Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ «РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

К.Л. Восканян, А.Д. Кузнецов, О.С. Сероухова

АВТОМАТИЧЕСКИЕ МЕТЕОРОЛОГИЧЕСКИЕ СТАНЦИИ

Часть 2. Цифровая обработка данных автоматических метеорологических станций

Практикум

РГГМУ

Санкт-Петербург
2015

УДК 91(075.8)

Рецензент: Г.Г. Щукин, д-р физ.-мат. наук, проф., Военно-космическая академия им. А.Ф. Можайского.

Восканян К.Л., Кузнецов А.Д., Сероухова О.С. Автоматические метеорологические станции. Часть 2. Цифровая обработка данных автоматических метеорологических станций. Практикум. — СПб.: РГГМУ, 2015. — 99 с.

ISBN 978-5-86813-423-4

В учебном пособии излагаются методы анализа и статистической обработки временных рядов, получаемых с помощью автоматических метеорологических станций общего (используемых в системе Роскомгидромета) и специального (дорожные, авиационные и экологические АМС) назначения, а также лабораторные работы, которые будут способствовать получению студентами практических навыков анализа временных рядов на ПЭВМ и работе с архивами и базами метеорологических данных.

Пособие направлено на формирование у студентов знаний в объеме, необходимом для понимания основных принципов обработки и анализа временных рядов с данными метеорологических наблюдений. Предназначено для магистров и бакалавров, обучающихся по направлению «Прикладная гидрометеорология».

Voskanyan K.L., Kuznetsov A.D., Serouhova O.S. Automatic weather stations. Part 2. Digital processing of data from automatic weather stations. Tutorial. — St. Petersburg, RSHU Publishers, 2015. — 99 pp

The tutorial shows the methods of analysis and statistical processing of time series obtained by automatic weather stations for general and special purpose. The description of the laboratory work, the implementation of which will contribute to getting the students practical skills of time series analysis on a PC and work with archives and database of meteorological data.

The manual is aimed at developing students' knowledge to the extent necessary for an understanding of the basic principles of processing and analysis of time-series data of meteorological observations. It is intended for masters and bachelors enrolled in the "Applied Hydrometeorology".

ISBN 978-5-86813-423-4

ВВЕДЕНИЕ
ВРЕМЕННЫЕ РЯДЫ МЕТЕОРОЛОГИЧЕСКИХ ВЕЛИЧИН

Существуют три вида лжи:
ложь, наглая ложь и статистика.
Бенджамин Дизраэли

Временной ряд (динамический ряд) — это расположенные в хронологической последовательности измеренные или заданные одна или конечное множество случайных (например, метеорологических) величин, соответствующих определенным моментам времени. В первом случае говорят об *одномерном временном ряде*, во втором — о *многомерном временном ряде*.

Временной ряд включает в себя два обязательных элемента — отметку времени и значение показателя ряда, полученное тем или иным способом и соответствующее указанной отметке времени. Каждый временной ряд рассматривается как выборочная реализация из бесконечной генеральной совокупности. Такая генеральная совокупность генерируется некоторым стохастическим процессом, на который могут оказывать влияние множество факторов. Если время непрерывно, временной ряд называется *непрерывным*. Если время изменяется дискретно, то временной ряд *дискретен*. В дискретном временном ряде его значения могут быть взяты как через равные, так и через неравные промежутки времени. В первом случае временной ряд называется *эквидистантным*, а во втором — *неэквидистантным*. Промежутки времени, через которые берется временной ряд, называют *интервалом дискретизации* (шагом дискретизации) Δt.

В зависимости от свойств различных выборок статистического ряда временные ряды принято разделять на следующие:

– *стационарные* временные ряды;
– *нестационарные* временные ряды.

Одномерный временной ряд называется *стационарным*, если его вероятностные характеристики (статистические параметры случайной величины: среднее значение, дисперсия и т.п.) постоянны (не изменяются с течением времени). Например, можно рассматривать стационарность временного ряда по математическому ожиданию. Временной ряд называется нестационарным, если хотя бы одна из его вероятностных характеристик непостоянна. Нестационарные временные ряды для решения

задачи прогнозирования часто приводятся к стационарным при помощи разностного оператора.

В зависимости от характера описываемого процесса временные ряды можно разделить на два следующих класса:

– временные ряды с длинной памятью;
– временные ряды с короткой памятью.

Критерием разделения временных рядов на указанные классы может, в частности, служить их автокорреляционная функция. Для временных рядов с длинной памятью их автокорреляционная функция убывает медленно. К временным рядам с короткой памятью относят временные ряды, автокорреляционная функция которых быстро убывает. Временные ряды с длинной памятью генерируют многие физические процессы. Например, временные ряды приземной температуры воздуха.

Кроме того, важной характеристикой временных рядов является их длина. Для решения многих классов задач важно знать законы, которые проявляются на фоне каких-либо длительных тенденций, например, при рассмотрении климатических характеристик. Временные ряды для сравнительно короткого временного интервала важны для краткосрочного и текущего прогнозирования, когда для оценки статистических свойств приходится работать с небольшим количеством измерений. Короткими временными рядами называются ряды, имеющие 20–50 наблюдений. Проблемы исследования таких рядов связаны, прежде всего, с надежностью получаемых оценок таких статистических характеристик, как, например, функция автоковариации.

Можно выделить две основные цели статистического анализа временных рядов:

– определение природы ряда (оценка статистических параметров и выделение детерминированной и случайной составляющих);
– использование полученных оценок для целей прогнозирования.

Выявление структуры временного ряда необходимо для того, чтобы построить математическую модель того явления, которое является источником анализируемого временного ряда.

К основным этапам анализа временного ряда можно отнести:

– графическое представление временного ряда;
– обнаружение временных разрывов временного ряда;
– обнаружение и устранение выбросов временного ряда;
– сглаживание временного ряда, при котором несистематические компоненты взаимно погашают друг друга (отфильтровывание шумов);
– расчет (оценка) основных статистических характеристик временного ряда (среднего, дисперсии, автокорреляционной функции, функции распределения и т.д.);

– выделение детерминированных составляющих временного ряда (тренд, сезонность, циклические составляющие);
– исследование случайной составляющей временного ряда;
– прогнозирование развития рассматриваемого процесса на основе имеющегося временного ряда.

Обычно в поведении временного ряда выявляют две основные тенденции: тренд и периодические колебания. Для стационарного временного ряда эти характеристики не меняются во времени.

Важным моментом исследования временных рядов является определение доверительных интервалов для полученных статистических характеристик. Здесь при выборочном исследовании генеральной совокупности и формулировании статистических выводов часто возникают этические проблемы. Основная из них — как согласуются доверительные интервалы и точечные оценки выборочных статистик. Публикация точечных оценок без указания соответствующих доверительных интервалов (как правило, имеющих 95 %-ный доверительный уровень) и объема выборки, на основе которых они получены, может вызвать недоразумения. Необходимо понимать, что в любых статистических исследованиях во главу угла должны быть поставлены не точечные, а интервальные оценки. Кроме того, особое внимание следует уделять правильному выбору объемов выборки. Чтобы доказать обоснованность полученных точечных оценок, необходимо указывать объем выборки, на основе которой они получены, границы доверительного интервала и его уровень значимости.

1. ПОКАЗАТЕЛИ ПОЛОЖЕНИЯ

1.1. Среднее арифметическое значение

Среднее арифметическое значение временного ряда $X_{ср}$ характеризует центр тяжести исследуемой характеристики или точку ее равновесия при различных колебаниях. Рассчитывается по формуле

$$X_{ср} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad (1.1)$$

где N — длина временного ряда (количество значений в нем).

Пример 1.1. Пример эквидистантного временного ряда

Таблица 1.1

№№ п/п	Значение t, °C
1	12,5
2	12,7
3	22,2
4	24,5
5	25,8
6	20,0

Пример 1.2. Для значений температуры из табл. 1.1:

$$X_{ср} = 117{,}7/6 = 19{,}6\ ^\circ\text{C}.$$

1.1.1. Расчет доверительного интервала для среднего арифметического значения

В статистике существует два вида оценок: точечные и интервальные.

Точечная оценка представляет собой отдельную выборочную статистику, которая используется для оценки параметра генеральной совокупности. Например, выборочное среднее — это точечная оценка математического ожидания генеральной совокупности.

При оценке параметров генеральной совокупности следует иметь в виду, что выборочные статистики, например среднее значение, зависят

от конкретных выборок. Чтобы учесть этот факт, для получения интервальной оценки, например, математического ожидания генеральной совокупности, анализируют распределение выборочных средних. Построенный интервал характеризуется определенным доверительным уровнем, который представляет собой вероятность того, что истинный параметр генеральной совокупности оценен правильно.

Введем кратко определение доверительного интервала: *доверительный интервал* — это *интервал значений* статистической характеристики временного ряда, соответствующий доверительной области статистического критерия.

Доверительный интервал:

1) оценивает некоторый параметр числовой выборки непосредственно по данным самой выборки;
2) «накрывает» значение этого параметра с вероятностью α.

Чтобы определить доверительный интервал, необходимо привлечь теорию проверки статистических гипотез.

Предположим, что результаты измерения распределены по нормальному закону со средним значением $X_{ср}$. Для того чтобы определить доверительный интервал для выборочного среднего арифметического значения измеряемой величины при неизвестной дисперсии генеральной совокупности, но при известной ее оценке по данным имеющейся выборки δ^2 (реальный случай), в первую очередь необходимо найти значение *критерия Стьюдента* $t_{кр}$. Затем вычислить значение Δx, с помощью которого и определяется доверительный интервал:

$$\Delta x = t_{кр} \frac{\sigma}{\sqrt{N}}, \tag{1.2}$$

где σ — стандартное (среднее квадратическое) отклонение временного ряда; N — количество значений временного ряда. Собственно доверительный интервал для математического ожидания $X_{ср}$ определяется с помощью следующего выражения:

$$\left(X_{ср} - \Delta x\right) < X_{ср} < \left(X_{ср} + \Delta x\right), \tag{1.3}$$

где $X_{ср}$ — среднее значение выборки.

1.1.2. Расчет критерия Стьюдента с использованием табличного процессора Excel

Значение критерия Стьюдента можно вычислить в табличном процессоре *Excel*, используя статистическую функцию СТЬЮДРАСПОБР.

При этом если задан уровень значимости α и N — длина временного ряда, то в указанной функции необходимо задавать следующие параметры:

$$= \text{СТЬЮДРАСПОБР}\ (1 - \alpha,\ N - 1).$$

Важно учесть, что при проведении расчетов в функции СТЬЮДРАСПОБР задается не уровень значимости α, а значение $(1 - \alpha)$, и не значение N, а значение $(N - 1)$!!!

Пример 1.2. Для данных, представленных в табл. 1.1 $N = 6$, $X_{ср} = 19{,}6$, $\sigma = 5{,}79$ и $\alpha = 0{,}95$, $t_{кр}$ = СТЬЮДРАСПОБР (0,05; 6) = 2,57. Тогда $\Delta x = 2{,}57 \cdot 5{,}79/2{,}45 \approx 6{,}1$.

Следовательно, при 95 % в доверительном интервале

$$13{,}5\ ^\circ\text{C} < X_{ср} < 25{,}7\ ^\circ\text{C}.$$

Пример 1.3. При уровне значимости $\alpha = 0{,}99$ и $N = 7$

$$\text{СТЬЮДРАСПОБР}\ (0{,}01;\ 6) = 3{,}707.$$

Пример 1.4. По данным 7 измерений некоторой величины найдена средняя результатов измерений, равная 30, и выборочная дисперсия, равная 36. Найдите границы, в которых с надежностью 0,99 заключено истинное значение измеряемой величины.

Найдем $t_{кр}\ (\alpha = 0{,}99;\ N - 1 = 6) \approx 3{,}71$. Тогда доверительные границы для интервала, заключающего истинное значение измеряемой величины, можно найти как

$$30 - \frac{3{,}71 \cdot 6}{\sqrt{7}} \le X_{ср} \le 30 + \frac{3/71 \cdot 6}{\sqrt{7}} \quad \text{или} \quad 21{,}587 \le X_{ср} \le 38{,}413.$$

Смысл полученного результата: если взять 100 различных выборок, то в 99 из них математическое ожидание будет находиться в пределах данного интервала, а в 1 из них — нет.

1.1.3. Проверка гипотезы о равенстве средних значений

Часто возникает ситуация, когда необходимо сравнить характеристики двух качественно одинаковых выборок, например, среднюю температуру за один и тот же период в разные годы. Естественно, числовые значения будут различаться, но являются ли эти различия значимыми?

Сформулируем нулевую гипотезу $H_0: \bar{x}_1 = \bar{x}_2$ и альтернативную $H_1: \bar{x}_1 \neq \bar{x}_2$. Для проверки гипотезы выберем t-критерий Стьюдента и рассчитаем *эмпирическое значение t-критерия Стьюдента*:

$$t^* = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{N_1 D_1 + N_2 D_2}} \sqrt{\frac{N_1 N_2 \left(N_1 + N_2 - 2\right)}{N_1 + N_2}}, \qquad (1.4)$$

где D_1 и D_2 — дисперсии двух частей выборки соответственно; N_1 и N_2 — длины соответствующих частей ряда.

Определим *критическое значение критерия Стьюдента* $t_{кр}$ по уровню значимости α и числу степеней свободы $\nu = N_1 + N_1 - 2$, где N_1 и N_2 — длины соответствующих частей ряда. Сравним эмпирическое t^* и критическое $t_{кр}$ значение критерия.

Если эмпирическое значение больше критического (по модулю), нулевая гипотеза отвергается. Следовательно, различия в средних значениях двух выборок статистически значимы (при заданном α), т.е. средние значения не равны друг другу.

1.2. Медиана и мода

Медианой (*Me*) называется значение признака, приходящегося на середину ранжированного (упорядоченного по возрастанию) ряда. Главное свойство медианы заключается в том, что сумма абсолютных отклонений членов ряда от медианы есть величина наименьшая:

$$\sum_{i=1}^{N} \left|x_i - Me\right| = \min. \qquad (1.5)$$

Для коротких временных рядов ($N < 30$) медиану используют вместо среднего арифметического значения.

Пример 1.5. Ранжированный временной ряд (для данных из табл. 1.1)

№№ п/п	Значение t, °C
1	12,5
2	12,7
3	20
4	22,2
5	24,5
6	25,8

Так как имеется четное количество чисел в числовом ряду, то можно найти среднее значение 2-х чисел, находящихся посередине ряда: 12,5;

12,7; **20**; **22,2**; 24,5; 25,8. Тогда, чтобы найти медиану, необходимо вычислить среднее значение 20 и 22,2:

$$20 + 22{,}2 = 44{,}2 \quad \text{и} \quad Me = 44{,}2/2 = 22{,}1.$$

Это и есть медиана данного числового ряда. При этом для этого же ряда

$$X_{\text{ср}} = 117{,}7/6 = 19{,}6.$$

Модой (*Mo*) называется наиболее часто встречающаяся в ряду величина (Мода = типичность). Иногда в совокупности встречается более чем одна мода (например: 6, 2, 6, 6, 8, 9, 9, 9, 10; мода = 6 и 9). В этом случае можно сказать, что совокупность *мультимодальна*. Из структурных средних величин только мода обладает таким уникальным свойством. Как правило, *мультимодальность* указывает на то, что набор данных не подчиняется нормальному распределению.

Мода как средняя величина употребляется чаще для данных, имеющих нечисловую природу. В экономике при экспертной оценке с помощью моды определяют наиболее популярные типы продукта, что учитывается при прогнозе продаж или планировании их производства. Так, например, среди перечисленных цветов автомобилей (*белый, черный, синий металлик, белый, синий металлик, белый*) мода будет равна *белому* цвету.

Для интервального ряда мода определяется по формуле:

$$Mo = X_{Mo} + h_{Mo}\left(f_{Mo} - f_{Mo-1}\right)/\left(f_{Mo} - f_{Mo-1}\right) + \left(f_{Mo} - f_{Mo+1}\right). \qquad (1.6)$$

где X_{Mo} — левая граница модального интервала; h_{Mo} — длина модального интервала; f_{Mo-1} — частота премодального интервала; f_{Mo} — частота модального интервала; f_{Mo+1} — частота послемодального интервала.

1.3. Описательная статистика в табличном процессоре *Excel*

Значение всех приведенных ранее статистических характеристик можно вычислить с помощью имеющейся в пакете *Excel* в опции Анализ данных функции Описательная статистика. Для этого необходимо сформировать столбец со значениями x_i, указать его начало и конец в окне «Входной интервал» и указать адрес ячейки, начиная с которой пакетом *Excel*, будут записаны основные статистические характеристики заданного ряда.

2. ПОКАЗАТЕЛИ РАЗБРОСА

2.1. Дисперсия

Дисперсия D характеризует среднее рассеяние значений ряда от среднего арифметического значения. Рассчитываются по формулам соответственно:

$$D = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - x_{\text{ср}}\right)^2, \tag{2.1}$$

$$\sigma = \sqrt{D}. \tag{2.2}$$

Выборочной дисперсией называют среднее арифметическое отклонения квадратов значений временного ряда от их среднего значения для данной выборки. Выборочная дисперсия является смещенной оценкой генеральной дисперсии, т.е. математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии. Для исправления выборочной дисперсии достаточно умножить ее на дробь $N/(N-1)$ и тогда получим исправленную дисперсию.

Стандартное отклонение σ (или выборочное среднее квадратическое отклонение) — корень квадратный из значения выборочной дисперсии.

