–1,339	–0,679	–2,021	4,969	7,682
–1,310	1,987	–1,338	5,992	8,801
–2,750	1,757	–0,001	7,998	3,598
–2,711	–0,003	–3,701	2,448	9,571
–0,312	–0,232	–0,493	7,261	6,103
–0,473	0,479	–0,777	6,835	3,133
–0,591	1,888	–2,824	3,765	7,208
–0,226	–0,824	–3,436	2,846	6,351
0,268	–0,984	–3,105	3,342	7,265
1,555	–1,230	–1,626	5,562	8,047
–1,777	0,350	–0,155	7,767	8,432
–1,360	0,712	–3,837	2,244	5,783
–0,772	0,554	–2,158	4,764	7,924
0,986	2,695	–0,433	7,35	9,407
0,474	1,559	–1,285	6,072	5,438
1,108	1,779	–1,284	6,075	4,975
–2,525	1,458	–0,253	7,621	7,069
–2,945	–1,248	–3,619	2,571	5,136
1,077	–1,850	–3,091	3,363	5,213
–2,882	1,165	–0,99	6,514	5,106
11	12	13	14	15
4,145	3,99	4,671	5,286	6,339
3,415	4,840	4,211	7,166	5,407
3,712	4,200	5,316	6,958	7,610
2,573	4,960	3,453	7,694	6,266

Продолжение табл. 1.1

11	12	13	14	15
4,364	3,080	7,472	4,059	8,130
2,961	3,690	5,322	6,519	6,505
1,744	4,67	4,887	5,70	5,802
3,788	2,040	6,422	5,420	6,763
2,536	4,870	7,177	5,624	6,474
3,791	3,440	5,424	5,638	6,57
2,521	3,680	5,594	6,722	7,614
2,748	3,720	5,704	5,843	5,925
2,532	5,690	5,405	6,478	7,276
1,999	3,710	4,277	5,762	7,22
3,094	4,390	5,450	5,215	7,586
2,466	6,040	4,254	4,916	5,452
1,587	2,890	3,947	5,637	7,648
2,452	6,530	3,690	6,665	5,327
1,965	4,080	4,303	6,109	6,686
1,307	2,930	5,463	5,183	7,104
2,988	3,680	6,714	5,616	6,106
2,711	4,790	3,986	6,850	7,868
3,189	3,580	4,557	5,404	5,323
2,357	3,450	4,490	7,434	8,967
2,179	3,760	5,401	5,559	7,955
5,304	3,420	5,383	5,252	6,413
2,578	3,530	6,714	6,853	6,332
3,023	3,780	4,757	7,121	7,152
2,621	3,330	4,411	7,779	6,120
3,261	3,520	3,939	6,435	6,320
0,244	3,230	4,531	5,182	6,034
2,584	3,500	5,208	4,803	6,184
2,202	4,010	4,163	5,555	7,12
3,240	3,440	4,108	5,166	7,575
2,984	3,690	5,585	6,825	6,373

Продолжение табл. 1.1

11	12	13	14	15
3,637	2,760	4,200	6,239	5,766
2,433	3,220	5,499	7,573	7,004
2,871	6,680	5,700	6,745	7,379
3,042	4,220	4,339	5,608	7,099
–0,012	5,830	4,693	6,367	6,160
2,981	3,520	3,651	6,198	5,857
2,225	4,280	5,492	4,235	6,309
3,609	2,560	4,601	5,675	7,550
2,672	4,030	4,058	4,888	6,887
1,417	3,530	5,057	6,901	5,642
2,277	4,140	4,771	7,311	5,214
4,209	3,640	5,516	7,167	6,319
2,743	3,090	4,949	5,011	5,360
4,683	4,070	4,581	6,489	7,306
4,864	4,420	6,770	5,525	7,808
3,674	3,950	4,121	5,682	6,482
1,740	3,810	3,451	5,797	6,578
1,565	4,520	4,278	5,914	6,843
3,327	3,440	4,435	4,522	7,257
5,104	2,950	5,115	6,494	6,852
1,073	4,180	6,516	7,053	4,716
2,873	4,860	4,255	5,595	7,332
3,391	3,460	2,142	6,777	7,087
3,999	2,930	5,549	4,768	6,840
2,906	3,340	4,339	7,222	7,467
3,061	5,020	4,117	6,332	7,179
4,207	2,830	5,044	5,922	5,469
3,281	4,520	4,872	6,373	5,685
1,843	3,260	6,426	6,230	6,957
4,418	3,80	4,871	5,703	7,942
1,697	4,270	5,621	6,124	7,618

Продолжение табл. 1.1

11	12	13	14	15
1,604	4,330	3,106	6,850	6,841
2,980	3,420	3,907	6,913	6,424
2,411	3,310	3,85	5,046	7,983
2,664	3,180	5,605	6,781	7,993
5,291	3,610	5,084	8,304	6,233
2,199	3,410	5,223	5,951	6,163
2,261	2,520	5,283	6,898	6,604
3,086	2,830	4,700	7,884	5,307
3,053	5,030	3,919	5,504	7,327
3,480	5,100	3,081	5,906	6,512
4,150	4,700	4,605	5,357	6,751
3,164	3,790	5,762	6,876	7,511
3,957	4,590	5,177	4,789	8,730
3,708	4,530	5,542	6,185	5,665
6,973	9,143	–2,681	–0,604	–2,014
16	17	18	19	20
9,994	6,838	–1,412	–1,518	–2,363
6,519	8,726	0,118	–1,325	–3,453
9,614	8,194	–0,720	–0,766	–1,465
7,292	9,887	–0,968	–3,786	–3,904
7,068	10,605	–2,064	–1,640	–2,396
9,121	8,796	–1,000	–3,325	–2,993
8,279	9,576	–1,004	–1,154	–1,150
6,320	8,288	–0,742	–1,992	–1,236
8,035	7,181	–0,199	–1,212	–2,251
9,345	7,391	1,788	–1,220	–3,059
8,989	10,870	–1,429	–2,060	–2,720
8,027	10,303	–2,473	–1,541	–4,002
9,521	7,249	–1,75	–2,785	–1,540
8,035	7,704	–2,159	–3,480	–2,791
8,919	7,910	–0,751	–2,721	–3,330

Продолжение табл. 1.1

16	17	18	19	20
8,008	10,538	–2,245	–1,535	–2,590
8,876	9,694	–0,421	–0,892	–3,006
7,385	9,024	–1,342	–1,997	–1,876
7,273	8,733	–0,937	–2,138	–3,854
7,656	9,791	–0,995	–1,204	–2,991
8,654	7,779	–1,815	–2,467	–2,840
8,236	6,607	–0,882	–1,436	–2,478
6,189	9,742	–2,405	–1,414	–2,876
7,677	10,206	–2,201	–3,854	–2,041
8,387	8,451	–2,581	–2,508	–3,346
9,193	8,760	–1,284	–2,117	–2,560
7,271	9,479	–0,346	–1,662	–3,184
6,688	8,212	–1,177	–4,250	–2,285
8,316	9,529	1,060	–2,847	–2,152
8,820	9,607	–0,417	–2,139	–3,165
8,738	8,609	0,209	–1,567	–2,422
9,488	8,897	–2,716	–2,425	–2,790
6,685	9,654	–0,927	–0,548	–1,370
8,771	9,797	–0,981	–2,008	–1,930
8,564	6,688	–1,723	–2,283	–3,706
8,296	9,497	–1,398	–0,641	–3,444
8,977	9,940	–0,553	–2,840	–3,111
8,856	9,225	–2,111	–2,996	–2,625
7,106	7,939	–1,300	–2,333	–2,535
8,931	8,821	–0,859	–2,583	–3,314
9,384	8,011	–1,152	–1,784	–3,345
8,579	9,607	0,326	–1,762	–2,911
6,312	9,299	0,096	–1,933	–3,926
9,315	9,380	–0,995	–2,317	–1,322
6,635	8,429	–1,064	0,316	–5,109
7,650	10,411	1,102	–1,054	–2,050

Продолжение табл. 1.1

16	17	18	19	20
6,255	9,367	–1,066	–1,919	–4,496
8,353	7,243	–0,782	–0,493	–2,990
7,682	9,150	–0,732	–0,566	–3,419
6,787	9,823	–1,754	–1,022	–1,893
6,744	8,347	–1,432	–1,497	–3,218
8,455	8,418	–0,529	–2,355	–3,687
11,177	10,454	–2,390	–2,042	–4,197
10,787	8,938	–0,027	–2,516	–3,762
7,071	6,785	–1,098	0,458	–2,066
8,678	8,290	0,169	–0,492	–2,446
8,253	9,744	0,166	–1,255	–2,971
7,223	8,539	–0,260	–2,506	–2,175
6,845	8,604	–1,800	–0,635	–3,952
9,396	11,459	–1,717	–1,498	–2,838
7,744	7,266	–1,079	–1,754	–4,177
8,913	9,721	–2,695	–3,306	–2,161
8,553	9,018	–3,010	–2,909	–2,295
7,653	9,985	–1,980	–3,714	–2,769
7,610	9,816	–2,138	–1,286	–2,973
8,248	8,166	–1,017	–3,150	–2,801
8,904	9,457	–2,337	–3,062	–3,374
6,920	9,084	–0,648	–1,999	–0,437
8,559	8,705	–0,936	–1,833	–2,084
8,616	7,469	–0,263	–1,501	–5,213
9,408	7,813	–0,045	–2,973	–1,363
8,005	9,366	–2,692	–1,177	–3,339
7,097	8,391	–0,289	–1,409	–4,66
6,885	8,563	0,829	–2,697	–2,792
8,059	9,343	–0,331	–3,504	–4,077
7,023	7,742	0,069	–1,503	–3,529
7,848	8,247	–0,619	–3,041	–3,627

Продолжение табл. 1.1

16	17	18	19	20
7,514	9,744	–0,434	–1,72	–2,405
7,664	9,814	0,751	–3,376	–2,583
21	22	23	24	25
0,482	0,084	2,159	0,265	1,271
0,193	0,703	1,080	1,090	0,894
0,155	0,121	1,438	0,402	1,641
0,013	0,544	0,591	0,518	0,282
1,260	0,983	0,171	0,603	1,013
0,150	2,078	2,959	1,127	0,207
0,078	1,086	0,666	2,294	0,837
2,303	0,588	0,503	1,225	1,819
0,711	1,230	0,223	1,736	1,523
0,206	0,198	0,742	1,812	1,864
0,425	1,663	0,639	0,898	1,589
1,364	1,988	0,441	0,077	0,402
0,163	0,779	0,212	2,925	0,221
1,221	0,305	0,273	0,019	0,129
1,259	1,682	2,021	0,181	1,331
2,423	3,319	0,349	0,014	0,529
1,794	0,405	1,925	1,230	1,233
1,067	0,631	0,337	0,087	0,543
3,159	0,093	1,179	1,422	0,237
0,544	0,069	1,063	1,163	3,575
0,921	0,001	0,312	0,328	0,427
1,438	0,065	1,894	3,086	3,270
1,077	0,014	1,519	0,128	0,306
0,910	0,033	0,003	1,229	1,360
4,527	0,566	0,903	2,503	2,705
0,735	0,754	0,416	0,588	1,856
2,641	2,070	1,813	0,143	0,559
0,468	0,636	1,573	1,095	0,823

Продолжение табл. 1.1

21	22	23	24	25
1,990	0,119	0,199	0,284	1,175
0,066	1,069	0,181	0,929	0,209
0,035	1,904	0,086	1,942	0,513
0,041	1,475	0,193	1,239	2,711
0,502	2,498	0,762	0,298	1,289
0,771	1,121	0,309	0,379	1,562
0,642	1,182	0,012	0,595	1,260
3,791	2,595	0,420	0,343	1,411
0,263	1,373	2,831	1,363	0,052
2,213	1,853	0,020	1,099	0,522
0,715	1,276	3,677	1,916	1,497
0,846	1,824	0,352	1,166	1,023
0,476	3,272	0,409	0,485	3,251
2,329	0,679	0,001	0,106	0,176
0,050	1,729	1,622	0,510	2,127
0,716	0,870	1,103	0,340	3,111
0,921	0,564	0,371	0,201	0,018
2,956	0,734	2,630	1,829	0,684
0,802	0,339	1,390	1,294	2,468
0,692	0,644	1,329	0,644	0,600
2,661	0,317	1,691	0,027	1,095
0,714	2,326	0,313	0,219	1,487
1,240	0,682	0,176	1,853	0,898
0,116	1,594	0,300	0,917	0,048
0,910	0,453	1,154	0,330	0,95
1,249	3,021	2,273	0,430	0,882
0,071	2,015	0,507	1,140	1,237
2,781	1,031	1,516	0,339	0,885
0,846	0,028	0,055	1,662	0,119
0,182	0,550	2,005	0,199	0,435
0,550	3,991	1,928	1,427	0,009

Продолжение табл. 1.1

21	22	23	24	25
0,563	0,956	1,050	2,370	0,497
0,002	1,693	3,846	1,553	1,526
0,405	0,230	0,480	0,727	1,875
0,154	0,146	0,790	0,195	0,087
0,135	0,035	0,035	0,429	1,407
1,447	0,104	2,504	0,144	1,007
0,444	0,145	0,186	1,174	0,711
0,034	0,572	0,339	0,407	0,115
0,238	2,050	0,799	1,100	2,483
0,070	3,424	1,193	0,945	0,474
0,116	0,095	0,946	0,028	0,031
0,572	1,226	1,385	0,287	3,373
1,530	0,190	1,711	0,099	1,278
0,786	1,132	0,092	0,374	0,056
0,601	1,541	0,097	2,987	1,056
0,045	1,223	0,249	0,290	1,488
1,044	0,387	0,652	0,314	0,631
0,769	1,796	0,809	0,333	0,715
0,120	0,311	0,722	1,446	1,738
1,418	0,118	1,714	0,699	0,357
3,598	1,569	0,267	0,458	0,218
26	27	28	29	30
0,468	1,062	0,813	0,106	0,368
0,833	0,673	1,308	0,152	0,226
0,493	2,253	0,533	0,52	0,096
0,921	1,006	0,165	0,256	3,770
1,527	0,373	0,274	1,843	0,284
1,255	0,504	3,463	0,081	0,855
0,534	0,073	2,091	3,135	0,347
0,165	1,055	0,671	0,125	1,657
0,683	2,056	6,708	0,265	4,294

Продолжение табл. 1.1

26	27	28	29	30
1,373	0,110	0,626	0,436	2,126
0,434	1,137	1,777	0,610	0,339
0,356	0,844	2,589	1,251	0,692
0,248	0,540	0,107	0,438	0,548
0,683	0,847	0,464	1,299	0,380
0,100	0,973	0,714	0,075	0,021
0,663	0,526	0,566	0,684	0,918
3,869	1,700	3,582	0,762	0,374
0,047	0,081	0,368	0,075	0,992
0,091	1,070	0,954	1,904	2,371
0,685	1,161	1,292	0,349	0,278
1,631	0,509	0,555	0,945	0,190
1,417	0,323	1,858	1,303	2,480
0,034	4,513	1,506	0,197	0,708
0,188	0,607	3,873	0,454	0,021
0,121	1,375	0,460	6,000	1,151
1,214	0,415	1,528	0,674	0,299
1,257	0,165	0,53	1,219	0,004
0,639	3,383	3,771	0,164	0,64
0,161	0,633	0,124	0,216	0,498
9,478	0,915	0,128	0,231	0,048
0,151	1,004	2,570	3,188	0,997
1,614	0,232	0,688	0,038	0,148
0,079	0,088	1,626	0,804	0,360
0,037	1,233	0,643	1,176	0,038
1,079	0,325	1,046	0,043	0,228
0,272	1,780	0,949	1,075	0,000
0,555	0,083	0,101	0,780	0,131
1,652	0,056	1,126	0,942	4,388
0,292	1,337	0,399	0,049	0,055
0,256	0,14	0,723	1,037	0,256

Продолжение табл. 1.1

26	27	28	29	30
0,977	0,059	2,161	0,261	1,148
0,447	0,843	2,533	0,866	0,067
0,460	0,931	0,239	4,589	0,247
0,017	0,623	0,449	0,723	0,305
0,308	0,182	0,673	1,019	0,063
0,197	0,392	1,343	0,322	0,983
0,141	1,999	2,791	0,969	0,285
0,491	0,549	0,670	1,965	0,63
1,372	2,138	0,172	0,419	1,148
0,307	0,842	1,628	2,599	0,253
0,347	0,361	0,901	0,457	0,955
0,448	0,105	1,875	0,026	3,229
1,250	2,042	1,980	0,468	0,156
0,556	0,293	0,178	0,728	0,239
1,040	0,555	0,428	0,627	1,895
0,444	0,243	0,380	0,108	1,711
1,280	2,003	2,146	0,483	2,311
0,787	2,278	2,037	0,473	0,619
1,693	0,655	6,321	0,011	0,829
0,161	1,305	0,624	0,146	0,597
1,102	1,331	0,704	0,402	0,103
1,085	0,227	0,407	0,188	0,137
2,995	0,286	0,000	2,460	1,179
2,850	0,918	2,594	0,063	0,475
0,621	1,040	0,131	0,813	3,833
0,682	0,701	0,216	3,961	0,671
0,730	0,251	1,224	0,509	0,844
0,589	1,448	1,960	0,737	0,155
0,425	1,594	1,498	0,496	1,467
0,093	1,871	0,522	0,327	1,063
1,408	0,755	0,040	0,254	0,262

Окончание табл. 1.1

26	27	28	29	30
1,114	0,612	3,203	0,922	1,272
0,808	0,672	0,775	0,352	0,647
0,227	0,063	0,115	0,088	0,038
0,364	0,34	0,388	1,055	1,403
0,197	0,280	0,387	1,266	0,328
2,354	0,369	0,065	0,543	0,172
4,501	1,895	2,352	1,187	0,07
0,204	3,503	1,482	1,151	1,884
3,748	0,457	0,284	1,201	0,601

После выбора варианта задания следует приступить к его решению.