Пример 2.1. Расчет выборочной дисперсии для временного ряда значений температуры воздуха из табл. 1.1, разд. 1.

	x	$\bar{x}$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	D
	12,5		-7,1	50,4	
	12,7		-6,9	47,6	
	22,2		0,4	0,2	
	24,5	19,6	2,6	6,8	167,4/6 = **33,5**
	25,8		4,9	24,0	
	20,0		6,2	38,4	
∑	117,7		0	167,4	

2.1.1. Расчет доверительного интервала для дисперсии

По аналогии с доверительным интервалом для математического ожидания находится доверительный интервал для дисперсии временного ряда:

$$D \cdot \Delta D_1 \prec D \prec D \cdot \Delta D_2, \tag{2.3}$$

где

$$\Delta D_1 = \left(N - 1\right) / \chi_1^2, \tag{2.4}$$

$$\Delta D_2 = \left(N - 1\right) / \chi_2^2, \tag{2.5}$$

В выражениях (2.4 и 2.5) χ_1^2, χ_2^2 — значения *критерия Пирсона* по уровням значимости α и $(1 - \alpha)$ и числу степеней свободы $\nu = N - 1$, где N — длина ряда.

Значение критерия Пирсона χ_1^2 определяется по уровню значимости $\alpha = 0{,}95$ и числу степеней свободы $\nu = N - 1$.

Значение критерия Пирсона χ_2^2 определяется по уровню значимости $(1 - \alpha) = 0{,}95$ и числу степеней свободы $\nu = N - 1$.

2.1.2. Расчет критерия Пирсона с использованием табличного процессора Excel

Значение критерия Пирсона можно вычислить с помощью статистической функции ХИ2ОБР. При этом если задан уровень значимости α и N — длина временного ряда, то в указанной функции необходимо задавать следующие параметры:

$$= \text{ХИ2ОБР}\ (1 - \alpha,\ N - 1)$$

Задается не уровень значимости α, а значение $(1 - \alpha)$!!!

Пример 2.2. Определим значения критериев Пирсона χ_1^2 и χ_2^2 при уровне значимости $\alpha = 0{,}95$ и $N = 7$.

$$\chi_1^2(\alpha = 0{,}95,\ N = 7) = \text{ХИ2ОБР}\ (0{,}05;\ 6) = 12{,}59,$$

$$\chi_2^2(\alpha = 0{,}05,\ N = 7) = \text{ХИ2ОБР}\ (0{,}95;\ 6) = 1{,}635.$$

Пример 2.3. По данным $N = 8$ испытаний найдено значение оценки для среднеквадратического отклонения $\sigma = 12$. Найти с вероятностью 0,95 ширину доверительного интервала, построенного для оценки дисперсии.

Доверительный интервал для неизвестной дисперсии генеральной совокупности можно найти по формуле:

$$\frac{N \cdot \sigma^2}{\chi_1^2} \le D \le \frac{N \cdot \sigma^2}{\chi_2^2}. \tag{2.6}$$

Подставляем

$$\chi_1^2(\alpha = 0{,}95,\ N = 8) = \text{ХИ2ОБР}\ (0{,}05;\ 7) = 14{,}067,$$

$$\chi_2^2(\alpha = 0{,}05,\ N = 8) = \text{ХИ2ОБР}\ (0{,}95;\ 7) = 2{,}167,$$

и получаем:

$$\frac{8 \cdot 144}{14{,}067} \le D \le \frac{8 \cdot 144}{2{,}167}$$

или

$$81{,}894 \le D \le 531{,}611.$$

Тогда ширина доверительного интервала *для дисперсии* равна 81,894–531,611.

Пример 2.4. По выборке объема $N = 25$ найдено среднее квадратическое отклонение $\sigma = 0{,}8$. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение с надежностью 0,95.

По данным: $\alpha = 0{,}95$ и $N = 25$.

$$\chi_1^2(\alpha = 0{,}95,\ N = 25) = \text{ХИ2ОБР}\ (0{,}05;\ 24) = 36{,}415,$$

$$\chi_2^2(\alpha = 0{,}05,\ N = 25) = \text{ХИ2ОБР}\ (0{,}95;\ 24) = 13{,}848,$$

получаем *для дисперсии* (см. формулу (2.6)):

$$\frac{25 \cdot 0{,}8 \cdot 0{,}8}{36{,}415} \le D \le \frac{25 \cdot 0{,}8 \cdot 0{,}8}{13{,}848}.$$

Искомый доверительный интервал *для среднеквадратического отклонения*:

$$0{,}663 < \sigma = 0{,}8 < 1{,}075.$$

2.1.3. Проверка гипотезы о равенстве дисперсий

Также как и средние значения, можно сравнить степень изменчивости характеристики в двух выборках, т.е. их дисперсию.

Сформулируем нулевую гипотезу $H_0 : D_1 = D_2$ и альтернативную $H_1 : D_1 \neq D_2$. Для проверки гипотезы используется параметрический *F-критерий Фишера*. Рассчитаем его *эмпирическое значение*:

$$F^* = D_1/D_2 \text{ или } D_2/D_1$$
(необходимо выбрать то отношение, для которого $F^* > 1$),

где D_1 и D_2 — дисперсии двух частей выборки, соответственно.

Определим *критическое значение* критерия Фишера $F_{кр}$ по уровню значимости $\alpha = 0{,}95$ и числам степеней свободы $\nu_1 = N_1 - 1$ и $\nu_2 = N_2 - 1$, где N_1 и N_2 — длины соответствующих частей ряда. Сравним эмпирическое и критическое значение критерия.

Если эмпирическое значение больше критического: $F_{эмп} > F_{кр}$, *то при заданном уровне значимости* α *нулевая гипотеза о равенстве дисперсий отвергается*:

$$F_{эмп} > F_{кр} \rightarrow D_1 \neq D_2.$$

Следовательно, в этом случае при заданном α различия в двух дисперсиях (т.е. в степени изменчивости) двух выборок статистически значимы (дисперсии считаются различными).

Критическое значение критерия Фишера можно определять по специальной таблице, исходя из уровня значимости α и степеней свободы числителя $(N_1 - 1)$ и знаменателя $(N_2 - 1)$. Кроме того, критическое значение критерия Фишера $F_{кр}$ можно рассчитать в табличном процессоре *Excel* с помощью функции FРАСПОБР.

2.1.4. Расчет критерия Фишера в табличном процессоре Excel

Значение критерия Фишера можно вычислить с помощью статистической функции FРАСПОБР. При этом если задан уровень значимости α и N_1, N_2 — длины двух временных рядов, то в указанной функции необходимо задавать следующие параметры:

$$= \text{FРАСПОБР}\,(\alpha,\ N_1 - 1,\ N_2 - 1).$$

При этом N_1 *должно соответствовать большей дисперсии, а* N_2 *— меньшей*! При $\alpha = 0{,}05$, $N_1 = 32$, $N_2 = 33$

$$\text{FРАСПОБР}_1\,(0{,}05;\ 31;\ 32) = 1{,}810, \quad \text{FРАСПОБР}_2\,(0{,}05;\ 31;\ 33) = 1{,}816.$$

Пример 2.5. Проиллюстрируем применение критерия Фишера на следующем примере. Дисперсия первого показателя составила 6,17 ($N_1 = 32$), а для второго 4,41 ($N_2 = 33$). Определим, можно ли считать уровень дисперсий примерно одинаковым для данных выборок на уровне значимости 0,05.

Определим *эмпирическое* значение критерия Фишера: $F_{эмп} = 6{,}17/4{,}41 \approx 1{,}4$. При этом критическое значение критерия Фишера $F_{кр}(0{,}05; 31; 32) \approx 2$.

Таким образом, $F_{эмп} = 1{,}4 < 2 = F_{кр}$, поэтому нулевая гипотеза о равенстве генеральных дисперсий на уровне значимости 0,05 принимается — дисперсии не различимы.

Пример 2.6. Для двух выборок, содержащих одинаковое число значений $N = 10$, получены значения средних величин, которые достоверно не различались друг от друга: $X_{ср_1} = 60{,}6$, $X_{ср_2} = 63{,}6$ (величина *t*-критерия Стьюдента оказалась равной 0,347). Однако есть ли при этом различия в степени однородности этих выборок? Для этого необходимо сравнить дисперсии тестовых оценок в обоих классах по критерию Фишера.

Рассчитав дисперсии для переменных *X* и *Y*, получаем:

$$D_1 = 527{,}83, \quad D_2 = 160{,}43.$$

Тогда, по формуле для расчета по *F*-критерию Фишера, находим:

$$F_{эмп} = 527{,}8/160{,}43 = 3{,}29.$$

Для *F*-критерия Фишера при степенях свободы в обоих случаях равных $10 - 1 = 9$ находим $F_{кр}$:

– для $P \leqslant 0{,}05$ $F_{кр} = 3{,}18$,
– для $P \leqslant 0{,}01$ $F_{кр} = 5{,}35$.

Таким образом, полученная величина $F_{эмп}$ попала в зону неопределенности (см. рис. 2.1):

$$3{,}18 < 3{,}29 < 5{,}35.$$

В терминах статистических гипотез можно утверждать, что гипотеза о сходстве дисперсий на уровне 5 % может быть отвергнута, а на уровне 1 % гипотеза о сходстве дисперсий принимается. Вместе с тем, можно утверждать, что по степени однородности имеются различия между этими выборками.

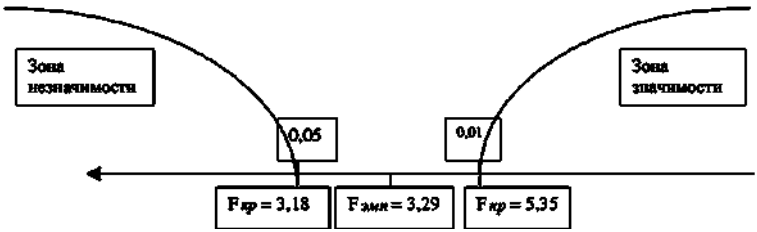


Рис. 2.1. Графическая иллюстрация применения критерия Фишера

2.2. Размах и коэффициент вариации

Размах вариации R характеризует максимальный разброс значений ряда:

$$R = \max - \min. \tag{2.7}$$

Изменчивость характеристики исследуемого временного ряда характеризует *коэффициент вариации C*, который можно рассчитать по формуле:

$$C = \frac{\sigma}{\overline{x}} 100\%. \tag{2.8}$$

Значение коэффициента вариации показывает, является ли данный временной ряд однородным ($C \prec 33\ \%$) или же изменчивость велика ($C \succ 33\ \%$). Коэффициент вариации позволяет оценить, велика или нет изменчивость ряда: если он составляет более 33 % — изменчивость значительна, а если меньше — изменчивость мала и выборка может считаться однородной. Коэффициент вариации также используется для сравнения изменчивости *двух выборок с разными единицами измерения* (например, температуры воздуха и атмосферного давления).

Пример 2.7. Для временного ряда, представленного в примере 1.1, раздел 1, размах вариации будет вычисляться как:

$$R = 25{,}8 - 12{,}5 = 13{,}3,$$

а коэффициент вариации будет равен:

$$C = \frac{\sqrt{33{,}5}}{19{,}6} 100\% = 29{,}5\%.$$

3. ПОКАЗАТЕЛИ, ОПИСЫВАЮЩИЕ ЗАКОН РАСПРЕДЕЛЕНИЯ

3.1. Функция распределения

Одной из основных идей в статистике является понятие функции плотности распределения (плотности вероятности). Функция распределения показывает соотношение между возможными значениями случайной величины и вероятностями их появления.

Функция распределения может быть симметричной или асимметричной. Чтобы описать форму распределения, необходимо вычислить его среднее значение и медиану. Если эти два показателя совпадают, переменная считается симметрично распределенной. Если среднее значение переменной больше медианы, ее распределение имеет положительную асимметрию (рис. 3.1). Если медиана больше среднего значения, распределение переменной имеет отрицательную асимметрию. Положительная асимметрия возникает, когда среднее значение увеличивается до необычайно высоких значений. Отрицательная асимметрия возникает, когда среднее значение уменьшается до необычайно малых значений. Переменная является симметрично распределенной, если она не принимает никаких экстремальных значений ни в одном из направлений, так что большие и малые значения переменной уравновешивают друг друга.

Рис. 3.1. Три вида распределений

Данные, изображенные на шкале А, имеют отрицательную асимметрию. На этом рисунке виден длинный «хвост» и перекос влево, вызванные наличием необычно малых значений. Эти крайне малые величины смещают среднее значение влево, и оно становится меньше медианы. Данные, изображенные на шкале Б, распределены симметрично. Левая и правая половины распределения являются своими зеркальными отражениями. Большие и малые величины уравновешивают друг друга, а среднее значение и медиана равны между собой. Данные, изображенные

на шкале В, имеют положительную асимметрию. На этом рисунке виден длинный «хвост» и перекос вправо, вызванные наличием необычайно высоких значений. Эти слишком большие величины смещают среднее значение вправо и оно становится больше медианы.

Функция распределения, полученная опытным путем, называется *Эмпирической (ЭФР)*. Она рассчитывается по конкретной выборке из генеральной совокупности. В гидрометеорологии в качестве ЭФР фигурирует функция, называемая *повторяемостью*.

Интегральная ЭФР характеризует вероятность появления величины, меньше заданной. Она рассчитывается путем последовательного суммирования частот (вероятности) на всех интервалах. К характеристикам ЭФР относят также *асимметрию* и *эксцесс*.

3.2. Гистограмма

Для изображения эмпирической функции распределения (ЭФР) широко используется гистограмма. *Гистограмма* — это *столбчатая диаграмма*, где по оси абсцисс откладываются значения интервалов, а частоты представлены прямоугольниками, построенными на соответствующих интервалах и имеющими высоту, пропорциональную частоте. Для расчета гистограммы предварительно формируется несколько интервалов (карманов) изменчивости переменной в исследуемом временном ряду, а затем рассчитывается количество значений переменной, попадающих в каждый интервал (частота). По гистограмме определяются *моды* как локальные максимумы ЭФР («вершины холмов»). Мод бывает одна или несколько. Соответственно, распределение является *одномодальным* или *многомодальным*. Моды характеризуют некоторые квазистационарные, т.е. наиболее устойчивые состояния. Эмпирическая функция распределения может иметь «хвосты», т.е. некоторое небольшое число наблюдений значительно больше (положительный «хвост») или меньше (отрицательный «хвост») среднего значения.

3.2.1. Определение числа интервалов при построении гистограммы

При построении гистограммы необходимо определить число интервалов k (карманов или групп), на которые будет разбита выборочная совокупность. Определение числа интервалов связано с объемом выборки, однако зависит не только от объема выборки, но и от вида закона распределения выборочной совокупности.

Для определения числа интервалов k или их длины h часто применяют следующие формулы.

Формула *Стерджесса*:

$$k = 1 + \log_2 N = 1 + 3{,}322 \ \lg N, \tag{3.1}$$

где N — объем выборки. Результат округляют до ближайшего целого числа.

Формула *Скотта:*

$$h = 3{,}5\sigma N^{-1/3}, \tag{3.2}$$

где h — длина интервала; σ — стандартное отклонение.

Формула *Брукса* и *Каррузера:*

$$k = 5 \ \lg N. \tag{3.3}$$

Еще несколько рекомендуемых соотношений:

$$k = \sqrt{N}, \quad k = 4 \lg N, \quad k = 5 \lg - 5. \tag{3.4}$$

В табл. 3.1 представлено сопоставление результатов, которые дает использование различных подходов к определению оптимального числа интервалов гистограммы в зависимости от объем выборки.

Таблица 3.1

Число интервалов гистограммы *k*

N	Оптимальное число интервалов k		
	$1 + \log_2 N$	$\sqrt[3]{\frac{2N}{3}}$	$\frac{N}{6k^3}\left[\ln(2k-1) + \frac{2k}{2k-1}\right] = 1$
500	10	7	7
1000	11	9	9
5000	13	15	15
10000	14	20	19
50000	17	35	32
100000	18	45	41
200000	19	58	51
500000	20	80	70
1000000	21	102	87

При больших N формула Скотта и информационный критерий рекомендуют значительно большее число интервалов по сравнению с формулой Стерджесса.

В табл. 3.2 представлены рекомендации ВНИИ метрологии по выбору оптимального числа интервалов в зависимости от объема выборки.

Таблица 3.2

Рекомендация ВНИИ метрологии

N	*k*
40–100	7–9
100–500	8–12
500–1000	10–16
1000–10000	12–22

При больших объемах выборок *N* разброс значений достаточно велик, поэтому на практике при выборе числа интервалов больше руководствуются разумными соображениями, выбирая их так, чтобы в интервалы попадали наблюдения числом не менее 5–10.

Задание *k* позволяет определить ширину каждого интервала гистограммы:

$$h = \left(X_{\max} - X_{\min}\right)/k, \tag{3.5}$$

где $X_{\max}$ и $X_{\min}$ — максимальное и минимальное значения в выборке. Результат округляют до ближайшего целого числа.

3.2.2. Построение гистограммы в табличном процессоре Excel 2003

Чтобы создать гистограмму в *Excel*, нужно воспользоваться пакетом анализа. Рассмотрим последовательность необходимых для этого действий. Сразу же подчеркнем, что приведенная последовательность действий не является единственно возможной.

1. На листе в один столбец вводятся исходные данные (если они отсутствовали на имеющемся листе).
2. С помощью опций Сервис – Анализ данных – Описательная статистика определяются максимальное и минимальное значения в используемой выборке: $X_{\max}$ и $X_{\min}$.
3. Одним из указанных выше способов по длине выборки *N* определяется число интервалов (карманов) в гистограмме *k*.
4. Вычисляется ширина каждого интервала гистограммы (ширина кармана):

$$h = \left(X_{\max} - X_{\min}\right)/k,$$

Вычисленное таким образом значение *h* округляется до удобной для расчета границ интервалов величины.

5. Далее в отдельную колонку, содержащую k строк, записываются границы интервала (интервал карманов). Например: 5, 10, 15, 20. За начало первого интервала A_0 принимается значение

$$A_0 = X_{\min} - h/2. \tag{3.6}$$

За конец j-го интервала принимается значение A_j, представляющее собой верхнюю границу j-го интервала и начало $(j + 1)$-го интервала:

$$A_j = A_{(j-1)} + h. \tag{3.7}$$

Построение шкалы интервалов продолжается до тех пор, пока величина A_j удовлетворяет соотношению

$$A_j < X_{\max} + h/2. \tag{3.8}$$

6. В отдельный столбец вводятся интервалы A_j в возрастающем порядке.
7. С помощью опций Сервис – Анализ данных – Гистограмма открывается рабочее окно, представленное на рис. 3.2 (вид окна уже после заполнения).

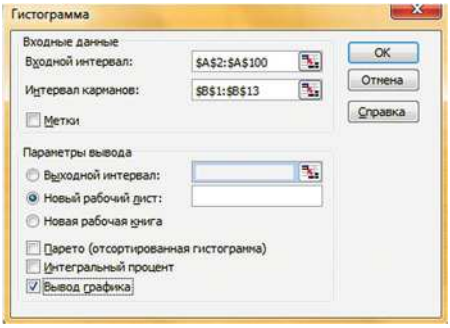


Рис. 3.2. Окно ввода информации для построения гистограммы

8. Заполняются строки: Входные данные, Входной интервал и Интервал карманов. Входной интервал: указываются начало и конец столбца с данных на имеющемся листе. Интервал карманов: указываются начало и конец данных столбца со значениями интервалов A_j, т.е. вводится ссылка на диапазон ячеек, который содержит значения интервалов.
9. В группе «Параметры вывода» выбирается местоположение выходных данных.
10. Нажимается кнопка OK. В указанном выходном поле появится график гистограммы.

Пример 3.1. Графическое представление гистограмм

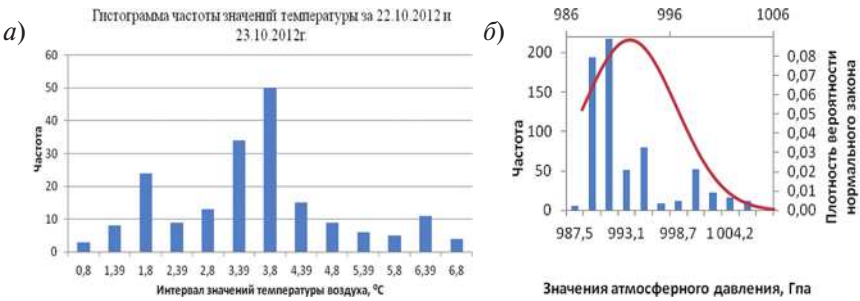


Рис. 3.3. Гистограммы распределения значений температуры воздуха (*а*) и значений атмосферного давления (с аппроксимацией гистограммы значений атмосферного давления нормальной функцией распределения) (*б*)

3.3. Асимметрия и эксцесс

Асимметрия (*As*) характеризует симметричность эмпирической функции распределения (ЭФР) относительно среднего значения и рассчитывается по формуле:

$$As = \frac{1}{N\sigma^3}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^3, \tag{3.9}$$

или

$$As = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\left(\frac{x_i - x_{\text{ср}}}{\sigma}\right)^3. \tag{3.10}$$

Значение $As = 0$ — при полной симметрии ЭФР относительно среднего значения. Если $As > 0$, то ЭФР обладает положительным «хвостом» и основная масса наблюдений (а также медиана) меньше среднего значения (это означает, что преобладают данные с меньшими значениями по сравнению со средним арифметическим). Если $As < 0$, то ЭФР обладает отрицательным «хвостом» и основная масса наблюдений (а также медиана) больше среднего значения (это означает, что преобладают данные с большими значениями).

Кроме того, большие (больше 1) значения асимметрии (по модулю) свидетельствуют о наличии в ряду «выбросов», которые могут быть ошибками наблюдения или наблюдаемыми катастрофическими эффектами.

Пример 3.2. Схема расчета асимметрии для временного ряда температуры воздуха.

	X	$\bar{x}$	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^3$	D	σ	σ^3	As
	12,5		−7,1	50,4	−357,9	33,5	5,8	193,7	−312,8/(193,7·5,8) = −0,3
	12,7		−6,9	47,6	−328,5				
	20		0,4	0,2	0,1				
	22,2	19,6	2,6	6,8	17,6				
	24,5		4,9	24,0	117,6				
	25,8		6,2	38,4	238,3				
Σ	117,7		0	167,4	−312,8				

Эксцесс (*Ex*) характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением и рассчитывается по формуле:

$$Ex = \left[\frac{1}{N\sigma^4} \sum_{i=1}^{N} \left(x_i - \bar{x} \right)^4 \right] - 3, \tag{3.11}$$

или

$$Ex = \left[\frac{N\left(N+1\right)}{\left(N-1\right)\left(N-2\right)\left(N-3\right)} \sum_{i=1}^{N} \left(\frac{x_i - x_{\text{ср}}}{\sigma} \right)^4 \right] - \frac{3\left(N-1\right)^2}{\left(N-2\right)\left(N-3\right)}. \tag{3.12}$$

Если $Ex > 0$, то ЭФР является островершинной и, как правило, у нее наблюдается два равнозначных «хвоста». Если $Ex < 0$, то ЭФР является плосковершинной и распределение стремится к случайному распределению.

Так же, как и для асимметрии, большие (больше 1) значения эксцесса свидетельствуют о наличии в ряду «выбросов», которые могут быть ошибками наблюдения или наблюдаемыми катастрофическими эффектами.

Пример 3.3. Схема расчета эксцесса для временного ряда температуры воздуха.

	X	$\bar{x}$	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^4$	D	σ	σ^4	Ex
	12,5		−7,1	50,4	2541,2	33,5	5,8	1120,8	[6907,7/(1120,8·6)]−3 = −1,8
	12,7		−6,9	47,6	2266,7				
	20		0,4	0,2	0,03				
	22,2	19,6	2,6	6,8	45,7				
	24,5		4,9	24,0	576,5				
	25,8		6,2	38,4	1477,6				
Σ	117,7		0	167,4	6907,7				

Эмпирическое правило. В большинстве ситуаций крупная доля наблюдений концентрируется вокруг медианы, образуя кластер. В наборах данных, имеющих положительную асимметрию, этот кластер расположен левее (т.е. ниже) математического ожидания, а в наборах, имеющих отрицательную асимметрию, этот кластер расположен правее (т.е. выше)

математического ожидания. У симметричных данных математическое ожидание и медиана совпадают, а наблюдения концентрируются вокруг математического ожидания, формируя колоколообразное распределение.

Если распределение не имеет ярко выраженной асимметрии, а данные концентрируются вокруг некоего центра тяжести, для оценки изменчивости можно применять эмпирическое правило, которое гласит: если данные имеют колоколообразное распределение, то приблизительно 68 % наблюдений отстоят от математического ожидания не более чем на одно стандартное отклонение. Приблизительно 95 % наблюдений отстоят от математического ожидания не более чем на два стандартных отклонения и 99,7 % наблюдений отстоят от математического ожидания не более чем на три стандартных отклонения.

Таким образом, стандартное отклонение, представляющее собой оценку среднего колебания вокруг математического ожидания, помогает понять, как распределены наблюдения, и идентифицировать выбросы. Из эмпирического правила следует, что для колоколообразных распределений лишь одно значение из двадцати отличается от математического ожидания больше, чем на два стандартных отклонения. Следовательно, значения, лежащие за пределами интервала $\mu \pm 2\sigma$, можно считать выбросами. Кроме того, только три из 1000 наблюдений отличаются от математического ожидания больше чем на три стандартных отклонения. Таким образом, значения, лежащие за пределами интервала $\mu \pm 3\sigma$ практически всегда являются выбросами.

Для распределений, имеющих сильную асимметрию или не имеющих колоколообразной формы, можно применять эмпирическое правило Бьенамэ – Чебышева. Более ста лет назад математики Бьенамэ и Чебышев независимо друг от друга открыли полезное свойство стандартного отклонения. Они обнаружили, что для любого набора данных, независимо от формы распределения, процент наблюдений, лежащих на расстоянии, не превышающем k стандартных отклонений от математического ожидания, не меньше $(1 - 1/k^2) \cdot 100\ \%$.

Например, если $k = 2$, то правило Бьенамэ – Чебышева гласит, что как минимум $(1 - (1/2)^2) \cdot 100\ \% = 75\ \%$ наблюдений должно лежать в интервале $\mu \pm 2\sigma$. Это правило справедливо для любого k, превышающего единицу.

Правило Бьенамэ – Чебышева носит весьма общий характер и справедливо для распределений любого вида. Оно указывает минимальное количество наблюдений, расстояние от которых до математического ожидания не превышает заданной величины. Однако, если распределение имеет колоколообразную форму, эмпирическое правило более точно оценивает концентрацию данных вокруг математического ожидания.

4. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

4.1. Автокорреляционная функция

Прогнозируемость временного ряда возможна лишь тогда, когда существует вероятностная (или аналитическая) связь последующих значений ряда от предыдущих. И в том, и в другом случаях для прогнозирования временного ряда необходимо построить его математическую модель. В том случае, когда предполагается использование вероятностной (статистической) связи, прогнозируемость (стационарного) временного ряда определяется с помощью автокорреляционной функции (АКФ).

Автокорреляционная функция $\rho(m)$ для временного ряда $X(k)$, имеющего среднее значение μ и дисперсию σ^2 — это величины коэффициентов корреляции последующих значений временного ряда с предыдущими значениями при сдвиге последних на величину m:

$$\rho(m) = M\left[\left(X(k)-\mu\right)\left(X(k+m)-\mu\right)\right]/\sigma^2. \qquad (4.1)$$

Очевидно, что $\rho(0) \equiv 1$, поскольку это корреляция временного ряда на самого себя.

Стационарный временной ряд прогнозируем, если для $m > 0$ существует $\rho(m) \neq 0$. Стационарный временной ряд не прогнозируем, если для любого $m > 0$ величина $\rho(m) = 0$. Такой ряд называют «*белым шумом*».

Пример 4.1. Графически зависимость функции $\rho(m)$ от m может иметь вид — см. рис. 4.1.

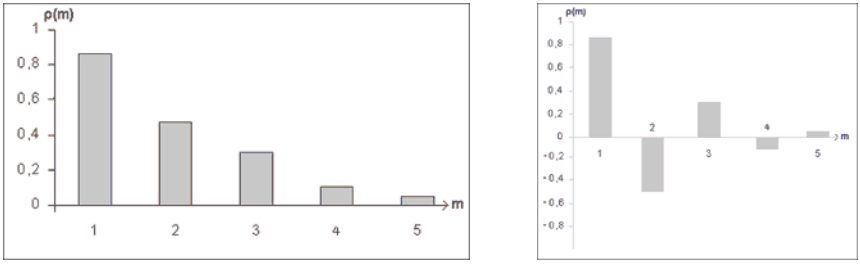


Рис. 4.1. Примеры графического представления функции $\rho(m)$: $m = 1, 2, ..., 5$

Оценивание автокорреляционной функции осуществляется по реализации (фрагменту) временного ряда. Если реализация содержит N значений, то выражение для оценки автокорреляционной функции имеет вид:

$$r_m = \sum_{i=1}^{N-m} \frac{\left[\left(X_i - X_{\text{ср}}\right)\left(X_{i+m} - X_{\text{ср}}\right)\right]}{\left(N-1\right)\sigma^2}, \qquad (4.2)$$

где r_m — оценка автокорреляционной функции; $X_{\text{ср}}$ — среднее значение для данной реализации временного ряда, состоящего из N членов; σ^2 — оценка дисперсии для данной реализации временного ряда.

Необходимо отметить, что оценки автокорреляционной функции имеют смысл при выполнении соотношения: $m < 0{,}1N$. Следовательно, при проверке прогнозируемости временного ряда длина реализации должна быть не менее 20–30 наблюдений.

После оценивания автокорреляционной функции необходимо для каждого m проверить гипотезу о равенстве нулю соответствующего коэффициента корреляции! Для этого в программах статистической обработки данных для каждой из оценок коэффициентов r_m вычисляются критические значения, которые на графике оценки автокорреляционной функции приобретают вид контрольных границ.

Пример 4.2. Результаты оценки значимости значений автокорреляционной функции с использованием критерия Стьюдента при уровне значимости $\alpha = 0{,}95$ (рис. 4.2).

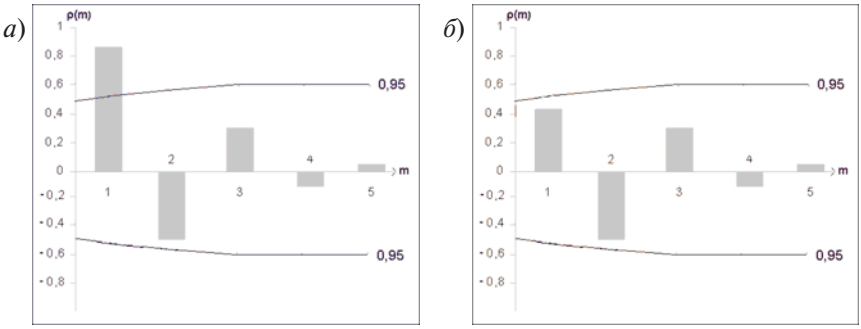


Рис. 4.2. Примеры графического представления результатов оценивания функции $\rho(m)$: m = 1, 2, ..., 5

Согласно приведенному графику на рис. 4.2*а*, значимой величиной является только r_1. На основе такой оценки можно сделать вывод о том, что в принципе данный ряд прогнозируем. Рис. 4.2*б* иллюстрирует случай

непрогнозируемого по имеющейся реализации ряда, так как все значения r_m для любого $m > 0$ незначимы.

4.2. Коэффициент корреляции Пирсона

4.2.1. Линейный коэффициент корреляции Пирсона

При наличии двух и более синхронных временных рядов между ними возможен специфический вид связи, обусловленный влиянием различных факторов. Эту связь называют стохастической, то есть вероятностной. Стохастическая связь между значениями y, полученными при разных уровнях фактора x, называется корреляционной. Мерой силы, или тесноты такой связи между переменными служит коэффициент корреляции r_{xy}:

$$r_{xy} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - x_{\text{ср}})(y_i - y_{\text{ср}})}{\sigma(x)\sigma(y)}, \tag{4.3}$$

где $x_{\text{ср}}$ и $y_{\text{ср}}$ являются оценками среднего, а $\sigma(x)$ и $\sigma(y)$ — среднеквадратические отклонения величин x и y.

Однако в целях оптимизации вычислений для расчета линейного коэффициента корреляции Пирсона используют получаемый с помощью преобразований аналог этой формулы, не требующий предварительного расчета среднего и среднеквадратического отклонения:

$$r_{xy} = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right]}}. \tag{4.4}$$

Значения коэффициента корреляции r_{xy} могут изменяться в пределах от -1 до $+1$. Чем ближе r_{xy} по абсолютной величине к единице, тем теснее связь. Переменные будут связаны линейной зависимостью при крайних значениях: $r_{xy} = -1$, $r_{xy} = +1$. Если значение $r_{xy} = 0$, то связь между переменными отсутствует. Когда $0 < |r_{xy}| < 1$, связь между переменными является стохастической, т.е. включает в себя функциональную составляющую, которая в большей или меньшей мере замаскирована влиянием случайных факторов (рис. 4.3).

Знак коэффициента корреляции показывает направление зависимости: «+» — прямая зависимость, «−» — обратная зависимость.

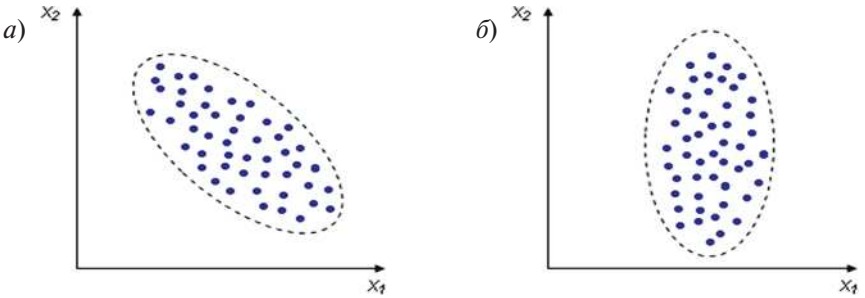


Рис. 4.3. Отрицательная корреляция (*а*) и отсутствие корреляции (*б*)

Степень корреляционной связи может быть оценена по *шкале Чеддока*:

$0{,}1 < |r_{xy}| < 0{,}3$ — слабая;

$0{,}1 < |r_{xy}| < 0{,}5$ — умеренная;

$0{,}5 < |r_{xy}| < 0{,}7$ — заметная;

$0{,}7 < |r_{xy}| < 0{,}9$ — высокая;

$0{,}9 < |r_{xy}| < 1$ — весьма высокая.

Методами корреляционного анализа решаются следующие задачи:

1. Взаимосвязь. Есть ли взаимосвязь между параметрами.
2. Прогнозирование. Если известно поведение одного параметра, то можно предсказать поведение другого параметра, коррелирующего с первым.
3. Классификация и идентификация объектов. Корреляционный анализ помогает подобрать набор независимых признаков для классификации.