1.2. Пример «ручного» выполнения работы

Подобрать закон распределения для непрерывной случайной величины, которая представлена выборкой из 50 значений (табл. 1.2).

Таблица 1.2

0,248	–1,487	–0,059	–0,078	0,194
1,701	0,072	0,318	0,899	–1,302
0,415	0,129	–1,088	0,318	–0,992
1,531	1,409	–0,267	–0,292	–0,056
–1,754	–0,934	0,512	0,085	–0,323
–1,382	–2,364	–1,085	0,367	0,529
–0,443	1,73	0,349	–0,882	0,757
–0,29	–0,449	–0,329	0,13	–0,882
–0,957	–0,095	–0,361	–0,54	0,229
–1,255	0,634	0,49	–1,088	1,028

Подбор закона распределения осуществляется в несколько этапов:

1. Группируем выборку, то есть представляем ее в компактном виде. Для этого организуем таблицу со следующими графами (табл. 1.3):

– номер интервала;

– левая граница интервала;

– правая граница интервала;

– центр интервала;

– абсолютная частота;

– относительная частота;

– нормированная частота.

Приведём некоторые рекомендации по выбору числа интервалов группировки m. На практике, как правило, используют значение m в границах $6–8 \le m \le 20–25$. Уточнить значение m для выборки объёма n можно, например, по формуле $m = 1 + \log_2 n$ либо $m = 1 + 3,32 \lg n$. Соответствующие оценки для нашего примера ($n = 50$) дают $6 \le m \le 7$, рекомендуем выбрать $m = 7$.

Далее заполняем таблицу, выполнив следующие действия:

– Находим $x_{\min} = –2,364$; $x_{\max} = 1,730$. Размах выборки: $d = x_{\max} – x_{\min} = 4,094$. Длина интервала группировки: $h = d/m = 0,585$.

– Записываем в таблицу границы интервалов. Первый интервал: $[x_{\min}, x_{\min} + h = x_1] = [–2,364; –1,779]$; второй: $(x_1, x_1 + h] = (–1,779; –1,194]$ и так далее. Здесь мы учли, что каждая внутренняя граница входит в два соседних интервала. Поэтому для определённости считаем, что каждый интервал содержит свою правую границу, то есть замкнут справа. В особом положении оказывается первый интервал – он замкнут и слева, и справа. Эта особенность будет учтена ниже, в п. 3. Заметим, что замыкать интервалы можно было и слева – принципиально ничего не меняется, только особым (замкнутым с двух сторон) при этом станет последний интервал.

– Центры интервалов: $x_k^* = (x_{k-1} + x_k)/2$, где x_{k-1} – левая, а x_k – правая границы интервала.

– Для каждого интервала подсчитываем абсолютные частоты n_k – число выборочных данных, попадающих в k-й интервал.

– Вычисляем относительные частоты $W_k = n_k / n$, где n – количество выборочных данных.

– Вычисляем нормированные частоты $H_k = W_k / h$, здесь h – длина интервала группировки.

Выполнив указанные выше действия для нашего примера, получим следующую таблицу результатов (табл. 1.3).

Таблица 1.3

№ инт.	x_{k-1}	x_k	x_k^*	n_k	W_k	H_k
1	–2,364	–1,779	–2,072	1	0,02	0,034
2	–1,779	–1,194	–1,487	5	0,1	0,171
3	–1,194	–0,609	–0,902	8	0,16	0,274
4	–0,609	–0,025	–0,317	13	0,26	0,445
5	–0,025	0,56	0,268	15	0,3	0,513
6	0,56	1,145	0,853	4	0,08	0,137
7	1,145	1,73	1,438	4	0,08	0,137

Таким образом, выборка приведена в компактную форму: пятьдесят исходных значений x_1, x_2,..., x_{50} представлены всего 14 числами, а именно: для каждого k-го интервала (k = 1,…, 7) это его центр x_k^* и нормированная частота H_k (при равноинтервальной группировке вместо H_k можно использовать W_k).

В результате группировки данных упрощаются дальнейшие вычисления, особенно при больших объемах выборки.

Представление об эмпирической функции плотности даёт гистограмма (рис. 1.1): на оси OX нужно отложить все 7 интервалов и на каждом из них построить прямоугольник с высотой H_k (W_k или n_k при равноинтервальной группировке).

Гистограмма

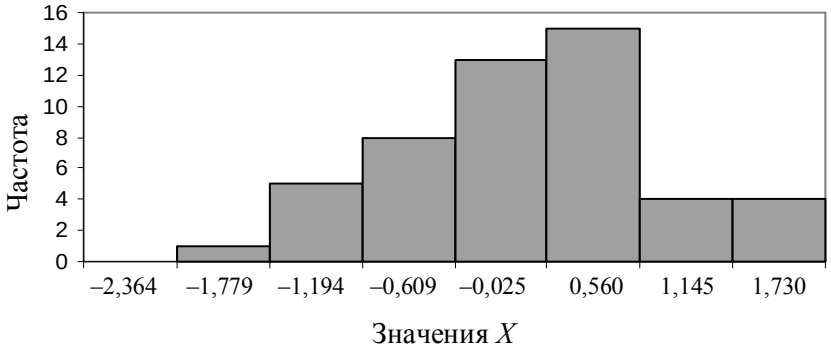

Рис. 1.1

Гистограмму визуально сравнивают с типовыми вариантами (рис. 1.2) для показательного, равномерного и нормального законов распределения соответственно.

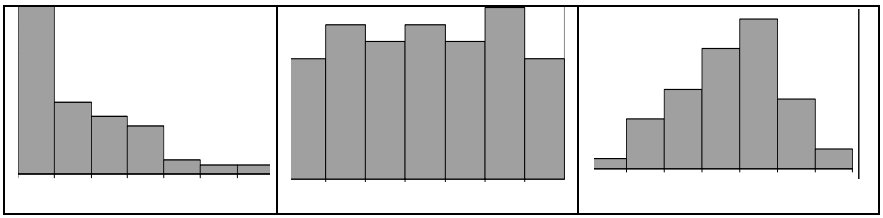

Рис. 1.2

В рассматриваемом примере сравнив данные рис. 1.1 и рис. 1.2, выдвигаем гипотезу: ***представленная выборкой случайная величина распределена по нормальному закону.***

Группировка выборочных данных значительно сокращает время и средства на их статистическую обработку. Наряду с этим явным выигрышем имеются некоторые потери информации при группировке: индивидуальность выборочных данных теряется после усреднения. При исследовании социальных яв-

лений – это плохо, так как индивидуальное мнение каждого отдельного человека заменяют некоторым осреднённым по группе мнением «типичного представителя». Напротив, при обработке данных физического эксперимента усреднение измерений по интервалу может оказаться вполне целесообразным, так как оно в некоторой степени сглаживает неизбежные ошибки отдельных измерений.

2. Числовые характеристики случайной величины.

Выборочное среднее

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{50}\sum_{i=1}^{50} x_i = -0{,}139.$$

Среднее значение по сгруппированным данным

$$\bar{x}_{gr} = \frac{1}{n}\sum_{k=1}^{m} x_k^* \cdot n_k = \frac{1}{50}\sum_{k=1}^{7} x_k^* \cdot n_k =$$

$$= (-2{,}072 - 1{,}4875 - 0{,}9028 - 0{,}31713 + 0{,}26815 +$$
$$+0{,}8534 + 1{,}4384)/50 = -0{,}153.$$

Относительная погрешность в вычислении среднего за счет замены выборки вариационным рядом

$$\delta = \left|1 - \frac{\bar{x}_{gr}}{\bar{x}}\right| 100\ \% \approx 10\ \%.$$

Дисперсия по выборке и по сгруппированным данным соответственно:

$$D = \frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})^2 = 0{,}794,\ D_{gr} = \frac{1}{50}\sum_{k=1}^{7}(x_k^* - \bar{x})^2 n_k = 0{,}685.$$

Относительная погрешность в вычислении дисперсии при замене выборки вариационным рядом составляет около 14 %.

Исправленное среднее квадратическое отклонение

$$\sigma_{gr} = \sqrt{\frac{n}{n-1} D_{gr}} = 0{,}836.$$

Числовые характеристики случайной величины, выражающие различные свойства закона распределения, необходимы

(а зачастую и достаточны) для решения многих практически важных задач. Кроме того, подавляющее большинство используемых в статистических приложениях теоретических законов распределения (биномиальный, нормальный, показательный…) могут быть однозначно восстановлены по одной-двум своим числовым характеристикам, например, по среднему значению и дисперсии.

3. Построение теоретического аналога для гипотетического закона распределения.

Запишем теоретическую функцию плотности $f(x)$ и теоретическую функцию распределения $F(x)$ согласно выдвинутой гипотезе о виде распределения случайной величины. При этом в качестве параметров распределения используем выборочные числовые характеристики, найденные выше. Для большинства практически важных распределений достаточно знать два параметра μ и σ – среднее значение и среднее квадратическое отклонение.

В п. 2 эти параметры были вычислены в двух подходах: по исходной выборке и по сгруппированным данным. Какой вариант выбрать? Первый, более точный вариант, берут при решении задачи с использованием специализированных компьютерных программ (MS Excel или OO Calc). В нашем случае при «ручных» расчетах числовые характеристики μ и σ проще вычислять по сгруппированным данным

$$\mu = \overline{x}_{gr} = \frac{1}{n}\sum_{k=1}^{m} x_k^* n_k = \frac{1}{50}\sum_{k=1}^{7} x_k^* n_k \ ,$$

$$\sigma = \sigma_{gr} = \sqrt{\frac{n}{n-1} D_{gr}} = \sqrt{\frac{50}{49} D_{gr}} \ .$$

Найденные параметры «привязывают» теоретические функции $f(x)$ и $F(x)$ к наблюдаемой области изменения случайной величины, то есть к выборочным данным.

Напомним функцию плотности распределения $f(x)$ и функцию распределения $F(x)$ для практически важных случаев. $f(x)$

и $F(x)$ потребуются в дальнейшем при вычислении вероятности попадания случайной величины в заданный интервал значений.

Показательное распределение характеризуется одним параметром: $\lambda = 1/\mu$. В этом случае функция плотности распределения и функция распределения:

$$f(x) = \begin{cases} 0, x \prec 0 \\ \lambda \cdot e^{-\lambda \cdot x}, x \geq 0 \end{cases}, F(x) = \begin{cases} 0, x \prec 0 \\ 1 - e^{-\lambda \cdot x}, x \geq 0 \end{cases}.$$

Равномерное распределение имеет два параметра:

$$a = \mu - \sqrt{3} \cdot \sigma, \quad b = \mu + \sqrt{3} \cdot \sigma .$$

Функция плотности и функция распределения в этом случае:

$$f(x) = \begin{cases} 0; x < a, x > b \\ \dfrac{1}{b-a}; a \leq x \leq b \end{cases}, F(x) = \begin{cases} 0; x < a, x > b \\ \dfrac{x-a}{b-a}; a \leq x \leq b \end{cases}.$$

Два параметра μ и σ характеризуют нормальное распределение, теоретические функции $f(x)$ и $F(x)$ для него:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], F(x) = 0,5 + \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Здесь $\Phi(z)$ – функция Лапласа, где $z = (x - \mu) / \sigma$. Таблицу значений $\Phi(z)$ можно найти в любой книге по теории вероятностей и математической статистике. Следует учитывать, что функция Лапласа – нечётная: $\Phi(-z) = -\Phi(z)$, а $\Phi(z > 4) \approx 0,5$.

В нашем примере точечные оценки параметров равны: $\mu = -0,153$, $\sigma = 0,836$. Тогда теоретические функции $f(x)$ и $F(x)$:

$$f(x) = \frac{1}{0,836\sqrt{2\pi}} \exp\left[\frac{-(x+0,153)^2}{2 \cdot (0,836)^2}\right], F(x) = 0,5 + \Phi\left(\frac{x+0,153}{0,836}\right).$$

Приведённые выше формулы наглядно представляют, как выполняется «подгонка» закона распределения случайной величины под результаты эксперимента. Во-первых, в качестве параметров теоретического закона распределения использованы выборочные числовые характеристики изучаемой случайной

величины. Во-вторых, в качестве аргумента $F(x)$ указывают наблюдаемые в эксперименте, то есть выборочные значения случайной величины.

Напомним, что в «ручных» расчетах числовые характеристики μ и σ рекомендовано вычислять по сгруппированным данным. Тогда, в рамках единого подхода, и значения аргумента $F(x)$ следует выбирать по сгруппированному варианту выборочных данных. Условимся для определённости, что в качестве значений аргумента x будем использовать правую границу каждого интервала. При этом последний интервал обеспечит расчёт $F(x)$ в точке x_{max}, а вот минимальное значение x_{min} выпадает из рассмотрения, так как для первого интервала его правая граница $x_1 = x_{min} + h$ (замена правой границы на левую не решает проблему, так как в этом случае из рассмотрения выпадет x_{max}). Поэтому введём дополнительный (фиктивный) нулевой интервал, который обеспечит вычисление $F(x)$ в точке x_{min}.

Результаты расчета теоретической функции распределения оформлены в виде таблицы (табл. 1.4), в которой каждый интервал (первая графа) представлен своей правой границей (вторая графа). В третьей графе приведены значения теоретической функции распределения $F(x)$, отвечающие правым границам интервалов группировки.

Таблица 1.4

№ инт.	Правая граница	$F(x)$
0	–2,364	0,004
1	–1,779	0,025
2	–1,194	0,104
3	–0,609	0,291
4	–0,025	0,564
5	0,56	0,805
6	1,145	0,942
7	1,73	0,989

4. Проверка основной гипотезы H_0 того, что исследуемая случайная величина X распределена по предполагаемому в п. 1 закону.

С этой целью можно использовать критерий Пирсона, который для предполагаемого закона распределения сопоставляет выборочные n_k и теоретические n'_k частоты по всем m интервалам группировки

$$x^2 = \sum_{k=1}^{m} \frac{(n_k - n'_k)^2}{n'_k}.$$

Наблюдаемую величину χ^2 сравнивают со значением $\chi^2_{кр}$, которое определим по таблице критических точек χ^2-распределения. Для этого зададимся уровнем значимости α (это вероятность отвергнуть основную гипотезу H_0, когда она верна) и числом степеней свободы $v = m - r - 1$ (m – число интервалов группировки; r – число параметров предполагаемого распределения).

Если $\chi^2 < \chi^2_{кр}$, то нет оснований отвергать нулевую гипотезу H_0, то есть расхождение наблюдаемых и теоретических частот незначительное. Иными словами, выдвинутая гипотеза о виде распределения генеральной совокупности не противоречит данным наблюдений.

Вычислим наблюдаемое значение статистики Пирсона в нашем примере. Так как выборочные частоты n_k уже определены на первом этапе, найдём теоретические частоты n'_k. С этой целью оформим таблицу со следующими графами (табл. 1.5).

Теоретическая вероятность P_k попадания значений случайной величины в k-й интервал равна разности значений теоретической функции распределения на границах этого интервала: $P(a \le X \le b) = F(b) - F(a)$.

В частности, для показательного распределения:

$$P_k = \exp(-\lambda \cdot x_{k-1}) - \exp(-\lambda \cdot x_k).$$

Для равномерного:

$$P_k = (x_k - x_{k-1})/(b-a).$$

В случае нормального распределения:

$$P_k = \Phi\left(\frac{x_k - \mu}{\sigma}\right) - \Phi\left(\frac{x_{k-1} - \mu}{\sigma}\right).$$

Значения P_k для рассматриваемого примера приведены во второй графе табл. 1.5.

Таблица 1.5

№ инт.	Теоретическая вероятность	Теоретическая частота	Статистика Пирсона
1	0,021	1,047	0,002
2	0,079	3,973	0,265
3	0,187	9,327	0,189
7	0,271	13,544	0,023
5	0,244	12,198	0,644
6	0,136	6,797	1,151
7	0,047	2,344	1,17

Теоретические частоты n'_k равны произведению соответствующих вероятностей P_k на объем выборки: $n'_k = P_k \cdot n$. Результаты для нашего примера ($n = 50$) даны в третьей графе табл. 1.5.

В последней графе приведены значения отдельных слагаемых для наблюдаемого значения статистики Пирсона. Просуммировав их, найдем $\chi^2 = 3{,}444$. Критическое значение для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $v = m - r - 1 = 4$ составляет $\chi^2_{кр} = 9{,}488$.

Заключение: так как наблюдаемое значение статистики Пирсона меньше критического, то гипотеза о нормальном распределении случайной величины принимается на заданном уровне значимости.