Слабыми сторонами линейного коэффициента Пирсона являются:

1. Неустойчивость к выбросам.
2. С помощью коэффициента корреляции Пирсона можно определить степень линейной зависимости между величинами, другие виды взаимосвязей выявляются методами регрессионного анализа.
3. Необходимо понимать различие понятий «независимость» и «некоррелированность». Из первого следует второе, но не наоборот.

Расчет коэффициента корреляции Пирсона предполагает, что переменные x и y распределены нормально.

Пример 4.3. «Слабые стороны» линейного коэффициента корреляции Пирсона иллюстрирует рис. 4.4. На рисунке представлены 4 выборки двух параметров с $N = 11$, имеющие один и тот же коэффициент корреляции $r_{xy} = 0{,}81$.

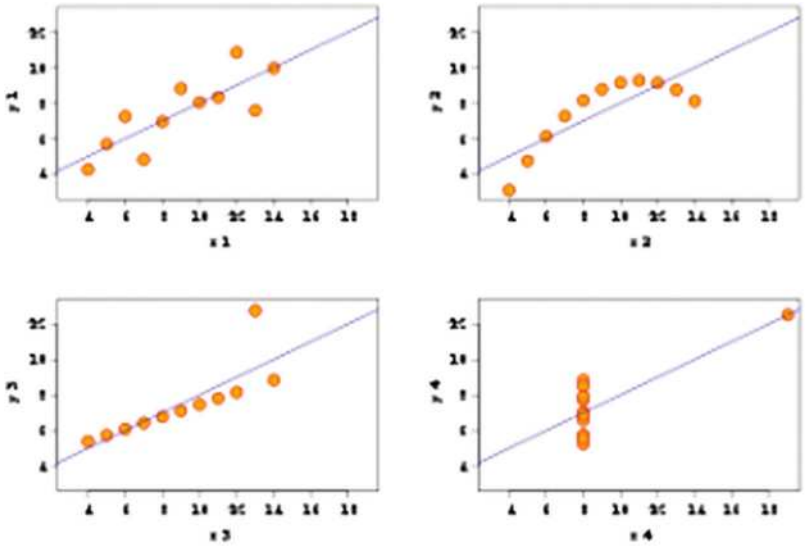


Рис. 4.4. Четыре выборки объемом 11, имеющие один и тот же коэффициент корреляции 0,81

Пример 4.4. Исследовалась зависимость рождаемости в небольших городах от количества аистов в них. Коэффициент корреляции Пирсона оказался близок к 0,8. Наличие стохастической связи в этом случае не обусловлено существованием причинной связи.

4.2.2. Частный коэффициент корреляции

Исключить влияние третьей переменной позволяет частный коэффициент корреляции. Он позволяет оценить корреляционную связь между двумя переменными, исключив влияние других переменных.

Для случая трех переменных частный коэффициент корреляции между случайными величинами X и Y при исключении влияния случайной величины Z $r_{xy/z}$ определяется по формуле:

$$r_{xy/z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)}}. \qquad (4.5)$$

4.2.3. Корреляционный анализ в табличном процессоре Excel

Чтобы рассчитать коэффициент корреляции в *Excel*, нужно воспользоваться функцией *Excel* КОРРЕЛ (массив 1; массив 2). Здесь массив 1 — это ячейка интервала значений для первой переменной x_i, а массив 2 — это второй интервал ячеек со значениями y_i. В результате получаем коэффициент корреляции между временными рядами, находящимися в ячейках массив 1 и массив 2.

4.3. Оценка значимости коэффициента корреляции Пирсона

Коэффициент корреляции r является случайной величиной, поскольку вычисляется из случайных величин. Следовательно, для него можно выдвигать и проверять различные гипотезы. Пусть производится выборка объемом N из двух временных рядов: X_i и Y_i, где $i = 1, 2, ..., N$. По этим данным вычисляется выборочный коэффициент корреляции r_{xy}. Необходимо проверить, коррелированны между собой или нет параметры этих двух временных рядов, содержащихся в исследуемой выборке.

Для проверки этой гипотезы необходимо знать распределение величины r_{xy} — значения коэффициента корреляции. Однако собственное распределение величины r_{xy} довольно сложное, поэтому для этой величины применяется следующее преобразование:

$$t_{\text{в}} = \frac{r_{xy}\sqrt{N-2}}{\sqrt{1-r_{xy}^2}}. \tag{4.6}$$

После расчета величины $t_{\text{в}}$ для оценки значимости коэффициента корреляции используют критерий Стьюдента. При заданном уровне значимости α находим величину $t_{\text{кр}}(N-2;\alpha)$.

Если модуль выборочного значения критерия $t_{\text{в}}$ превосходит $t_{\text{кр}}$, то гипотеза о равенстве коэффициента корреляции нулю отвергается и выборочный коэффициент корреляции считается статистически значимым. В противном случае, т.е. если $|t_{\text{в}}| < t_{\text{кр}}$, гипотеза принимается — выборочный коэффициент корреляции считается статистически незначимым:

– $|t_{\text{в}}| < t_{\text{кр}}$: корреляция *незначима* при заданном уровне значимости α;
– $|t_{\text{в}}| > t_{\text{кр}}$: корреляция *значима* при заданном уровне значимости α.

Интервальная оценка для коэффициента корреляции (доверительный интервал):

$$r_{xy} \pm t_{\text{кр}}\sqrt{\frac{1-r_{xy}^2}{N-2}}. \tag{4.7}$$

Пример 4.5. Для $N = 7$ и $\alpha = 0{,}05$: $t_{кр}(5; 0{,}05) = 2{,}571$. Если $r_{xy} = 0{,}971$, то доверительный интервал для коэффициента корреляции будет определяться следующим соотношением:

$$0{,}971 \pm 2{,}571\left[\left(1-0{,}971^2\right)/\left(7-2\right)\right]^{1/2},$$

т.е. при заданном уровне значимости доверительный интервал для коэффициента корреляции лежит в интервале от 0,695 до 1.

Для проверки гипотезы о значимом отличии коэффициента корреляции от нуля (т.е. существует взаимосвязь между величинами) тестовая статистика может быть вычислена по следующей формуле:

$$t_в = \left[0{,}5\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) - \frac{abs\left(r_{xy}\right)}{2\left(N-1\right)}\right]\sqrt{N-3}. \qquad (4.8)$$

Полученное таким образом значение $t_в$ сравнивается с табличным значением коэффициента Стьюдента $t_{кр}(\alpha, N = \infty)$. Для $\alpha = 0{,}95$ $t_{кр} = 1{,}96$.

Если тестовая статистика $t_в$ больше табличного значения $t_{кр}$, то коэффициент значимо отличается от нуля. Из анализа формулы (4.8) видно, что чем больше измерений N, тем лучше (больше тестовая статистика, вероятнее, что коэффициент значимо отличается от нуля).

Для проверки гипотезы о значимом отличии двух коэффициентов корреляции, полученных для двух выборок из одной и той же генеральной совокупности, тестовая статистика вычисляется по формуле:

$$t_в = \left[0{,}5\ln\left(\frac{\left(1+r_1\right)\left(1-r_2\right)}{\left(1-r_1\right)\left(1+r_2\right)}\right)\right]\frac{1}{\sqrt{\dfrac{1}{N_1-3}+\dfrac{1}{N_2-3}}}. \qquad (4.9)$$

Полученное таким образом значение $t_в$ сравнивается с табличным значением коэффициента Стьюдента $t_{кр}(\alpha, N = \infty)$. Если тестовая статистика $t_в$ больше табличного значения $t_{кр}$, то два коэффициента значимо отличаются друг от друга.

Пример 4.6. Оценка значимости коэффициентов корреляции приземных значений температуры и относительной влажности (r_{tf}), и температуры и атмосферного давления (r_{tp}).

	r_{xy}	N	σ_r	$t_в$	$t_{кр}$
r_{tf}	−0,88	85	0,025	35,2	1,98
r_{tp}	0,05	1430	0,026	1,92	1,96

Анализируя полученные результаты, надо отметить, что в первом случае межрядовая корреляция r_{tf} значима (носит неслучайный характер),

поскольку $|t_в| > t_{кр}$. Во втором случае для межрядовой корреляции $r_{tp}: |t_в| > t_{кр}$. Это свидетельствует о том, что отклонение от нуля значения коэффициента корреляции r_{tp} случайно и, следовательно, незначимо.

4.4. Множественный коэффициент корреляции

4.4.1. Коэффициент частной корреляции

С помощью коэффициента корреляции выявляется связь между двумя признаками, один из которых можно рассматривать как результативный, а другой соответственно — как факторный. Но в действительности, если одна величина коррелированна с другой, то это может являться отражением того, что они обе коррелированы с третьей величиной или с совокупностью величин. Поэтому, чтобы оценить истинную корреляцию между переменными, нужно «очистить» их от влияния третьих переменных. Для решения этой проблемы вводится понятие частной корреляции. Корреляция между двумя переменными, вычисленная после устранения влияния всех других переменных, называется *частной корреляцией*.

Частный коэффициент корреляции первого и второго признаков при исключении влияния третьего оценивает тесноту корреляционной связи между первым и вторым признаками при фиксированном значении третьего признака. Оценку влияния на первый признак изменений третьего признака при постоянных значениях второго признака можно рассчитать следующим образом:

$$R_{13,2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{32}^2\right)}}. \qquad (4.10)$$

В более общем случае для оценки влияния на переменную y нескольких переменных x_i, $i = 1, \ldots, m$, используют матрицу коэффициентов парной корреляции:

$$R = \begin{bmatrix} 1 & r_{yx_1} & \ldots & r_{yx_m} \\ r_{yx_1} & 1 & \ldots & r_{x_2x_m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{yx_m} & r_{x_2x_m} & \ldots & 1 \end{bmatrix}. \qquad (4.11)$$

4.4.2. Расчет матрицы коэффициентов парной корреляции в табличном процессоре Excel

Для расчета матрицы коэффициентов парной корреляции R нужно обратиться к средствам анализа данных: меню Сервис – Анализ

данных – Корреляция. В появившемся окне необходимо указать координаты левой верхней ячейки и правой нижней ячейки, что позволит указать все столбцы и строки, содержащие исследуемые временные ряды. На рис. 4.5 входной интервал показывает, что для расчета матрицы будут использоваться данные, расположенные в колонках C, D, E, A, G и в строках 1–500. Элементы корреляционной матрицы будут представлены на новом рабочем листе.

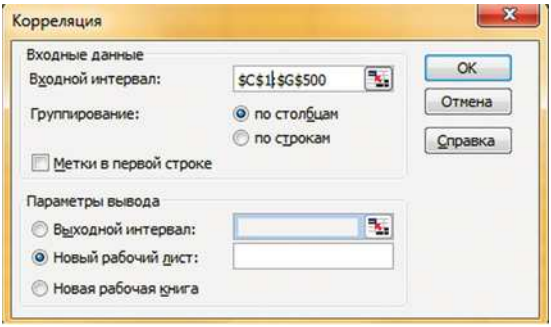


Рис. 4.5. Пример задания входного интервала

В качестве исследуемых временных рядов в данном примере использовались 500 значений следующих метеорологических величин (дискретность — 15 мин):

1) температуры воздуха, *T*;
2) атмосферного давления, *P*;
3) парциального давления водяного пара, *E*;
4) относительной влажности, *F*;
5) скорости ветра, *V*.

На рис. 4.6 представлен фрагмент этих данных. Здесь же указанны размерности метеорологических величин.

#	Time	T, C	P, hPa	e, hPa	RH, %	V, m/s
1	01.07.2012 00:00:00	14.6	1011.6	10.1	61.0	5.0
2	01.07.2012 00:00:10	14.6	1011.6	10.3	62.0	4.0
3	01.07.2012 00:00:20	14.6	1011.6	10.3	62.0	5.0
4	01.07.2012 00:00:30	14.6	1011.7	10.3	62.0	4.0
5	01.07.2012 00:00:40	14.6	1011.6	10.3	62.0	4.0
6	01.07.2012 00:00:50	14.6	1011.6	10.3	62.0	4.0
7	01.07.2012 00:01:00	14.6	1011.6	10.3	62.0	3.0
8	01.07.2012 00:01:10	14.6	1011.6	10.3	62.0	4.0
9	01.07.2012 00:01:20	14.6	1011.6	10.3	62.0	3.0
10	01.07.2012 00:01:30	14.6	1011.6	10.3	62.0	3.0

Рис. 4.6. Фрагмент временных рядов, использованных для расчета матрицы коэффициентов парной корреляции

В результате расчетов выдается таблица, содержащая элементы корреляционной матрицы: коэффициенты корреляции для каждой возможной пары переменных измерений в диапазоне от −1 до +1 включительно. Поскольку корреляционная матрица симметрична, то выводятся только значения, находящиеся ниже главной диагонали. Для рассматриваемого примера форма представления расчетных данных представлена на рис. 4.7.

	A	B	C	D	E	F
1		*T*	*P*	*E*	*F*	*V*
2	T	1				
3	P	-0.2562402	1			
4	E	-0.1685514	0.122111436	1		
5	F	-0.73519	0.24534263	0.781802921	1	
6	V	-0.0391343	0.205270093	-0.144026371	-0.083252696	1

Рис. 4.7. Матрица коэффициентов парной корреляции

Как уже отмечалось ранее, после расчета коэффициентов корреляции необходимо оценить их значимость. Необходимые для этого статистические величины представлены на рис. 4.8. Они были получены также средствами анализа данных: меню Сервис – Анализа данных – Описательная статистика.

1	*T*		*P*		*E*		*F*		*V*	
2										
3	Среднее	14.5296	Среднее	1011.42	Среднее	10.2692	Среднее	62.184	Среднее	2.78
4	Стандартн	0.003831	Стандартн	0.00503	Стандартн	0.003812	Стандартн	0.029632	Стандартн	0.038603
5	Медиана	14.5	Медиана	1011.5	Медиана	10.3	Медиана	62	Медиана	3
6	Мода	14.5	Мода	1011.5	Мода	10.2	Мода	62	Мода	3
7	Стандартн	0.085665	Стандартн	0.112468	Стандартн	0.08524	Стандартн	0.662587	Стандартн	0.863186
8	Дисперсия	0.007339	Дисперсия	0.012649	Дисперсия	0.007266	Дисперсия	0.439022	Дисперсия	0.74509
9	Эксцесс	-0.09685	Эксцесс	-0.70575	Эксцесс	-0.29594	Эксцесс	-0.0698	Эксцесс	0.239463
10	Асимметр	0.524428	Асимметр	-0.47597	Асимметр	0.109227	Асимметр	-0.38764	Асимметр	0.102337
11	Интервал	0.4	Интервал	0.5	Интервал	0.5	Интервал	3	Интервал	5
12	Минимум	14.4	Минимум	1011.2	Минимум	10	Минимум	60	Минимум	0
13	Максимум	14.8	Максимум	1011.7	Максимум	10.5	Максимум	63	Максимум	5
14	Сумма	7264.8	Сумма	505710.2	Сумма	5134.6	Сумма	31092	Сумма	1390
15	Счет	500	Счет	500	Счет	500	Счет	500	Счет	500
16	Наибольш	14.8	Наибольш	1011.7	Наибольш	10.5	Наибольш	63	Наибольш	5
17	Наименьш	14.4	Наименьш	1011.2	Наименьш	10	Наименьш	60	Наименьш	0
18	Уровень н	0.007527	Уровень н	0.009882	Уровень н	0.00749	Уровень н	0.058218	Уровень н	0.075844

Рис. 4.8. Статистические характеристики четырех временных рядов

5. ВРЕМЕННОЙ ТРЕНД

5.1. Выявление временного тренда

Само понятие временного тренда предполагает, что он относится к временному ряду, содержащему значения некоторого параметра в различные моменты времени t_i, расположенные в хронологическом порядке. Собственно, *временной тренд* временного ряда — это гладкая функция $y(t)$, описывающая его долгосрочное поведение, а нахождение временного тренда — это задание вида этой функции и определение ее параметров (коэффициентов) по имеющимся значениям выборки исследуемого временного ряда.

Нужно подчеркнуть, что вид функции, описывающей временной тренд, не определяется однозначно самим рядом и является некоторым условным объектом, использующимся для более полного понимания особенностей рассматриваемого процесса. Кроме того, нужно отметить, что не существует «автоматического» способа обнаружения тренда во временном ряде. Легко определить временной тренд, если члены временного ряда монотонно изменяются во времени (значения временного ряда устойчиво возрастают или устойчиво убывают). В этом случае наличие тренда часто хорошо видно на графике. Анализировать такой ряд обычно нетрудно.

Другой случай, когда в поведении временного ряда не прослеживается такая длительная монотонность. Тогда можно осуществить проверку гипотезы о существовании временного тренда с использованием некоторых простых критериев или тестов.

Тест числа поворотных точек (пиков) основан на вычислении числа локальных максимумов в рассматриваемой выборке, содержащей N значений. Каждое значение ряда (x_i) сравнивается с двумя, рядом стоящими. Точка считается поворотной, если значение ряда в этой точке:

– либо одновременно больше и предыдущего и последующего значения;
– либо одновременно меньше и предыдущего и последующего значения.

Полученное таким образом число K сравнивается с идеальным значением $K_{\text{ид}}$, соответствующим случайному временному ряду. В случайном ряду должно выполняться строгое неравенство:

$$K_{\text{ид}} > \left\{ \frac{2(N-2)}{3} - 2\sqrt{\frac{16N-29}{90}} \right\},$$

где в данном случае фигурные скобки означают взятие целой части результата вычислений.

Если K меньше $K_{ид}$, то это может свидетельствовать в пользу наличия у исследуемого ряда временного тренда. Отклонение K от идеального значения $K_{ид}$ в большую сторону свидетельствует о значительной дисперсии и заметной отрицательной автокорреляции случайной компоненты. Справедливости ради следует также указать, что отклонение K в меньшую сторону может возникнуть не только при наличии тренда, но и в случае положительной автокорреляции членов ряда.

Особо сложным случаем является выделение временного тренда в том случае, когда за рассматриваемый промежуток времени вместо монотонности многократно меняется характер протекающего процесса. На рис. 5.1 в первом случае (*а*) прослеживается монотонность и ряд явно имеет временной тренд. Во втором случае (*б*) в момент времени t^* произошло резкое изменение тенденции. Значение t^* — это *точка бифуркации*, определяющая момент времени не изменения, а резкой смены характера протекающего процесса. Из анализа поведения представленных на рис. 5.1*б* трех прямых видно, как важно при задании временного тренда учесть как сам момент t^*, так и последующее изменение временной тенденции.

Ситуация, представленная на рис. 5.1*б*, иллюстрирует поведение так называемых прерванных временных рядов, когда на последовательность наблюдений в некоторый момент времени t^* воздействует некоторое внешнее событие. При этом можно выделить следующие типы воздействий:

– устойчивое скачкообразное (рис 5.1*б*);
– скачкообразное временное.

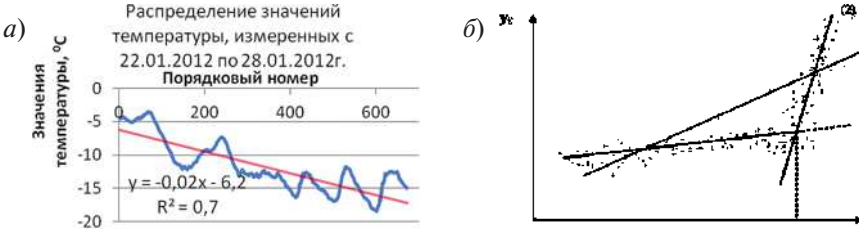


Рис. 5.1. Примеры использования линейной функции для описания временного тренда:
а — временной ряд с выраженной тенденцией к уменьшению своих значений во времени; *б* — временной ряд, имеющий точку бифуркации в момент t^*

Наиболее простым видом функции, описывающей временной тренд, является линейная функция. Уравнение линейного временного тренда может быть представлено формулой:

$$y(t) = a_1 t + a_0 + \varepsilon, \tag{5.1}$$

где t — время; a_0 и a_1 — коэффициенты тренда; ε — ошибка трендовых составляющих.

Коэффициенты a_0 и a_1 полинома (5.1) могут быть определены по имеющейся выборке $\{t_i, y_i\}$, где $i = 1, ..., N$, например, методом наименьших квадратов.

При использовании метода наименьших квадратов коэффициенты a_0 и a_1 находятся на основе решения следующей системы двух линейных алгебраических уравнений:

$$a_0 N + a_1 \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} y_i, \quad a_0 \sum_{i=1}^{N} t_i + a_1 \sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} y_i\, t_i. \tag{5.2}$$

Уравнение нелинейного (квадратичного) тренда:

$$y = a_2 t^2 + a_1 t + a_0 + \varepsilon. \tag{5.3}$$

5.2. Характеристики временного тренда

Коэффициент детерминации R^2. Термин R^2 в табличном процессоре *Excel* (для линейного тренда) характеризует вклад тренда в общую дисперсию ряда. Термин R^2 кроме коэффициента детерминизации также носит следующие названия: достоверность аппроксимации, квадрат смешанной корреляции.

Коэффициент детерминации R^2 для табличных данных: $\{t_i, y_i, i = 1, ..., N\}$ и уравнения временного тренда

$$y_{\text{сум}} = w(t) \tag{5.4}$$

определяется следующим соотношением:

$$R^2 = S_{\text{рег}} / S_{\text{общ}}. \tag{5.5}$$

Здесь

$$y_{\text{сум}} = \sum y_i, \quad S_{\text{общ}} = \sum \left(y_i - y_{\text{сум}}\right)^2, \quad S_{\text{рег}} = \sum \left[w(t_i) - y_{\text{сум}}\right]^2, \tag{5.6}$$

где суммирование по индексу i ведется от 1 до N.

Величина тренда T_r — изменение характеристики по линейному тренду за определенный промежуток времени:

$$T_r = \frac{w(t_N) - w(t_1)}{N}, \tag{5.7}$$

где N — длина ряда.

В линейном случае величина T_r определяется как:

$$T_r = \frac{(a_0 + a_1 t_N) - (a_0 + a_1 t_0)}{N} = \frac{a_1 (t_N - t_0)}{N} = a_1 h, \tag{5.8}$$

где h — временная дискретность временного ряда.

Отсюда видно, что величина линейного тренда определяется, прежде всего, коэффициентом регрессии a_1. Величина тренда равна коэффициенту a_1 линейного тренда и имеет размерность характеристики y на единицу дискретности.

5.3. Оценка значимости коэффициентов линейного временного тренда

При определении временного линейного тренда наиболее важным представляется оценка его значимости, т.е. насколько существенен его вклад в изменчивость временного процесса, поскольку вклад тренда в общую дисперсию временного ряда может быть как значимым, так и незначимым.

При оценке значимости линейного временного тренда используется критерий Стьюдента. Можно показать, что $t_r = t_{a_1}$. Это облегчает оценку значимости тренда, так как позволяет использовать уже рассмотренную ранее методику оценки значимости коэффициента корреляции к оценке значимости коэффициента a_1 в формуле (5.1). Иными словами, для оценки значимости линейного временного тренда достаточно проверить на значимость коэффициента корреляции r данных: $\{t_i, y_i, i = 1, ..., n\}$. В случае его незначимости считается, что тренда нет.

Записывается нулевая гипотеза по отношению к коэффициенту линейного тренда a_1 и коэффициенту корреляции r:

$$H_0 : |a_1| = 0, \quad H_0 : |r| = 0. \tag{5.9}$$

Для проверки этих гипотез рассчитывается выборочный критерий Стьюдента, причем можно показать, что $t_r = t_{a_1}$. Это облегчает оценку значимости тренда.

Тренд считается значимым, если оценки критерия Стьюдента превышают его критическое значение при заданном уровне значимости, т.е.

$$t > t_{\text{кр}}(\alpha, \nu = N - 1), \tag{5.10}$$

где $t_{кр}(\alpha, \nu = N - 1)$ — критическое значение статистики Стьюдента, соответствующее уровню значимости α и числу степеней свободы ν.

При оценке значимости нелинейного тренда рассчитывается корреляционное отношение, а затем осуществляется проверка нулевой гипотезы так же, как и для коэффициента корреляции. По величине коэффициента корреляции и корреляционного отношения легко определить коэффициент детерминации.

Если в исследуемой выборке и линейный и нелинейный тренды значимы, тогда при анализе предпочтение отдают нелинейному тренду, если он вносит значительно больший вклад (более чем на 5 %) в дисперсию выборки, или линейному — в обратном случае.

6. ВЫЯВЛЕНИЕ ГРУБЫХ ПОГРЕШНОСТЕЙ

В статистических рядах довольно часто можно обнаружить *выбросы* — резко выделяющиеся наблюдения, которые существенно отклоняются от распределения остальных выборочных данных. Эти данные могут отражать экстремальные свойства изучаемого явления (переменной) или быть обусловлены ошибками измерений или расчетов, возникающих в результате ручной или машинной обработки. В первом случае выбросы представляют особый интерес, поскольку они связаны обычно со стихийными природными процессами. Экстремальная аномалия обязательно должна учитываться в статистических расчетах, так как она отражает реально произошедшее катастрофическое событие.

Грубые ошибки могут приводить к существенному искажению получаемых результатов и соответственно к их неправильной интерпретации. В связи с этим выявление и исключение грубых ошибок (промахов), относящихся по характеру своего происхождения к случайным погрешностям, является важной задачей первичного анализа временных рядов.

Грубые ошибки в статистических данных должны выявляться, прежде всего, путем физического анализа и желательно в реальном режиме времени. Если они отличаются от основной массы данных на порядок и более, то их выявление и устранение не представляет особых затруднений и может быть осуществлено визуально. Значительно более сложной является задача нахождения промахов при ретроспективном анализе данных, особенно в тех случаях, когда они не слишком сильно отличаются от других результатов, а физический анализ процессов, приводящих к формированию сомнительных оценок в данных, оказывается невозможным. Очевидно, в этом случае без использования специальных статистических приемов не обойтись. Ниже рассмотрены наиболее простые методы.

Пусть внутри имеющейся выборки обнаружено измерение под номером $n - x_n$, которое резко выделяется среди значений имеющегося ряда. Необходимо решить вопрос о принадлежности x_n предыдущим $(n - 1)$ наблюдениям. Для решения этого вопроса сформируем отрезок из имеющейся выборки: $x_1, x_2, ..., x_n$ (последний элемент отрезка — исследуемое значение) и рассмотрим нулевую гипотезу:

$$H_0 : x = x_n.$$

Проверка этой гипотезы при условии нормальности исходных данных осуществляется с помощью критерия Стьюдента:

$$t^* = \left| x_{ср} - x_n \right| / \sigma, \tag{6.1}$$

где x_n — подлежащее проверке n-ое значение переменной величины x; $x_{ср}$ — среднее арифметическое значение; σ — среднее квадратическое отклонение.

Следует подчеркнуть, что используемые в соотношении (6.1) выборочные характеристики $x_{ср}$ и σ *вычисляются без учета* величины x_n! Затем проверяется выполнение неравенства:

$$t^* > t_{кр}\left(\alpha, \nu = n - 1\right), \tag{6.2}$$

где $t_{кр}(\alpha,\nu = n - 1)$ — критическое значение статистики Стьюдента, соответствующее уровню значимости α и числу степеней свободы ν.

Если это *неравенство выполняется*, то нулевая гипотеза отвергается и делается вывод, что резко отличающееся наблюдение x_n входит в противоречие с данной выборкой и поэтому может быть из нее *исключено*. Если это неравенство не выполняется, то мы можем полагать, что крайнее наблюдение x_n исключать нецелесообразно. После исключения крайнего значения данную процедуру можно повторить и для следующего по абсолютной величине максимального отклонения, но предварительно необходимо пересчитать $x_{ср}$ и σ для выборки нового объема.

Данный способ выявления грубых ошибок весьма прост и легко применим на практике, однако он имеет существенные недостатки. В частности, он оказывается нечувствительным, если в исследуемой выборке выбросы группируются вместе, но отстоят довольно далеко от основной массы наблюдений. Кроме того, далеко не всегда исходная выборка имеет нормальное распределение.

Пример 6.1. Выявление искусственного выброса для временного ряда значений температуры воздуха. Пусть имеется фрагмент временного ряда, содержащий $N = 13$ измерений температуры воздуха:

1,4; 1,4; 1,6; 1,6; 1,5; 1,8; 2; 2,3; 2,3; 2,4; 2,6; 3; 8.

Есть основание полагать, что $x_{13} = 8$ °C является выбросом. Для проверки этой гипотезы рассчитаем по 12-ти значениям выборки необходимые для проверки 13-го члена ряда параметры:

$$x_{\text{ср}} = 3{,}4, \quad \sigma = 1{,}4, \quad t^* = 3{,}3, \quad t_{\text{кр}}\left(\alpha = 0{,}95,\ \nu = 11\right) = 2{,}2.$$

По рассчитанным значениям критерия Стьюдента — (t^* и $t_{\text{кр}}$) проверяем выполнение неравенства 5.10 (разд. 5). Так как $t^* > t_{\text{кр}}$ (3,3 > 2,2), то неравенство (5.10) *выполняется* и нулевая гипотеза о том, что значение $x_{13} = 8$ принадлежит данной выборке при уровне значимости $\alpha = 0{,}95$, *отвергается*. Т. е. можно предположить, что наблюдение $x_n = 8$ °С входит в противоречие с данной выборкой и поэтому может быть из нее исключено.

7. ПРОВЕРКА СТАЦИОНАРНОСТИ ВРЕМЕННОГО РЯДА

Одной из первых задач анализа временных рядов является оценка стационарности имеющегося ряда наблюдений, потому что большинство последующих методов анализа требует, чтобы исследуемый ряд был стационарным. Под стационарностью понимают неизменность вероятностных (статистических) характеристик во времени. Соответственно, если характеристика испытывает изменения во времени, то можно говорить о нестационарности.

При классифицировании нестационарности временных рядов можно выделить три класса:

1. Нестационарность по математическому ожиданию, когда среднее значение характеристики за какой-либо период времени значительно отличается от её среднего значения за другой период;
2. Нестационарность по дисперсии, когда средняя изменчивость характеристики за какой-либо период времени значительно отличается от средней изменчивости за другой период;
3. Нестационарность по автокорреляционной функции (АКФ), когда в разные периоды времени у характеристики отмечается различная частотная структура.

Оценить стационарность можно на основании теории проверки статистических гипотез. Для этого временной ряд разбивается на части (периоды времени), для каждой из которых отдельно рассчитываются простые статистики (среднее, дисперсия, АКФ). Разбиение реализации на отдельные интервалы желательно осуществлять, исходя из закономерностей внутренней структуры рассматриваемого процесса, ибо каких-либо формальных критериев для этого нет. Затем попарно проводится проверка равенства этих характеристик для частей рядов. Если различия окажутся статистически значимыми, следовательно, временной ряд не стационарен по одной или по нескольким из рассматриваемых статистических характеристик. Если расхождение между вероятностными характеристиками для всех интервалов окажется незначимым, то делается вывод, что данный временной ряд является стационарным.

Общая схема проверки стационарности временного ряда:

1. Преобразовать статистический ряд, разбив его на две части.
2. Для каждой из частей рассчитать математическое ожидание, дисперсию и автокорреляционную функцию.

3. Проверить гипотезу о равенстве средних значений двух частей ряда. Сделать вывод о стационарности по математическому ожиданию.
4. Проверить гипотезу о равенстве дисперсий двух частей ряда. Сделать вывод о стационарности по дисперсии.
5. Проверить гипотезу о равенстве коэффициентов автокорреляции. Сделать вывод о стационарности по автокорреляционной функции.
6. Построить график всего временного ряда, на котором нанести:
   – отметку разбиения ряда на части;
   – отдельно для каждой части — среднее значение и стандартное отклонение.

Пример 7.1. Оценка стационарности временного ряда температуры воздуха за 22.10.2012–23.10.2012.

В качестве примера оценки стационарности временного ряда был взят временной ряд распределения температуры за 22.10.2012–23.10.2012, полученный с помощью автоматической метеорологической станции «Погода». Временной ряд был поделен на две равные части, для каждой из которых отдельно рассчитались среднее значение и дисперсия. Далее были построены доверительные интервалы для всех этих характеристик. Полученные результаты представлены в табл. 7.1.

Таблица 7.1

Результаты, полученные при оценке стационарности временного ряда температуры воздуха за 22.10.2012–23.10.2012

	$X_{ср}$	$(\bar{x} - \Delta x) < X_{ср} < (\bar{x} + \Delta x)$	σ	ΔD_1	ΔD_2	$D \cdot \Delta D_1 < D < D \cdot \Delta D_2$
Первая выборка	4,0	$3{,}6 < X_{ср} < 4{,}4$	1,4	0,8	1,3	$1{,}5 < 1{,}9 < 2{,}5$
Вторая выборка	2,7	$2{,}5 < X_{ср} < 2{,}9$	1,0	0,8	1,3	$0{,}8 < 1{,}0 < 1{,}3$
Оценка	Различия в средних значениях значимы, ряд нестационарен,			Различия в дисперсиях значимы, ряд нестационарен,		

Анализируя полученные данные, можно отметить, что расхождения между средними значениями и дисперсиями двух выборок значимо отличаются друг от друга. Это говорит о том, что рассматриваемый ряд является нестационарным как по среднему арифметическому значению, так и по дисперсии.

8. ФОРМИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ МЕТЕОРОЛОГИЧЕСКИХ ВЕЛИЧИН

8.1. Формирование модельных временных рядов

При проведении различных исследований, а также при тестировании различных программ для ПЭВМ, часто возникает необходимость в построении модельных временных рядов с заданными свойствами. Для такого построения рядов можно использовать достаточно простые соотношения, позволяющие выполнить такого рода моделирование. Рассмотрим формирование модельных временных рядов на примере приземной температуры воздуха.

Для моделирования временных рядов использовались следующие математические выражения.