1.3. Пример выполнения работы в MS Excel и OO Calc

Предварительное замечание. Для работы с большими массивами данных в MS Excel и OO Calc предусмотрен удобный инструмент – именование диапазона размещения данных в электронной таблице.

Чтобы присвоить имя диапазону, нужно:

– выделить его;

– ввести имя диапазона в окне *Имя*;

– нажать клавишу Enter.

В результате заметно облегчается процедура обращения к данным. Например, указание выборки как аргумента встроенной функции выполняется по команде ВСТАВКА – ИМЯ – ВСТАВИТЬ.

Преимущества использования данного инструмента становятся очевидными, если вспомнить, что реальные выборки содержат тысячи элементов.

Поименуем исходные данные в задаче как «выборка». Тогда, например, формула для вычисления наибольшего элемента выборки будет выглядеть так: = MAX (выборка).

Для расчета значений наиболее употребительных функций в Excel и Calc предусмотрен специальный сервис – Мастер функций, который запускается при нажатии кнопки со значком f_x на панели инструментов (либо из пункта меню окна ВСТАВКА – ФУНКЦИЯ). Встроенные функции разделены на категории (тематические группы): математические, финансовые, статистические и так далее. После указания конкретной функции данной категории щёлкаем по кнопке ОК, и имя функции заносится в строку формул вместе со скобками для указания параметров. Исполнение – по нажатию клавиши Enter. Заметим, что параметры также можно вводить в поля диалогового окна Мастера функций.

Реализация в Excel и Calc всех четырёх этапов решения задачи во многом аналогична. Поэтому только при возникновении

особенностей выполнения операций в этих двух табличных процессорах будем приводить для каждого из них отдельный вариант.

Подбор закона распределения случайной величины показан в виде фрагментов рабочего листа для каждого этапа.

1. Группировка выборочных данных.

Как в Excel, так и в Calc минимальный и максимальный элементы выборки находят встроенные функции min (выборка) и max (выборка) соответственно.

Далее диапазон выборочных данных d = max–min разбивают на интервалы группировки, границы которых записывают в рабочую таблицу (первые три графы табл. 1.6).

Таблица 1.6

а) минимум и максимум			
min =	–2,364		
max =	1,73		
б) размах выборки			
d = max–min	4,094		
в) число интервалов группировки			
m =	7		
г) длина интервала			
$h = d/m$	0,585		
Результаты группировки имеют вид			
номер интервала	левая граница	правая граница	частота
0		–2,364	
1	–2,364	–1,779	1
2	–1,779	–1,194	5
3	–1,194	–0,609	8
4	–0,609	–0,025	13
5	–0,025	0,56	15
6	0,56	1,145	4
7	1,145	1,73	4

Выборочные частоты (последняя графа табл. 1.6) удобно рассчитывать с помощью встроенной функции, которая в Excel называется ЧАСТОТА, а в Calc – FREQUENCY.

Последовательность действий:

– выделяем последнюю графу табл. 1.6 под результаты;

– заполняем диалоговое окно функции ЧАСТОТА или FREQUENCY. В качестве аргументов указываем выборку и правые границы интервалов группировки;

– даём команду на исполнение одновременным нажатием трех клавиш Ctrl + Shift + Enter.

Графическое представление выборочных частот в виде гистограммы (рис. 1.3) проще всего выполнить, следуя указаниям Мастера построения диаграмм (команда ВСТАВКА – ДИАГРАММА либо соответствующая кнопка стандартной панели инструментов).

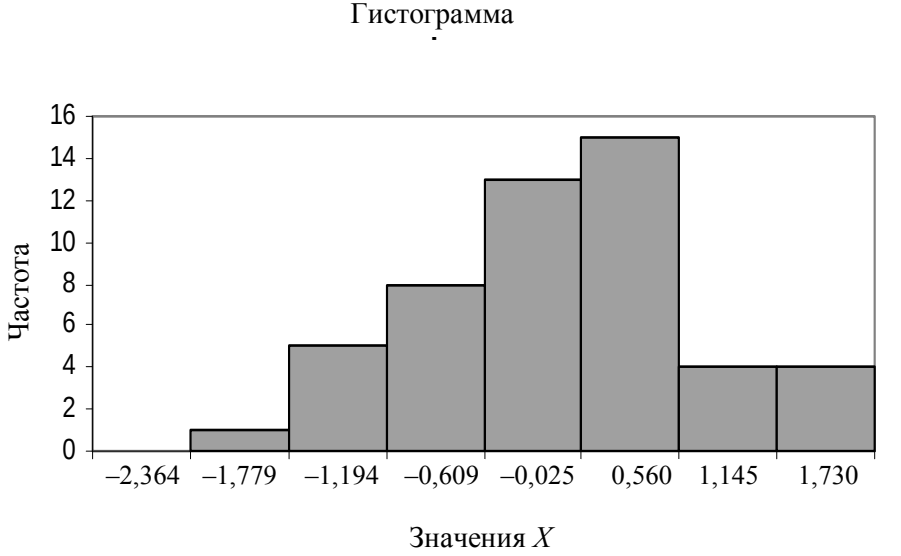

Рис. 1.3

По приведённой выше гистограмме выдвигаем гипотезу ***о нормальном*** распределении случайной величины X.

2. Основные числовые характеристики случайной величины.

Среднее значение μ и среднее квадратическое отклонение σ в Excel находят с помощью встроенных функций:

среднее значение μ: = СРЗНАЧ (выборка);

стандартное отклонение σ: = СТАНДОТКЛОН (выборка);

В этих же целях Calc использует:

среднее значение μ = AVERAGE (выборка);

стандартное отклонение σ = STDEV (выборка).

Как в Excel, так и в Calc названные выше функции следует искать в категории «Статистические».

Найденные числовые характеристики используем для вычисления параметров предполагаемой функции распределения для трёх рассматриваемых здесь вариантов:

нормальное: μ, σ, в нашем примере $\mu = -0{,}139$; $\sigma = 0{,}891$;

показательное: $\lambda = 1/\mu$;

равномерное: $\alpha = \mu - \text{корень}(3) * \sigma$; $b = \mu - \text{корень}(3) * \sigma$.

Заметим, что для вычисления квадратного корня в Calc следует использовать встроенную функцию SQRT(аргумент). Тогда параметры равномерного распределения в Calc:

$\alpha = \mu - \text{SQRT}(3) * \sigma$; $b = \mu - \text{SQRT}(3) * \sigma$.

3. Расчёт значений теоретической функции распределения $F(x)$ согласно предполагаемому закону.

Для нормального и показательного законов расчёт $F(x)$ в Excel ведём с помощью встроенных функций НОРМРАСП (x, μ, σ, истина) и ЭКСПРАСП (x, λ, истина) соответственно. В Calc расчёт $F(x)$ для нормального и показательного законов выполняют с помощью встроенных функций NORMDIST (x, μ, σ, истина) и EXPONDIST (x, λ, истина) соответственно.

Для равномерного распределения расчёт $F(x)$ ведём по формуле $F(x) = (x - a)/(b - a)$.

В качестве значений независимой переменной (x) указываем правые границы интервалов группировки.

В нашем примере для нормального закона распределения с параметрами $\mu = -0{,}139$ и $\sigma = 0{,}891$ получим следующие результаты (табл. 1.7).

Таблица 1.7

№ инт.	Правая граница	Функция распределения
0	–2,364	0,006
1	–1,779	0,033
2	–1,194	0,118
3	–0,609	0,299
4	–0,025	0,551
5	0,560	0,784
6	1,145	0,925
7	1,730	0,982

4. Проверка гипотезы о виде распределения случайной величины с помощью статистики Пирсона:

$$x^2 = \sum_{k=1}^{m} \frac{(n_k - n'_k)^2}{n'_k}.$$

Выборочные частоты n_k уже определены в самом начале, на этапе 1, теперь найдём теоретические частоты n'_k.

Теоретическая частота n'_k для k-го интервала равна произведению объёма выборки ($n = 50$) на P_k – вероятность попадания случайной величины в k-й интервал. Неизвестные величины P_k найдём, используя данные табл. 1.7. А именно, P_k вычисляются как разность значений теоретической функции распределения на границах k-го интервала, то есть $P_k = F_k - F_{k-1}$, здесь номер интервала $k = 1,..., m$.

Результаты описанных выше действий заносим в табл. 1.8. Данные этой же таблицы позволяют осуществить промежуточный контроль за правильностью выполнения задания. Очевидно, что сумма элементов второй графы должна быть приближённо равна единице, сумма элементов третьей графы должна быть приближённо равна объёму выборки.

Таблица 1.8

№ инт.	Теоретическая вероятность	Теоретически частота	Статистика Пирсона
1	0,027	1,330	0,082
2	0,085	4,268	0,126
3	0,181	9,033	0,118
7	0,252	12,615	0,012
5	0,233	11,628	0,978
6	0,141	7,074	1,336
7	0,057	2,84	0,474

Сумма элементов последней графы даёт наблюдаемое значение статистики Пирсона $\chi^2 = 3{,}125$.