1. Модельный временной ряд, формируемый по формуле (8.1), описывает изменение температуры воздуха (t — время, ч) и содержит *три составляющие*:
   - постоянную составляющую, равную 10 °С;
   - переменную составляющую, задаваемую синусоидой с амплитудой ±5 °С и периодом 24 ч;
   - случайную составляющую, которая имитирует случайную погрешность измерения с нулевым средним значением и среднеквадратичным отклонением 0,5 °С (вычисляется с использованием псевдослучайных чисел, равномерно распределенных на промежутке [0,1] и генерируемых командой *rnd*):

$$y(t) = 10 + 5 \cdot \sin\left(\frac{2\pi t}{24}\right) + 0{,}5\left(\sqrt{2\ln\left(\frac{1}{rnd(t)}\right)}\cos\left[2\pi\, rnd(t-1)\right]\right). \quad (8.1)$$

2. По формуле (8.2) можно сформировать временной ряд, у которого дополнительно появляется *четвертая составляющая*:
   - линейный временной тренд, имитирующий прогрев воздушной массы на 0,06 °С в час.

$$y(t) = 10 + 5 \cdot \sin\left(\frac{2\pi t}{24}\right) + 0{,}5\left(\sqrt{2\ln\left(\frac{1}{rnd(t)}\right)}\cos\left[2\pi\, rnd(t-1)\right]\right) + 0{,}06t. \quad (8.2)$$

3. По формуле (8.3) формируется временной ряд, у которого по сравнению с формулой (8.2) увеличивается амплитуда переменной составляющей — суточного изменения температуры воздуха (рис. 8.1):

$$y(t) = 10 + (5 + 0{,}03t) \cdot \sin\left(\frac{2\pi t}{24}\right) + \\ + 0{,}5\left(\sqrt{2\ln\left(\frac{1}{rnd(t)}\right)}\cos\left[2\pi\, rnd(t-1)\right]\right) + 0{,}06t. \qquad (8.3)$$

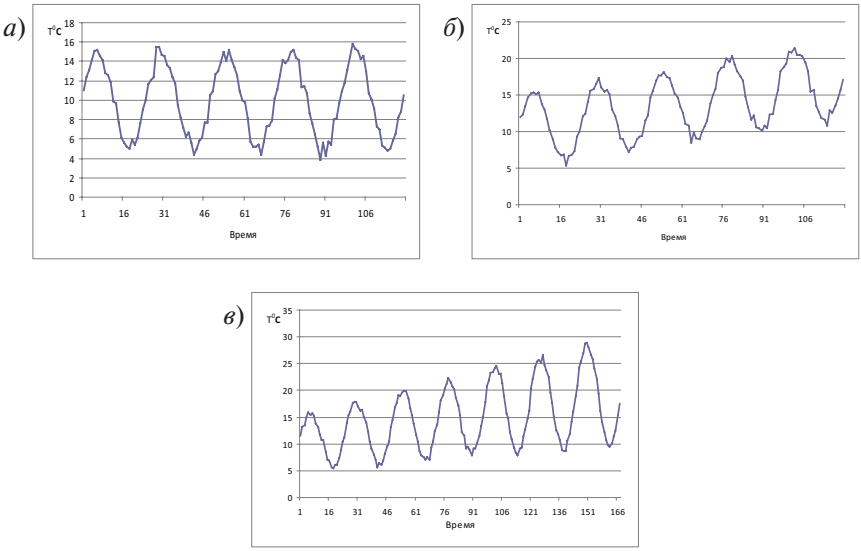


Рис. 8.1. Временные ряды температуры воздуха:
а — по формуле (8.1); *б* — по формуле (8.2); *в* — по формуле (8.3)

8.2. Формирование временных рядов по данным автоматических метеорологических станций

В процессе своей работы автоматическая метеорологическая станция «Погода», установленная в первом учебном корпусе РГГМУ, проводит измерения основных метеорологических величин с дискретностью 10 с. На основе этих измерений сформирован архив данных, содержащих следующие метеорологические параметры за 2009–2014 гг.:

- температура воздуха, °С;
- парциальное давление водяного пара, Па;
- атмосферное давление, Па;

– скорость ветра, м/с;
– направление ветра, градусы;
– зональная составляющая скорости ветра, м/с;
– меридиональная составляющая скорости ветра, м/с.

База метеорологических данных размещена на сервере. При отправке через сеть Интернет запроса по адресу: http://meteolab.rshu.ru:8080 пользователь получает ответ, представленный на рис. 8.2. С помощью этой страницы можно конкретизировать параметры запроса к базе метеорологических данных и получить временной ряд требуемых значений метеорологических величин в текстовом виде. Для этого, с помощью представленной на рис. 8.2 формы, можно указать, как иллюстрирует рис. 8.3, время начала блока данных, требуемую продолжительность блока и временную дискретность. Дискретность полученных данных может варьироваться от 10 с до 3 ч. Отмечая галочками, пользователь имеет возможность выбрать в списке из десяти параметров требуемые, которые затем и будут представлены в итоговом текстовом файле.

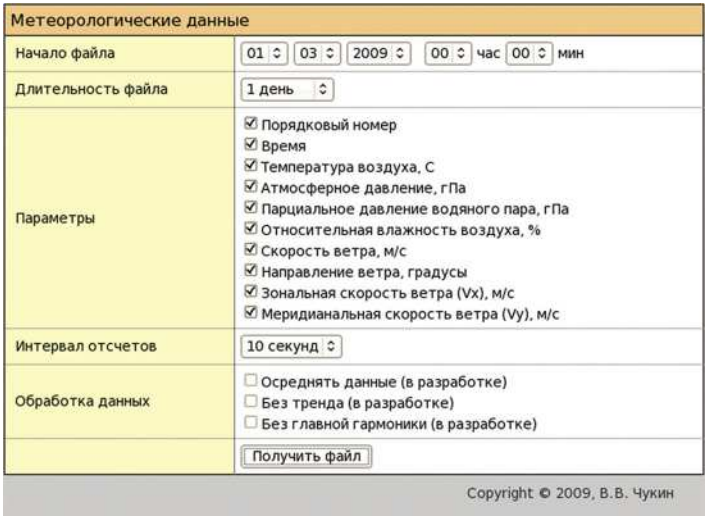


Рис. 8.2. Внешний вид пользовательского web-интерфейса

Рис. 8.3. Выбор начала блока данных, его продолжительности и дискретности временного ряда

После нажатия на кнопку Получить файл генерируется текстовый файл, который отображается в отдельном окне браузера и имеет вид, представленный на рис. 8.4.

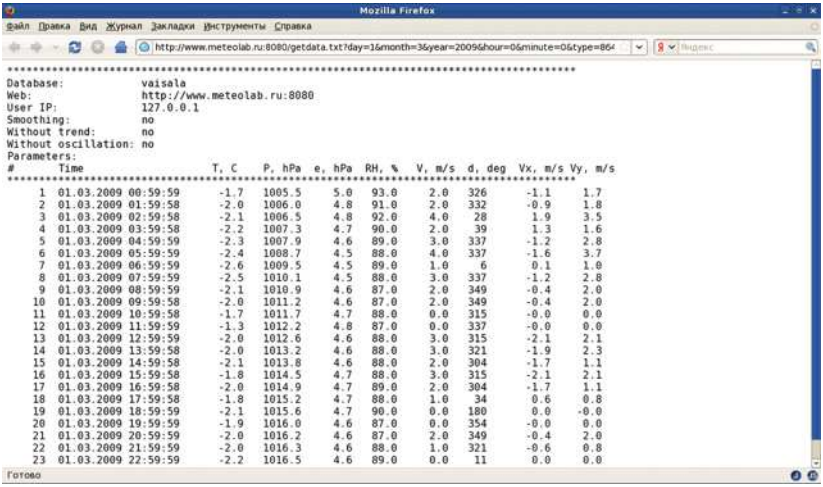


Рис. 8.4. Текстовый файл с метеорологическими данными

В первых строках текстового файла передается служебная информация: адрес ресурса в сети Интернет, IP-адрес пользователя и дополнительные параметры. Затем следует заголовок с указанием названий параметров и их размерности. Далее расположены строки со значениями метеорологических данных. Данный файл, присвоив ему соответствующее имя, необходимо сохранить для дальнейшего использования. Текстовый формат данных, получаемых из архива, может быть импортирован в табличный процессор *Excel*. Пример такого представления приведен в табл. 8.1.

Таблица 8.1

Файл с данными архива в формате пакета *Excel*

1	−12,2	999,9	2,1	87	2	56	1,7	1,1
2	−12,5	999,9	2,1	88	4	107	3,8	−1,2
3	−12,7	1000,1	2	88	2	107	1,9	−0,6
4	−12,8	1000,2	2	88	4	90	4	0
5	−12,9	1000,2	2	88	3	90	3	0
6	−13,3	1000,3	1,9	87	2	96	2	−0,2
7	−13,5	1000,4	1,9	86	3	112	2,8	−1,1
8	−13,8	1000,6	1,8	86	1	118	0,9	−0,5
9	−14,0	1000,6	1,8	86	4	84	4	0,4
10	−14,2	1000,8	1,8	86	3	73	2,9	0,9
11	−14,4	1000,8	1,7	86	4	51	3,1	2,5

Другой архив данных, полученных в Ленинградской области в поселке Воейково, можно получить, отправив запрос по адресу: http://www.fier867.0fees.net/iram/div.html. В ответ получаем стартовую страницу для выбора необходимых архивных данных, изображенную на рис. 8.5.

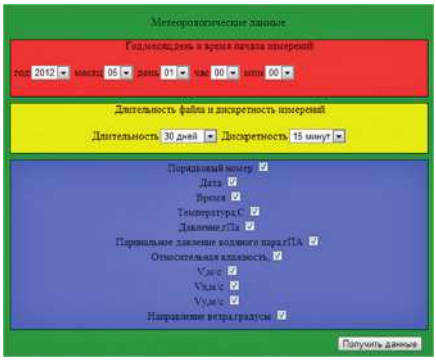


Рис. 8.5. Стартовая страница архива метеорологических данных Института радарной метеорологии (ИРАМ)

При работе с архивом ИРАМ существует возможность выбора даты (год, месяц, число) и конкретного времени начала необходимой выборки временного ряда метеорологических данных. Кроме того, можно ввести нужную длину временного ряда и дискретность, необходимую для дальнейшей обработки данных (рис. 8.6). Длительность ряда может быть выбрана в диапазоне от 15 мин до 30 дней, а дискретность измерений — от 1 мин до 6 ч, что позволяет использовать полученные временные ряды для решения широкого круга задач.

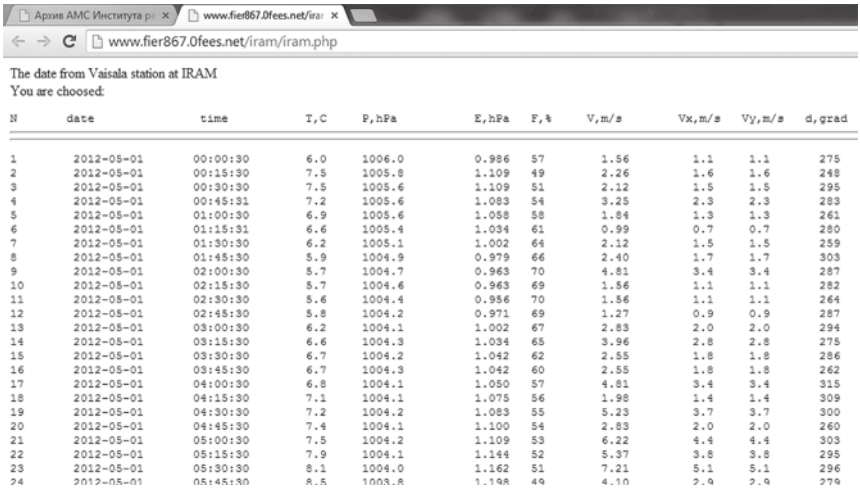
а)

б)

Рис. 8.6. Выбор параметров архивных данных:
а — длительность временного ряда; *б* — дискретность данных

Кроме того, существует возможность выбора перечня необходимых метеорологических величин (температура воздуха, атмосферное давление, парциальное давление, скорость ветра и др.).

При нажатии кнопки «Получить данные» генерируется файл с данными, которые отображаются в отдельном окне браузера и имеют вид, представленный на рис. 8.7.



N	date	time	T,C	P,hPa	E,hPa	F,%	V,m/s	Vx,m/s	Vy,m/s	d,grad
1	2012-05-01	00:00:30	6.0	1006.0	0.986	57	1.56	1.1	1.1	275
2	2012-05-01	00:15:30	7.5	1005.8	1.109	49	2.26	1.6	1.6	248
3	2012-05-01	00:30:30	7.5	1005.6	1.109	51	2.12	1.5	1.5	295
4	2012-05-01	00:45:31	7.2	1005.6	1.083	54	3.25	2.3	2.3	283
5	2012-05-01	01:00:30	6.9	1005.6	1.058	58	1.84	1.3	1.3	261
6	2012-05-01	01:15:31	6.6	1005.4	1.034	61	0.99	0.7	0.7	280
7	2012-05-01	01:30:30	6.2	1005.1	1.002	64	2.12	1.5	1.5	259
8	2012-05-01	01:45:30	5.9	1004.9	0.979	66	2.40	1.7	1.7	303
9	2012-05-01	02:00:30	5.7	1004.7	0.963	70	4.81	3.4	3.4	287
10	2012-05-01	02:15:30	5.7	1004.6	0.963	69	1.56	1.1	1.1	282
11	2012-05-01	02:30:30	5.6	1004.4	0.956	70	1.56	1.1	1.1	264
12	2012-05-01	02:45:30	5.8	1004.2	0.971	69	1.27	0.9	0.9	287
13	2012-05-01	03:00:30	6.2	1004.1	1.002	67	2.83	2.0	2.0	294
14	2012-05-01	03:15:30	6.6	1004.3	1.034	65	3.96	2.8	2.8	275
15	2012-05-01	03:30:30	6.7	1004.2	1.042	62	2.55	1.8	1.8	286
16	2012-05-01	03:45:30	6.7	1004.3	1.042	60	2.55	1.8	1.8	262
17	2012-05-01	04:00:30	6.8	1004.1	1.050	57	4.81	3.4	3.4	315
18	2012-05-01	04:15:30	7.1	1004.1	1.075	56	1.98	1.4	1.4	309
19	2012-05-01	04:30:30	7.2	1004.2	1.083	55	5.23	3.7	3.7	300
20	2012-05-01	04:45:30	7.4	1004.1	1.100	54	2.83	2.0	2.0	260
21	2012-05-01	05:00:30	7.5	1004.2	1.109	53	6.22	4.4	4.4	303
22	2012-05-01	05:15:30	7.9	1004.1	1.144	52	5.37	3.8	3.8	295
23	2012-05-01	05:30:30	8.1	1004.0	1.162	51	7.21	5.1	5.1	296
24	2012-05-01	05:45:30	8.5	1003.8	1.198	49	4.10	2.9	2.9	279

Рис. 8.7. Файл с метеорологическими данными

В верхней части файла передается служебная информация с указанием названия базы данных, затем следует заголовок с указанием названий выбранных параметров. Далее расположены столбцы со значениями метеорологических данных, готовых для дальнейшей обработки. Такие файлы пригодны для дальнейшей обработки в различных программах, в том числе в табличном процессоре *Excel*.

9. ЛАБОРАТОРНЫЕ РАБОТЫ

Как уже отмечалось ранее, временной ряд — это собранный в разные моменты времени материал о значении каких-либо параметров исследуемого процесса, т.е. конечную (по времени) реализацию случайной величины, расположенную в хронологическом порядке. Следует иметь в виду, что существуют принципиальные отличия временного ряда от последовательности наблюдений $x_1, x_2, ..., x_n$, образующих случайную выборку из этого временного ряда. Последнее означает, что в дальнейшем, осуществляя анализ, мы не сможем распространять все статистические свойства случайной выборки на весь временной ряд. Во временном ряду каждому отчету должно быть указано время измерения или номер измерения по порядку. При этом его значения могут быть взяты как через равные, так и через неравные промежутки времени, определяющие дискретность выполненных измерений. В первом случае временной ряд называется эквидистантным, а во втором — неэквидистантным.

Дискретность (Δ) представляет собой интервал времени, через который проводились измерения. Например, если измерения проводились через 2 с, то $\Delta = 2$ с; если через сутки — $\Delta = 24$ ч и т.п. Если сравнить данные, полученные с АМС с дискретностью $\Delta = 15$ мин и $\Delta = 3$ ч (рис. 9.1), то можно заметить, что при увеличении дискретности, мы теряем информацию о флуктуациях интересующей нас метеовеличины в то время, когда наблюдения не производились. Соответственно, чем больше дискретность, тем более осредненную характеристику мы получаем. Как правило, величина дискретности напрямую зависит от поставленных перед наблюдателем задач.

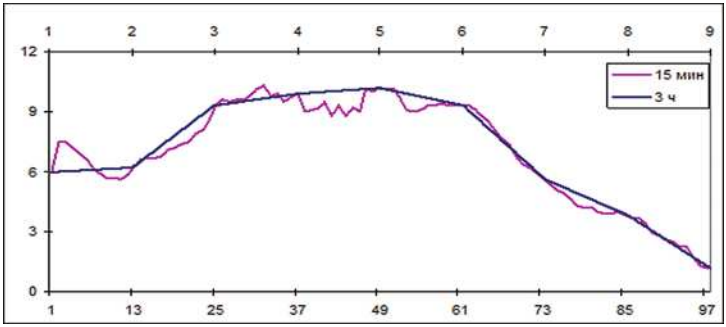


Рис. 9.1. Данные, полученные с АМС с различной дискретностью (15 мин и 3 ч)

Для того чтобы временные ряды, полученные с АМС, можно было использовать для исследований, прежде всего, необходимо перевести данные в цифровой формат, а также провести анализ ряда на разрывы и выбросы. Во временных рядах довольно часто можно обнаружить выбросы, связанные обычно с кратковременным сбоем работы аппаратуры. Как правило, в таком случае в ряду метеорологических данных наблюдается резко выделяющиеся значения, которые существенно отклоняются от распределения остальных выборочных данных. На рис. 9.2 представлен временной ряд температуры воздуха, содержащий выброс. На графике хорошо видно значение, отличающееся от основной массы данных. Устранение такого выброса не представляет особых затруднений и может быть осуществлено визуально.

Рис. 9.2. Фрагмент временного ряда температуры воздуха с выбросом

Разрывы, как правило, связаны с временным разрывом при регистрации данных, например, АМС не работала в течение нескольких часов и данные об изменении метеовеличины за этот промежуток отсутствуют (рис. 9.3).