С другой стороны, для заданного уровня значимости α и числа степеней свободы ν найдём критическое значение $\chi^2_{кр}$. В Excel величину $\chi^2_{кр}$ даёт встроенная функция ХИ2ОБР(α, ν), в Calc ей соответствует функция CHINV(α, ν).

В нашем случае $\alpha = 0{,}05$ и $\nu = m - r - 1 = 7 - 2 - 1 = 4$. Для этих значений параметров $\chi^2_{кр} = 9{,}488$.

Заключение: так как наблюдаемое значение статистики Пирсона меньше критического, то гипотеза о нормальном распределении случайной величины принимается на заданном уровне значимости.

2. УКАЗАНИЯ К ВЫПОЛНЕНИЮ ЗАДАНИЯ № 2

Достаточно часто в эксперименте наблюдается изменение не одной, а ***двух случайных переменных***. Тогда возникает естественный вопрос – связаны ли эти переменные между собой? Если они связаны, то как? Для ***детерминированных*** переменных положительный ответ устанавливает факт функциональной связи между переменными. Напротив, связь между ***случайными переменными*** проявляется как тенденция при массовых испытаниях. Такая связь называется ***статистической*** и проявляется в том, что изменение одной случайной величины сопровождается изменением закона распределения другой. В частности, если изменение одной случайной величины приводит к изменению только среднего значения другой (без нарушения закона распределения), случайные величины находятся в ***корреляционной зависимости***, которая является предметом изучения в ***корреляционном анализе***. При положительном заключении (то есть между изучаемыми переменными наблюдается достаточно тесная корреляция) переходят к определению ***формы*** этой связи, используя методы ***регрессионного анализа***.

При статистической зависимости каждому конкретному значению объясняющей (независимой, факторной) переменной соответствует некоторое вероятностное распределение объясняемой (зависимой, результирующей) переменной. Поэтому анализируют, как объясняющая переменная влияет на объясняемую переменную «в среднем», то есть определяют условное математическое ожидание результирующего показателя Y при фиксированном значении X объясняющей переменной: $M(Y|x) = F(X)$. Функция $F(X)$ называется ***функцией регрессии Y на X.***

Для отражения того факта, что реальные значения зависимой переменной не всегда совпадают с её условными математическими ожиданиями и могут отличаться при одном и том же значении объясняющей переменной, фактическая

зависимость должна быть дополнена некоторым случайным слагаемым $E(X)$:

$$Y(X) = M(Y|x) + E(X) = F(X) + E(X).$$

Это соотношение представляет сущность регрессионного подхода к математическому моделированию статистической связи двух случайных величин. Здесь $F(X)$ представляет ***доминирующую тенденцию*** в изменении зависимой переменной, а $E(X)$ описывает ***отклонения*** (остатки, «помехи») значений $Y(X)$ от функции регрессии.

В статистической практике мы никогда не располагаем информацией для точного описания функции регрессии $F(X)$, поэтому обычно ограничиваются поиском подходящей её аппроксимации $\hat{Y}(X)$. Выбор общего вида аппроксимирующей функции $\hat{Y}(X)$ основан, конечно, на анализе имеющихся статистических данных. В результате получают линейную, параболическую, гиперболическую и некоторые другие виды регрессий.

Таким образом, статистическое исследование зависимостей представляет функционирование реального объекта через набор количественных показателей (переменных). Корреляционный анализ позволяет выявить характер взаимосвязи между этими переменными. Для взаимосвязанных переменных определяют, какие из них являются результирующими (зависимыми, изучаемыми), а какие представляют условия испытания (являются независимыми, факторными). Задача регрессионного анализа – подбор по результатам наблюдений подходящей функции регрессии, которая описывает изменение результирующего показателя при изменении условий испытания.

Основные этапы выполнения работы № 2

Обычно регрессионный подход осваивают при изучении простейшей, то есть ***линейной связи двух случайных величин***.

Во-первых, линейная модель регрессии основательно проработана в теоретическом плане и поэтому поддержана компьютерной реализацией в большинстве статистических приложений. Во-вторых, линейную модель зачастую используют в качестве начального приближения в процессе построения более сложной и адекватной модели. В-третьих, довольно широкий класс ***нелинейных*** регрессий сводится к линейным путем тождественных математических преобразований. И, наконец, если для изучаемых случайных величин установлен совместный нормальный закон распределения, то доказано, что сама функция регрессии (а не её аппроксимация) имеет линейный вид.

Итак, рассмотрим простую модель линейной регрессии для двух случайных величин.

1. Корреляционный анализ.

Построению регрессионной модели предшествует корреляционный анализ, который оценивает ***наличие и силу статистической связи*** между исследуемыми переменными. Только при положительном заключении переходят к определению ***формы*** этой связи, то есть построению функции регрессии.

Для первичного визуального анализа экспериментальных данных строят ***корреляционное поле*** (его ещё называют ***диаграммой рассеяния***). Для этого на координатную плоскость наносят все пары наблюдений (xi, yi), $i = 1,\ldots, n$. Взаимное положение точек характеризует визуально форму кривой, представляющей эти точки наилучшим образом.

Тесноту (силу) линейной связи оценивает коэффициент корреляции r_{xy}, который вычисляют для выборочных данных по следующим формулам:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2}\cdot\sqrt{\overline{y^2} - \overline{y}^2}}.$$

Связь, оцениваемая величиной r_{xy}, симметрична относительно обеих переменных, а её знак поддаётся содержательной интерпретации. А именно, положительный коэффициент корреляции ($r_{xy} > 0$) означает одинаковый характер тенденции взаимосвязанного изменения случайных величин: с увеличением X наблюдается тенденция увеличения соответствующих индивидуальных значений Y. Отрицательное значение коэффициента корреляции ($r_{xy} < 0$) говорит о противоположной тенденции взаимосвязанного изменения случайных величин: с увеличением X наблюдается тенденция уменьшения соответствующих индивидуальных значений Y. При $r_{xy} = 0$ между X и Y отсутствует ***линейная*** связь.

Какую величину выборочного коэффициента корреляции r_{xy} следует считать достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными?

Для ответа на этот вопрос проверяется основная гипотеза H_0: $\rho = 0$ об отсутствии корреляционной связи X и Y (другими словами: выборочный коэффициент корреляции r_{xy} статистически незначимый и целиком обусловлен случайным колебанием выборки, на основе которой он вычислен).

При выполнении нулевой гипотезы случайная величина $t(\alpha, n-2) = \dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ распределена по закону Стьюдента с $n-2$ степенями свободы, поэтому если окажется, что $t\ (0{,}05;\ n-2) \le \left|\dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}\right|$ (здесь $t\ (0{,}05;\ n-2)$ – 5 % точка распределения Стьюдента с $n-2$ степенями свободы), то гипотеза об отсутствии корреляционной связи отвергается: коэффициент корреляции r_{xy} статистически значим (существенно отличен от нуля), и между X и Y существует линейная статистическая связь.

Если статистическая значимость коэффициента корреляции подтверждена, то оценку тесноты связи X и Y делают с использованием так называемой шкалы Чеддока (табл. 2.1).

Таблица 2.1

Величина $\|r\|$	0,1–0,3	0,3–0,5	0,5–0,7
Характеристика силы связи	слабая	умеренная	заметная
Величина $\|r\|$	0,7–0,9	0,9–0,99	–
Характеристика силы связи	высокая	весьма высокая	–

На практике дальнейшее построение регрессионной модели считают целесообразным при $|r_{xy}| > 0{,}7$.

2. Регрессионный анализ.

Пусть по результатам корреляционного анализа установлена надёжная линейная связь X и Y. Например, прямая линия на рисунке довольно хорошо представляет опытные данные, изображённые точками корреляционного поля (рис. 2.1).

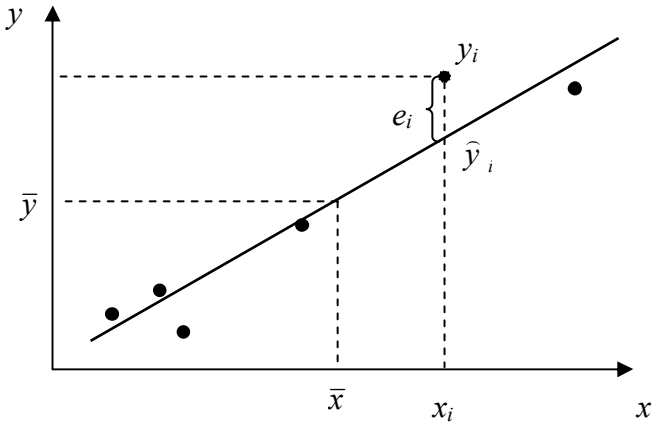

Рис. 2.1

В этом случае для аппроксимации экспериментальных значений (x_i , y_i), $i = 1,\ldots, n$ хорошо подходит линейная функция регрессии

$$\hat{y}(x) = \beta_0 + \beta_1 \cdot x .$$

Здесь β_0 и β_1 – параметры, подлежащие определению.