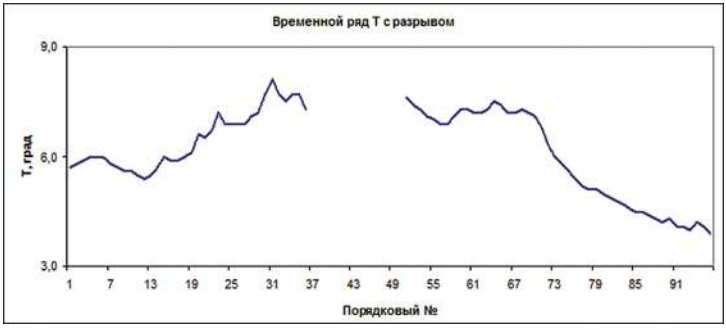


Рис. 9.3. Фрагмент временного ряда температуры воздуха с разрывом

В лабораторных работах рассмотрены методы использования временных рядов с метеоданными, полученными с АМС РГГМУ и ИРАМ, в

различных программах для дальнейших исследований. Для этого, прежде всего, необходимо перевести данные в цифровой формат и провести анализ ряда на разрывы и выбросы.

Лабораторная работа № 1
Формирование временных рядов от АМС РГГМУ и ИРАМ

Цель работы:

1. Сформировать временные ряды метеовеличин по данным АМС РГГМУ в текстовом виде с дискретностью 15 мин и 3 ч и сохранить их для дальнейшего использования. Даты начала и продолжительности временных рядов и используемый при этом метеорологический параметр или параметры указываются преподавателем.
2. Сформировать временные ряды метеовеличин по данным АМС ИРАМ в текстовом виде с дискретностью 15 мин и 3 ч и сохранить их для дальнейшего использования. Даты начала и продолжительности временных рядов и используемый при этом метеорологический параметр или параметры те же, что и в п. 1.
3. Перевести все полученные данные в пакет *Excel*, и подготовить файлы с данными измерений на АМС для дальнейшей обработки.

Порядок выполнения:

1. *Формирование временных рядов по данным АМС РГГМУ.* Запустить браузер. База метеорологических данных, измеренных АМС РГГМУ, размещена на сервере *meteolab.rshu.ru* и для доступа к данным используется *web*-интерфейс, реализованный на языке программирования *Java*. С помощью этой страницы можно отправить запрос к базе метеорологических данных и получить требуемые значения в текстовом виде.

В строке браузера набрать http://meteolab.rshu.ru:8080 и нажать *enter*. В ответ от сервера пользователь получает страницу, представленную на рис. 8.2 (см. разд. 8). Здесь содержится список из десяти параметров, которые пользователь имеет возможность получить в итоговом файле. Порядковый номер и время измерения относятся к обязательным параметрам (галочки не снимать). Отметить необходимые метеорологические параметры (например, температуру или влажность), снять галочки с остальных. Далее необходимо указать время начала блока данных, требуемую продолжительность блока и выбрать интервал отсчетов (дискретность Δ измерений) (см. разд. 8 рис. 8.3). Дискретность полученных данных может варьироваться от 10 с до 3 ч.

Нажать кнопку Получить файл. В отдельном окне браузера отобразится текстовый файл (см. разд. 8 рис. 8.4). В первых строках текстового файла передается служебная информация с указанием названия базы

данных, адреса ресурса в сети Интернет, *IP*-адресе пользователя и дополнительных параметрах предварительной обработки данных. Затем следует заголовок с указанием названий параметров. Далее расположены строки со значениями метеорологических данных. Для сохранения данных на диске компьютера необходимо в меню браузера Файл выбрать пункт Сохранить как... В появившемся окне указать каталог для сохранения файла, задать желаемое короткое имя файла (*k*12, *t*34 и т.д.) и нажать на кнопку Сохранить. Данный файл сохраняется для дальнейшего использования.

2. *Формирование временных рядов по данным АМС ИРАМ.* Архив, содержащий данные метеорологических наблюдений, проведенных в Ленинградской области в поселке Воейково, находится на сайте Института Радарной Метеорологии (ИРАМ). Запустить браузер. В строке браузера набрать http://www.fier867.0fees.net/iram/div.html или http://aiismeteo.rshu.ru и нажать *enter*. В ответ от сервера пользователь получает стартовую страницу, представленную на рис. 8.5 (см. разд. 8). Указать время начала блока данных и требуемую продолжительность блока.

Выбрать нужную длину временного ряда и дискретность, необходимую для дальнейшей обработки данных. Далее следует список из одиннадцати параметров, которые пользователь имеет возможность получить в итоговом файле. Порядковый номер, дата и время измерения относятся к обязательным параметрам (галочки не снимать). Отметить необходимые метеорологические параметры (например, температуру или давление), снять галочки с остальных.

Длительность ряда может быть выбрана в диапазоне от 15 мин до 30 дней, а дискретность измерений — от 1 мин до 6 ч (см. разд. 8 рис. 8.6).

Нажать кнопку Получить файл. В отдельном окне браузера отображается текстовый файл (см. разд. 8 рис. 8.7). В верхней части файла передается служебная информация с указанием названия базы данных, затем следует заголовок с указанием названий выбранных параметров. Далее расположены столбцы со значениями метеорологических данных, готовых для дальнейшей обработки. В открывшемся файле выделить полученные данные, от первого измерения до последнего. Щелкнуть правой кнопкой мыши и выбрать Копировать.

Открыть программу Блокнот. Щелкнуть правой кнопкой мыши по листу и выбрать пункт меню Вставить.

Для сохранения данных на диске компьютера необходимо в меню программы Блокнот Файл выбрать пункт Сохранить как... В появившемся окне указать каталог для сохранения файла, задать желаемое короткое имя файла и нажать на кнопку Сохранить. Такие файлы пригодны для дальнейшей обработки в различных программах, в том числе в пакете *Excel*.

3. *Перевод данных в пакет Excel и подготовка временных рядов данных измерений АМС для дальнейшей обработки.* Текстовый формат данных, получаемых из архива, весьма удобен для их импортирования и дальнейшего использования в пакете *Excel*, но требует конвертирования в цифровой формат при их использовании в программах, написанных на современных языках программирования.

Чтобы перевести архивные данные в формат *Excel*, необходимо:

1. Открыть новый документ *Excel*.
2. В меню программы во вкладке Данные выбрать пункт Импорт внешних данных и далее — Импортировать данные (рис. 1). В открывшемся окне найти сохраненный файл с метеоданными в формате *txt* и открыть его двойным щелчком левой кнопкой мыши (или нажать кнопку Открыть). После этого откроется окно Мастер текстов. При выборе формата поставить галку у пункта С разделителями, начать импорт с 11 строки (для базы данных ИРАМ — с 3 строки), язык выбрать Кириллица (DOS) и нажать Далее (рис. 2*а*).

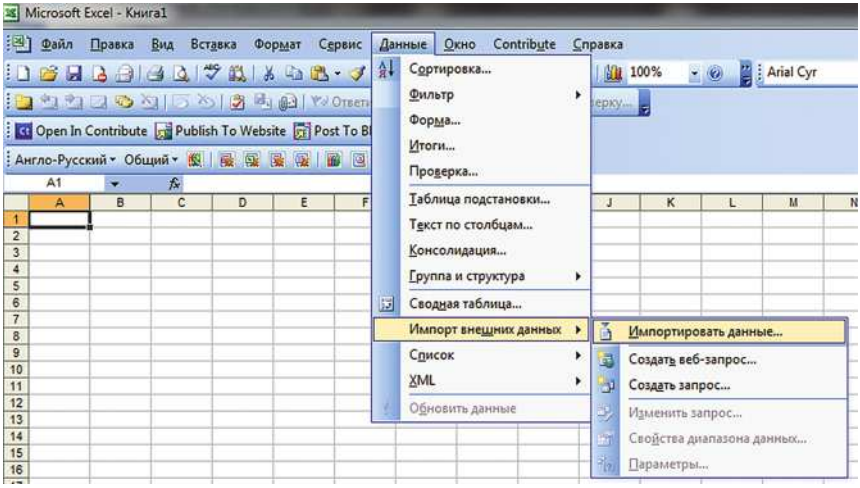
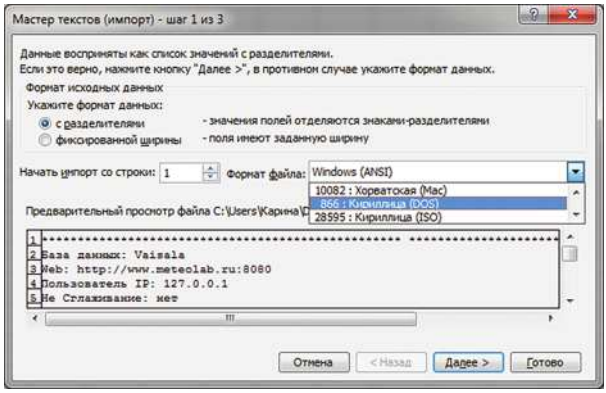


Рис. 1. Импорт внешних данных в программе *Excel* — начальный этап

3. Выбрать символ-разделитель Знак табуляции и Пробел (для базы данных ИРАМ только символ-разделитель Знак табуляции) и нажать Далее.
4. Открыть вкладку Подробнее и проверить Разделитель целой и дробной части (должна быть выбрана точка) (рис. 2*б*). После проверки нажать ОК.
5. Выбрать ячейку для начала данных и снова нажать ОК.

а)



б)

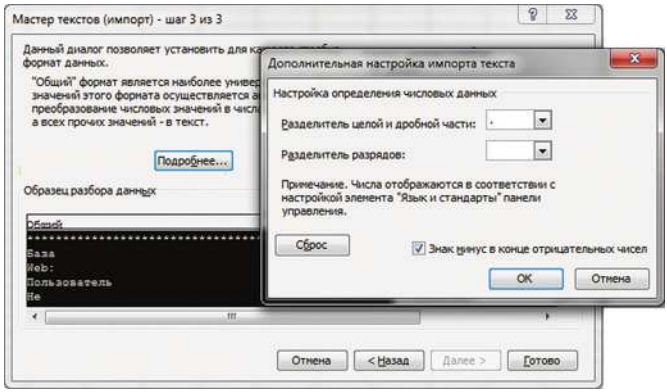


Рис. 2. Импорт внешних данных в программе *Excel* — промежуточный этап:
а — выбор формата исходных данных;
б — выбор разделителя целой и дробной части

Теперь выбранный временной ряд находится в пакете *Excel* (табл. 1) и может быть использован для дальнейшей обработки.

Таблица 1

Файл с данными архива в формате пакета *Excel* (фрагмент)

1	12.2	999.9	2.1	87	2	56	1.7	1.1
2	12.5	999.9	2.1	88	4	107	3.8	1.2
3	12.7	1000.1	2	88	2	107	1.9	0.6
4	12.8	1000.2	2	88	4	90	4	0
5	12.9	1000.2	2	88	3	90	3	0
6	13.3	1000.3	1.9	87	2	96	2	0.2
7	13.5	1000.4	1.9	86	3	112	2.8	1.1

Лабораторная работа № 2
Анализ временных рядов на выбросы и разрывы в пакете *Excel*

Цель работы:

1. Проанализировать временные ряды на разрывы и выбросы в программе *Excel*:
   – найти значение разницы между соседними данными метеовеличины (ΔT);
   – построить график зависимости данных исходного временного ряда и рассчитанных значений разницы между соседними данными метеовеличины (ΔT) от времени;
2. Проанализировать полученные результаты.

Порядок выполнения:

1. *Найти значение разницы между соседними данными метеовеличины* (ΔT).
Открыть чистый лист документа *Excel* и заполнить следующим образом:
   – столбец A содержит порядковый номер измерения;
   – столбец B — дату получения метеоданных;
   – столбец C — время измерения;
   – столбец D — значения метеовеличины (рис. 3*а*).
   – столбец E содержит значения разности метеоданных (например, температуры (ΔT)) между соседними значениями.

а)

	A	B	C	D	E
1	1249	16.05.2012	0:04:58	10.2	0.1
2	1250	16.05.2012	0:19:58	10.1	0
3	1251	16.05.2012	0:34:58	10.1	0.2
4	1252	16.05.2012	0:49:58	9.9	0
5	1253	16.05.2012	1:04:59	9.9	0.1
6	1254	16.05.2012	1:19:59	9.8	0
7	1255	16.05.2012	1:34:59	9.8	0
8	1256	16.05.2012	1:49:58	9.8	-0.1
9	1257	16.05.2012	2:04:58	9.9	-0.2

б)

C	D	E
0:04:58	10.2	=D1-D2
0:19:58	10.1	
0:34:58	10.1	
0:49:58	9.9	
1:04:59	9.9	
1:19:59	9.8	
1:34:59	9.8	

Рис. 3. Пример заполнения таблицы *Excel*:
а — введение данных; *б* — введение формулы (1)
для нахождения разности температур

Сосчитать разницу температур довольно легко, когда данных немного, однако, если количество измерений около 1000, процесс доставит уйму хлопот и займет много времени. Для того чтобы быстро заполнить столбец E, необходимо в ячейку E1 записать разности температур, определяемых выражением

$$\Delta T_1 = T_1 - T_2. \qquad (1)$$

Для этого необходимо выполнить следующие действия:

– выделить ячейку E1 нажатием на нее один раз левой кнопкой мыши;
– после выделения на клавиатуре набрать знак равенства =;
– левой кнопкой мыши щелкнуть на ячейку D1, при этом в ячейке E1 появляется часть формулы = D1 (значение D1 при этом будет выделено синим цветом);
– с клавиатуры ввести знак −;
– левой кнопкой мыши щелкнуть на ячейку D2, при этом в ячейке E1 появляется введенная формула = D1 − D2 (D1 выделено синим цветом, D2 — зеленым) (рис. 3*б*);
– нажать Enter. В ячейке E1 появилось сосчитанное программой значение ΔT;
– выделить ячейку E1 нажатием на нее один раз левой кнопкой мыши;
– подвести курсор в уголок выделенной ячейки и левой кнопкой мыши протянуть (нажали и не отпускаем) до предпоследнего значения в исходном временном ряду (рис. 4).

C	D	E
0:04:58	10.2	0.1
0:19:58	10.1	0
0:34:58	10.1	0.2
0:49:58	9.9	0
1:04:59	9.9	0.1
1:19:59	9.8	0
1:34:59	9.8	

Рис. 4. Нахождение разности температур для всего массива данных

2. *Построить график для анализа данных временных рядов РГГМУ на выбросы и разрывы в программе Excel.*

Выделить таблицу и в командном меню выбрать Вставка. В открывшемся списке выбрать строку Диаграмма. Откроется окно Мастер диаграмм (шаг 1 из 4): тип диаграммы. В этом окне выбрать вкладку Нестандартные, в списке — Графики (2 оси). Затем нажать Далее (рис. 5).

Открывается окно Мастер диаграмм (шаг 2 из 4): источник данных диаграммы. Во вкладке Диапазон данных в верхней части образец будущего графика, затем Диапазон — это выделенная таблица данных. В пункте Ряды выбрать В столбцах (рис. 6).

Затем открыть вкладку Ряд. В поле Ряд выбрать Ряд 1. В пункте Имя написать Исходные данные. В пункте Значения нажать на указатель в правом углу графы, выделить мышкой столбец D таблицы с данными (исходный временной ряд), нажать Enter (рис. 7).

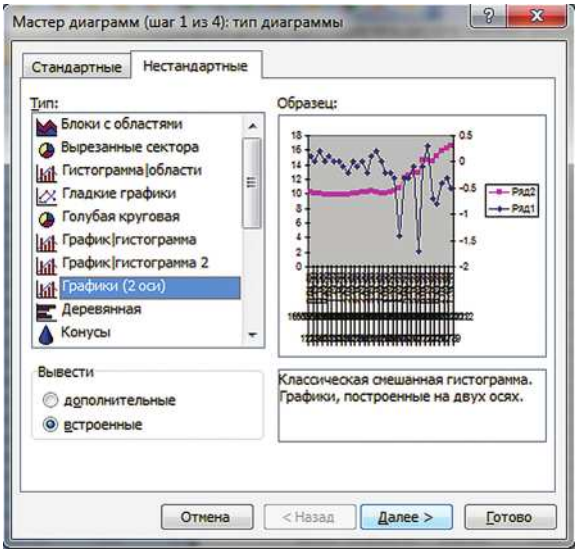


Рис. 5. Окно Мастер диаграмм (шаг 1 из 4)

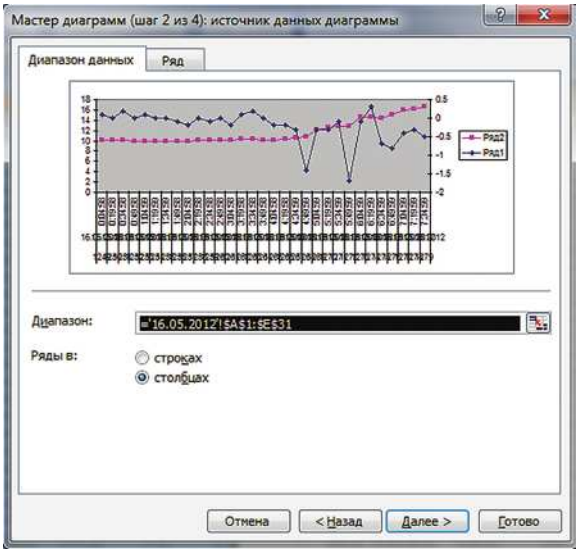


Рис. 6. Окно Мастер диаграмм (шаг 2 из 4) — вкладка Диапазон данных

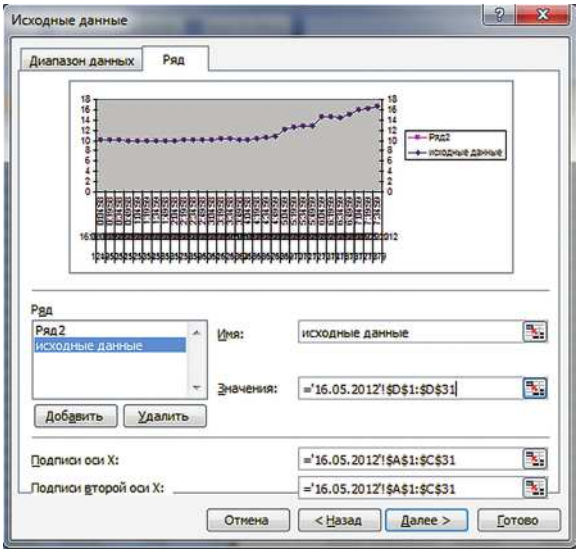


Рис. 7 Окно Мастер диаграмм (шаг 2 из 4) — вкладка Ряд

В поле Ряд выбрать Ряд 2. В пункте Имя написать Дельта Т. В пункте Значения нажать на указатель в правом углу графы, выделить мышкой столбец Е таблицы с данными (рассчитанные значения разницы температур), нажать *Enter*. В полях Подписи оси X и Подписи второй оси X нажать на указатель в правом углу графы, выделить мышкой столбец А таблицы с данными (порядковый номер измерения), нажать *Enter*. Затем нажать Далее.