С этой целью чаще всего используют метод наименьших квадратов (МНК): по выборочным данным находят оценки b_0 и b_1 для теоретических параметров β_0 и β_1, которые обеспечивают наименьшее суммарное отклонение точек (x_i, y_i), $i = 1,\ldots, n$ от выборочной функции регрессии $\hat{y} = b_0 + b_1 \cdot x_i$.

Применение МНК даёт следующие формулы для b_0 и b_1:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}, b_0 = \bar{y} - b_1 \cdot \bar{x} .$$

Для «ручных» расчётов по приведённым выше формулам удобно заготавливать таблицу, содержание которой определено в заголовках граф (табл. 2.2).

Таблица 2.2

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	$e_i^2 = (y_i - \hat{y}_i)^2$
сумма							
среднее							

При заполнении этой таблицы довольно часто ошибаются, считая, что $\overline{x^2} = \bar{x}^2$. Рекомендуем расписать несколько первых слагаемых при вычислении средних вручную, тогда легко убедиться в том, что $\overline{x^2} \neq \bar{x}^2$.

3. Оценка качества уравнения регрессии.

Общее качество уравнения регрессии оценивают с помощью ***коэффициента детерминации***

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}.$$

Измеряя ***долю*** общей дисперсии (знаменатель), которая объясняется уравнением регрессии (числитель), величина R^2 варьируется в пределах от нуля до единицы. Чем ближе R^2 к единице, тем лучше уравнение регрессии аппроксимирует исходные данные (тем ближе экспериментальные точки расположены к линии регрессии). Если $R^2 = 1$, все экспериментальные точки лежат на линии регрессии, что соответствует ***функциональной*** зависимости между X и Y. Если $R^2 = 0$, то изменение Y происходит не вследствие изменения X, а под воздействием случайных либо не учтённых в модели факторов.

Для ***парной линейной регрессии*** коэффициент детерминации равен квадрату коэффициента корреляции $R^2 = r_{xy}{}^2$, причём значимость коэффициента детерминации следует из значимости коэффициента корреляции.

4. Прогнозирование по уравнению регрессии.

Точечным прогнозом называется значение $\hat{y}_p$, найденное путём подстановки заданной величины x_p в уравнение регрессии, то есть $\hat{y}_p = b_0 + b_1 x_p$.

Точность прогноза убывает по мере удаления x_p от $\overline{x}$. Напротив, увеличение объёма выборки n способствует улучшению качества прогноза.

На практике уравнение регрессии считается пригодным для прогнозирования при $R^2 > 0{,}5$. В этом случае оно объясняет более половины дисперсии Y его зависимостью от X. Оставшаяся доля дисперсии Y, равная $(1 - R^2)$, объясняется либо случайными, либо не учтёнными в модели факторами.

2.1. Варианты второго задания

Варианты второго задания приведены в табл. 2.3.

Таблица 2.3

№ 1	*X*	74	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 2	*X*	70	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 3	*X*	74	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 4	*X*	71	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 5	*X*	75	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 6	*X*	76	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 7	*X*	77	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 8	*X*	78	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 9	*X*	69	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 10	*X*	70	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 11	*X*	71	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 12	*X*	72	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 13	*X*	73	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 14	*X*	74	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2

Окончание табл. 2.3

№ 15	*X*	75	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 16	*X*	76	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 17	*X*	71	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 18	*X*	72	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 19	*X*	73	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 20	*X*	70	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 21	*X*	74	86	117	125	150
	Y	0,5	0,7	1,1	1,5	2
№ 22	*X*	70	85	116	123	150
	Y	0,6	0,8	1,2	1,5	2
№ 23	*X*	74	88	118	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 24	*X*	71	86	110	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 25	*X*	75	86	111	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 26	*X*	76	85	113	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 27	*X*	77	88	115	120	145
	Y	0,5	0,8	1,1	1,5	1,9
№ 28	*X*	78	86	115	125	150
	Y	0,5	0,7	1	1,5	1,8
№ 29	*X*	69	86	114	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 30	*X*	70	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2

После выбора варианта задания следует приступить к его решению.

2.2. Пример «ручного» выполнения работы

Зависимость между уровнем помех *Y* и расстоянием от источника сигнала *X* представлена таблицей данных (табл. 2.4).

Таблица 2.4

X	2	6	10	14	18
Y	9	10	12	19	20

Необходимо оценить уровень помех на расстоянии в двадцать единиц.

1. Корреляционный анализ.

По виду корреляционного поля заключаем, что линейная функция пригодна для аппроксимации зависимости *Y* от *X* (рис. 2.2).

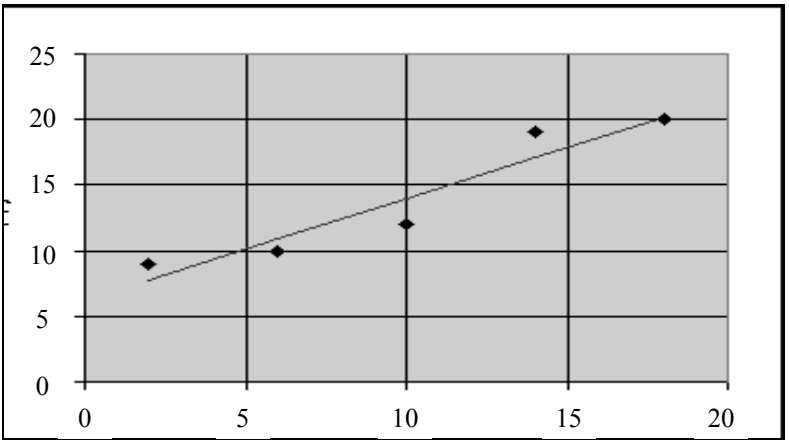

Рис. 2.2

Коэффициент корреляции

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = 0{,}952 .$$

Проверим его статистическую значимость.

Для уровня значимости α = 0,05 и числа степеней свободы $v = n - 2 = 3$ по таблице критических точек распределения Стьюдента найдем значение $t(\alpha,\ n - 2) = 3{,}18$. Наблюдаемое значение $t = 0{,}952 \cdot \sqrt{3} / \sqrt{1 - 0{,}952^2} \approx 5{,}39$, то есть коэффициент корреляции r_{xy} является существенным (статистически значим).

Заключение: между наблюдаемыми переменными есть положительная линейная корреляция: с увеличением расстояния от источника сигнала уровень помех в среднем возрастает. Теснота линейной связи оценивается по шкале Чеддока как весьма высокая.

2. Регрессионный анализ.

Для записи выборочной функции регрессии $\hat{y}_p = b_0 + b_1 x$ в явном виде оценим её коэффициенты. Промежуточные вычисления оформим в виде таблицы (табл. 2.5).

Таблица 2.5

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	e_i^2
1	2	9	4	18	81	7,8	1,44
2	6	10	36	60	100	10,9	0,81
3	10	12	100	120	144	14	4
4	14	19	196	266	361	17,1	3,61
5	18	20	324	360	400	20,2	0,04
сумма	50	70	660	824	1086	70	9,9
среднее	10	14	132	164,8	217,2	14	1,98

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = 0{,}775; b_0 = \bar{y} - b_1 \cdot \bar{x} = 6{,}25 .$$

Следовательно, выборочное уравнение регрессии имеет вид

$$\hat{y} = 6{,}25 + 0{,}775 \cdot x .$$

Угловой коэффициент b_1 показывает, что при увеличении расстояния от источника сигнала на 1 единицу уровень помех увеличится в среднем на 0,775 единиц. Постоянная b_0 показывает, что при нулевом расстоянии уровень помех составит в среднем 6,25 единиц. Знаки коэффициентов уравнения регрессии соответствуют физическому смыслу описываемого явления.

3. Оценка качества уравнения регрессии.

Вычислив коэффициент детерминации $R^2 = r_{xy}{}^2 \approx 0{,}91$, заключаем, что 91 % дисперсии (разброса) результативного признака объясняется уравнением регрессии, а доля необъяснённой дисперсии составляет лишь 9 %.

Значимость коэффициента детерминации следует из установленной выше значимости коэффициента корреляции.

Близость величины коэффициента R^2 к единице свидетельствует о высоком качестве уравнения регрессии.

4. Прогнозирование по уравнению регрессии.

Найдем точечный прогноз при $x_p = 20$ единиц:

$$\hat{y} = 6{,}25 + 0{,}775 \cdot 20 = 21{,}75 .$$

На расстоянии в 20 единиц от источника сигнала ожидаемый уровень помех будет составлять в среднем 21,75 единицы.

Вывод: при исследовании связи между уровнем помех в сигнале *Y* и расстоянием до его источника *X* по выборочным данным построено уравнение регрессии $\hat{y} = 6{,}25 + 0{,}775 \cdot x$, которое обладает хорошими статистическими свойствами и, следовательно, может быть рекомендовано для практического применения.

2.3. Пример выполнения работы в MS Excel и OO Calc

Построение уравнения регрессии и его статистический анализ, аналогичный по полноте рассмотренному выше «ручному» варианту, обеспечивают стандартные средства Excel и Calc, что заметно упрощает решение практических задач.

Корреляционное поле, представляющее экспериментальные данные, строим с помощью Мастера диаграмм (тип диаграммы – «Точечная») – рис. 2.3.

Корреляционное поле

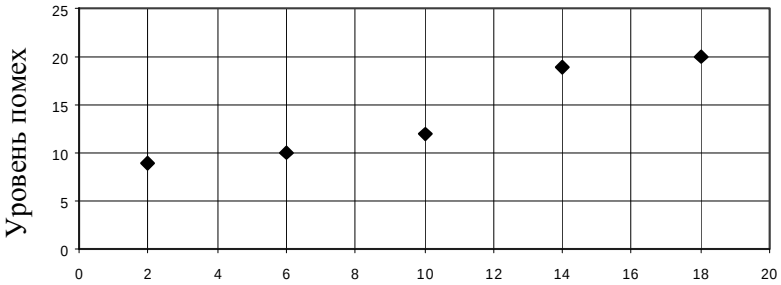

Рис. 2.3

Визуально определяем, что линейная функция пригодна для аппроксимации зависимости между *X* и *Y*.

Помечаем одну из точек диаграммы и из контекстного меню выбираем команду Excel: ДОБАВИТЬ ЛИНИЮ ТРЕНДА. В Calc аналогичная по смыслу команда – ВСТАВИТЬ КРИВУЮ РЕГРЕССИИ.

В диалоговом окне присутствуют две закладки, первая из них содержит образцы форм линии регрессии (тренда). В нашем примере заказываем линейную форму.

Переходим во вторую закладку – ПАРАМЕТРЫ. При использовании Excel устанавливаем флажки на опциях «Показать уравнение на диаграмме» и «Поместить на диаграмму величину достоверности аппроксимации». В Calc эти указания соответственно «Показать уравнение» и «Показать коэффициент корреляции R^2».

Результаты расчетов приведены на рис. 2.4: здесь изображена аппроксимирующая исходные данные линия регрессии, выписано её уравнение, а также указана величина коэффициента детерминации R^2.

Рис. 2.4

К этим результатам необходимо добавить только проверку на значимость коэффициента корреляции, его наблюдаемое значение – корень квадратный из коэффициента детерминации R^2. Тогда наблюдаемое значение t-статистики:

$$t = \frac{\sqrt{R^2(n-2)}}{\sqrt{1-R^2}} \approx \frac{\sqrt{3 \cdot 0{,}91}}{\sqrt{1-0{,}91}} \approx 5{,}4\,.$$

Для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $v = n - 2 = 3$ по таблице критических точек распределения

Стьюдента найдем предельное значение $t(\alpha, n-2) = 3{,}18$. Так как, $|t| > t(\alpha, n-2)$, то коэффициент корреляции является существенным (статистически значим).

Критические точки распределения Стьюдента для заданного уровня значимости α и числа степеней свободы $v = n - 2$ в Excel находят с помощью встроенной функции СТЬЮДРАСПОБР (α, v). Соответствующая ей в Calc функция – TINV(α, v).

Таким образом, полученные в электронных таблицах результаты подтверждают правильность «ручного» варианта решения поставленной задачи. Понятно, что вывод также сохраняется.

В заключение отметим, что табличные процессоры Excel и Calc располагают и другими более мощными инструментами обработки экспериментальных данных. Так, достаточно полный статистический материал для ***линейных по параметрам*** регрессионных моделей даёт встроенная функция ЛИНЕЙН (в Excel), в Calc аналогичная функция называется LINEST.

Для оценки тесноты линейной связи между изучаемыми признаками в Excel используют функцию КОРРЕЛ, а в Calc – CORREL.

Сложные ***многофакторные задачи***, нацеленные на планирование ***активного*** эксперимента, можно анализировать в Excel с применением инструмента РЕГРЕССИЯ из Пакета анализа.

ЗАКЛЮЧЕНИЕ

Исходя из учебных целей, в качестве базовых вычислительных инструментов выбраны MS Excel и OO Calculation.

MS Excel во многом более предпочтителен для студентов-заочников, которые на своих рабочих местах традиционно работают в Microsoft Office. С другой стороны, современная тенденция к использованию открытого программного продукта, в частности Open Office, требует формирования соответствующих навыков работы при обучении в вузе. В результате для каждого читателя обеспечена возможность выбора такого прикладного пакета, который он может использовать напрямую без излишних разъяснений общего характера. Особо отметим, что оба приложения как Excel, так и Calc располагают широкой гаммой инструментов, напрямую предназначенных для статистической обработки данных и численного экспериментирования.

Конечно, на рынке современных программных продуктов имеются и другие предложения, относящиеся, правда, к профессиональному классу. Из зарубежных статистических пакетов выделяются Statistica, S-Plus, SPSS, StatGraphics. Из отечественных – STADIA, Эвриста, Мезозавр, Олимп: СтатЭксперт, Статистик-Консультант, Forecast Expert. В перечисленных системах общего назначения присутствуют, как правило, средства описательной статистики, методы регрессионного анализа и некоторые инструменты исследования временных рядов.

При выборе конкретного варианта следует ориентироваться на специфику решаемых задач, стоимость программного продукта, соответствие его уровня квалификации практикующего инженера, доступность сопроводительной документации и некоторые другие моменты. В общем и целом отечественные вычислительные пакеты требуют относительно меньшей квалификации пользователей, чем профессиональные пакеты зарубежного производства.

СПИСОК ЛИТЕРАТУРЫ

1. *Гмурман, В.Е.* Теория вероятностей и математическая статистика: учеб. пособие для вузов / В.Е. Гмурман. – М. : Высшая школа, 2009. – 480 с.
2. *Просветов, Г.И.* Теория вероятностей и математическая статистика. Задачи и решения / Г.И. Просветов. – М. : Альфа Пресс, 2009. – 272 с.
3. *Дорофеева, Н.С.* Первичная обработка выборочных данных: учеб. пособие / Н.С. Дорофеева. – Томск : Изд-во Том. гос. архит.-строит. ун-та, 2009, – 61 с.
4. *Слободской, М.И.* Теория вероятностей и математическая статистика : метод. указания и варианты заданий / М.И. Слободской. – Томск : ТГАСУ, 2001. – 87 с.
5. *Протасов, К.В.* Статистический анализ экспериментальных данных / К.В. Протасов. – М. : Мир, 2005. – 142 с.
6. *Анализ* статистических данных с использованием Microsoft Excel для Office XP [пер. с англ.] / М.Р. Мидлтон ; под ред. Г.М. Кобелькова. – М. : БИНОМ. Лаборатория знаний, 2005. – 296 с.
7. *Тюрин, Ю.Н.* Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров ; под ред. В.Э. Фигурнова. – 3-е изд., перераб. и доп. – М. : ИНФРА – М, 2003. – 544 с.
8. *Чекотовский, Э.В.* Графический анализ статистических данных в Microsoft Excel 2000 / Э.В. Чекотовский. – М. : Издательский дом «Вильямс», 2002. – 464 с.
9. *Руководство* пользователя Open Office.org 2. – СПб : БХВ-Петербург, 2007. – 320 с.
10. *Макаров, Е.* Инженерные расчёты в MathCAD 14 / Е. Макаров. – СПб. : Питер, 2007. – 592 с.
11. *Боровиков, В.П.* Прогнозирование в системе Statistica в среде Windows / В.П. Боровиков, Г.И. Ивченко. – М. : Финансы и статистика, 2006. – 367 с.

Учебное издание

Мамонтов Геннадий Яковлевич
Иконникова Ирина Александровна

АНАЛИЗ ДАННЫХ В MS EXCEL И OO CALC

Учебно-методическое пособие

Редактор Е.А. Кулешова
Технический редактор А.А. Маракулина

Подписано в печать 07.10.2010 г. Формат 60×84.
Усл. печ. л. 3,49. Уч.-изд. л. 3,16. Тираж 140 экз. Зак. № 372.
Изд-во ТГАСУ, 634003, г. Томск, пл. Соляная, 2.
Отпечатано с оригинал-макета в ООО ТГАСУ.
634003, г. Томск, ул. Партизанская, 15.