Открывается окно Мастер диаграмм (шаг 3 из 4): параметры диаграммы. Во вкладке дать название диаграмме, оси X, и двум осям Y. Во вкладке Легенда, создать ее (поставив галочку в соответствующем окне) и определить ее положение на графике (рис. 8). Легенда нужна для того, чтобы сразу видеть, какие данные на графике что означают. Нажать Готово.

На завершающем шаге выбрать ячейку, где можно увидеть созданный график. Выбрав, нажать на клавишу Готово.

На готовом графике правой кнопкой мыши нажать на ось Y, в открывшемся списке выбрать Формат оси. В открывшемся окне (рис. 9) в пункте Ось X (категорий) пересекается в значении поставить минимальное значение по шкале Y. Остальные значения изменяются в зависимости от построенного графика (рис. 10).

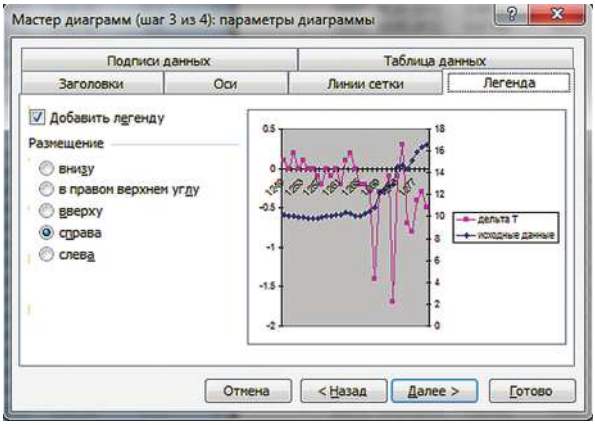


Рис. 8 Окно Мастер диаграмм (шаг 3 из 4)

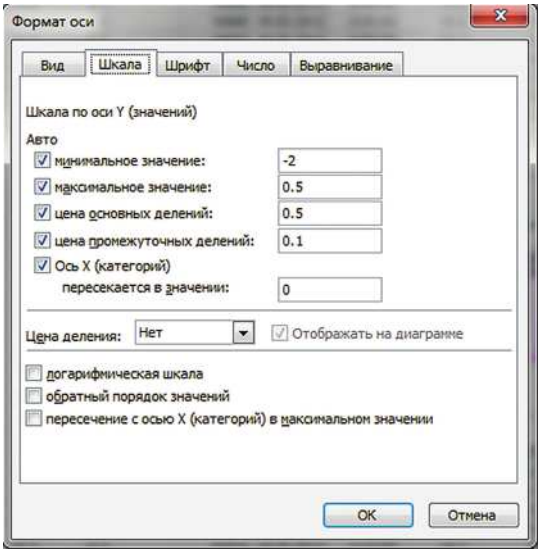


Рис. 9. Окно Формат оси

Рис. 10. Итоговый вид графика

Если вдруг, по какой-то причине, не устраивает результат, исправить график можно прямо в нем (щелчком правой кнопкой мыши из контекстного меню) или в главном меню Диаграмма. Подпункты этой вкладки содержат все рассмотренные выше шаги.

Далее необходимо проанализировать полученный график. Если во временном ряду присутствуют выбросы, то необходимо провести их исследование так, как это описано в разделе 6. Далее составить итоговую таблицу с указанием порядкового номера проверяемого измерения и эмпирического и критического значений критерия Стьюдента при уровне значимости $\alpha = 0{,}95$. Если во временном ряду выявлены разрывы, необходимо разбить исходный временной ряд на части, не содержащие временные разрывы. Этот простейший способ поможет избежать ошибки при оценке статистических характеристик ряда.

Аналогично выполняется анализ временных рядов на разрывы и выбросы в программе *Excel* для архивных данных, полученных с АМС ИРАМ.

Лабораторная работа № 3
Построение временных рядов с различной дискретностью в программе *Excel*

Цель работы:

1. Построение временных рядов с различной дискретностью в программе *Excel*.
2. Анализ полученного материала.

Порядок выполнения:

Данная лабораторная работа выполняется аналогичным образом для временных рядов, полученных с АМС РГГМУ и ИРАМ.

1. Открыть новый лист документа *Excel* и импортировать метеоданные из формата *txt* (см. лаб. раб. №1, п. 3) в следующем порядке:

– в столбец A поместить данные, полученные с дискретностью 15 мин;
– в столбец F импортировать данные с дискретностью 3 ч;
– (каждый массив данных должен содержать порядковый номер, дату и время измерения и, конечно, измеряемую метеовеличину).

2. Построить на одном графике две кривые — временные ряды с дискретностью 15 мин и 3 ч. Для этого выполнить следующие действия.

Построение графика выполнять аналогично лабораторной работе № 2 до работы с вкладкой Ряд.

Открыть вкладку Ряд. В поле Ряд выбрать Ряд 1. В пункте Имя написать 15 мин. В пункте Значения нажать на указатель в правом углу графы, выделить мышкой данные столбца D таблицы (исходный временной ряд с дискретностью 15 мин), нажать *Enter* (рис. 11). После этого под полем Ряд нажать Добавить, выбрать Ряд 2. В пункте Имя написать 3 ч. В пункте Значения нажать на указатель в правом углу графы, выделить мышкой столбец I таблицы с данными (исходный временной ряд с дискретностью 3 ч), нажать *Enter*. В поле Подписи оси X нажать на указатель в правом углу графы, выделить мышкой столбец A таблицы с данными (порядковый номер измерения). В поле Подписи второй оси X нажать на указатель в правом углу графы, выделить мышкой столбец F таблицы с данными (порядковый номер измерения), нажать *Enter*. Затем нажать Далее.

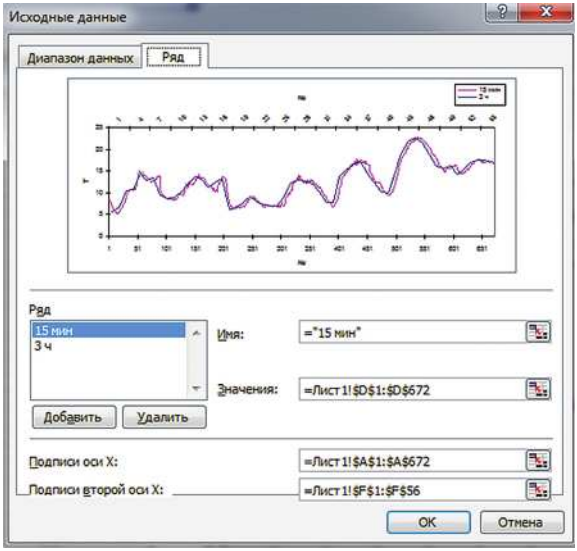


Рис. 11. Окно Мастер диаграмм (шаг 2 из 4) — вкладка Ряд

3. Открывается окно Мастер диаграмм (шаг 3 из 4): параметры диаграммы. Во вкладке Оси поставить галочку в пункте По вспомогательной оси, ось X (категорий) и убрать галочку в пункте По вспомогательной оси, ось Y (значений) (рис. 12)

Рис. 12. Окно Мастер диаграмм (шаг 3 из 4) — вкладка Оси

Во вкладке Заголовки дать название диаграмме, двум осям *x*, и оси *y* (при этом поле Вторая ось Y значений не заполняется) (рис. 13).

Во вкладке Легенда создать ее (поставив галочку в соответствующем окне) и определить ее положение на графике. Нажать Готово.

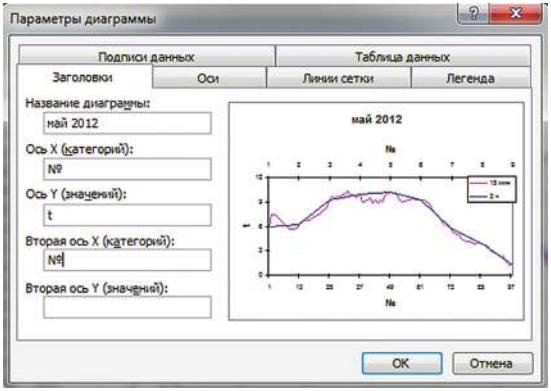


Рис. 13. Окно Мастер диаграмм (шаг 3 из 4)

4. На готовом графике правой кнопкой мыши нажать на верхнюю ось *x*, в открывшемся списке выбрать Формат оси. В открывшемся окне (рис. 14) выбрать вкладку Шкала, задать число категорий между подписями и делениями равным 3 (рис. 14*а*), снять галочку в пункте Пересечение с осью

Y (значений) между категориями и нажать OK. Затем проделать то же самое с нижней осью *x*, но число категорий задать равным 50 (рис. 14*б*).

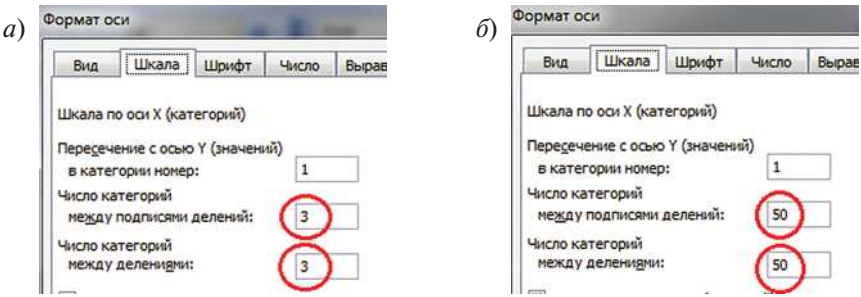


Рис. 14. Окно Формат оси — вкладка Шкала:
а — параметры верхней шкалы *x*; *б* — параметры нижней шкалы *x*

Если вдруг, по какой-то причине, не устраивает результат, исправить график можно прямо в нем (щелчком правой кнопкой мыши из контекстного меню) или в главном меню Диаграмма. Подпункты этой вкладки содержат все рассмотренные выше шаги.

5. Провести анализ готового графика (рис. 15).

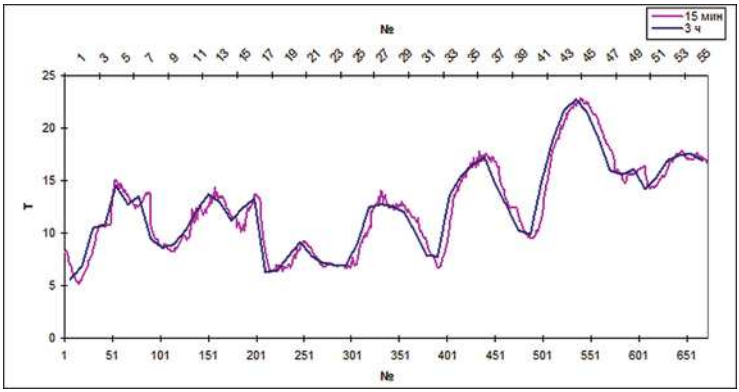


Рис. 15. Итоговый вид графика

Лабораторная работа № 4
Оценка статистических характеристик временных рядов от АМС РГГМУ и ИРАМ

Для расчета статистических характеристик временных рядов используются показатели описательной статистики. Все эти показатели

легко можно получить с помощью пакета Описательной статистики программы *Excel*. Программа дает возможность рассчитать следующие статистические характеристики: среднее значение ряда, стандартную ошибку, медиану, моду, стандартное отклонение, дисперсию выборки, эксцесс, асимметрию, интервал, минимальное и максимальное значения, количество обработанных значений.

Цель работы:

1. Рассчитать статистические характеристики временных рядов с дискретностью 15 мин и 3 ч в пакете *Excel*.
2. Сравнить и проанализировать полученные результаты.
3. Рассчитать доверительные интервалы для среднего и дисперсии выборки.
4. Построить гистограммы распределения временных рядов с дискретностью 15 мин и 3 ч. Провести анализ распределения временных рядов.

Порядок выполнения:

1. *Подготовка исходных данных*. Разбить исходный временной ряд с дискретностью 15 мин. на три фрагмента: первые 7 дней, вторые 7 дней и 14 дней, объединяющие первые и вторые 7 дней.

Например:

а) данные с 1 по 7 число месяца включительно (672 измерения);
б) данные с 8 по 14 число месяца включительно (672 измерения);
в) данные с 1 по 14 число месяца (1344 измерения).

Сохранить полученные ряды в формате *Excel*.

2. *Расчет статистических характеристик временных рядов с дискретностью 15 мин в пакете Excel*. Расчет выполняется по единой схеме для данных, полученных с АМС РГГМУ и ИРАМ.

Открыть сохраненный файл в формате *Excel*. Рассчитать статистические характеристики для всего блока данных (t, P, f, e и т.д.) по трем фрагментам исходного ряда данных (7 + 7 + 14) с помощью пакета Анализ данных. Для этого последовательно для каждого массива данных в меню программы выбрать Сервис, Анализ данных. В появившемся в отдельном окне списке выбрать строку Описательная статистика. Нажать ОК.

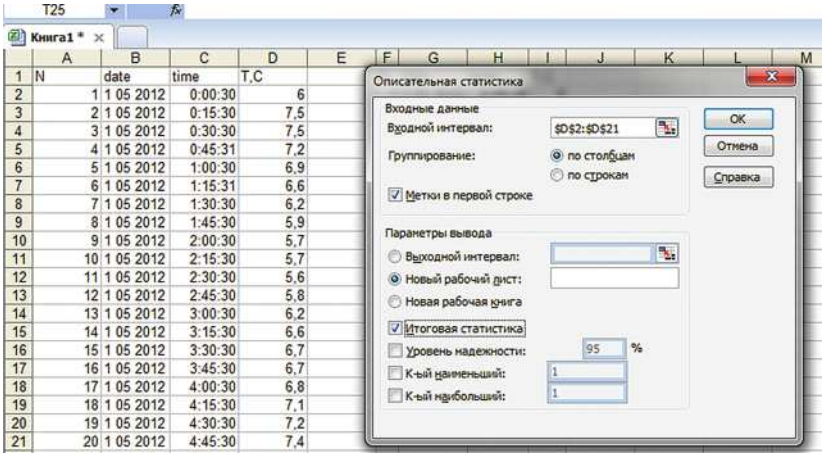
Во вкладке Описательная статистика (рис. 16) вводим входной интервал (значения блока данных за период). Выбираем Группирование по столбцам, Метки в первой строке параметры ввода: Новый рабочий лист. Ставим галку у строки Итоговая статистика. Нажимаем ОК.

Открывается новый лист с результатами подсчета программой статистических характеристик (рис. 16*б*). Округлить полученные значения.

Выделить столбец со значениями, нажав правую кнопку мышки, и в появившемся списке выбрать Формат ячеек. В открывшемся окне выбрать вкладку Число (рис. 17), формат — Числовой, ввести число десятичных знаков, необходимое для решения конкретной задачи (для температуры воздуха — 1, для относительной влажности — 0 и т.д.). Нажать OK.

Провести сравнительный анализ статистических характеристик рядов (7 + 7 + 14). Пример расчета статистических характеристик приведен в табл. 2.

а)



	A	B	C	D
1	N	date	time	T,C
2	1	1 05 2012	0:00:30	6
3	2	1 05 2012	0:15:30	7,5
4	3	1 05 2012	0:30:30	7,5
5	4	1 05 2012	0:45:31	7,2
6	5	1 05 2012	1:00:30	6,9
7	6	1 05 2012	1:15:31	6,6
8	7	1 05 2012	1:30:30	6,2
9	8	1 05 2012	1:45:30	5,9
10	9	1 05 2012	2:00:30	5,7
11	10	1 05 2012	2:15:30	5,7
12	11	1 05 2012	2:30:30	5,6
13	12	1 05 2012	2:45:30	5,8
14	13	1 05 2012	3:00:30	6,2
15	14	1 05 2012	3:15:30	6,6
16	15	1 05 2012	3:30:30	6,7
17	16	1 05 2012	3:45:30	6,7
18	17	1 05 2012	4:00:30	6,8
19	18	1 05 2012	4:15:30	7,1
20	19	1 05 2012	4:30:30	7,2
21	20	1 05 2012	4:45:30	7,4

б)

	G	H
	Столбец1	
	Среднее	5.75
	Стандартная ошибка	0.058387421
	Медиана	5.75
	Мода	6
	Стандартное отклонение	0.202259959
	Дисперсия выборки	0.040909091
	Эксцесс	-1.003851852
	Асимметричность	-0.197765293
	Интервал	0.6
	Минимум	5.4
	Максимум	6
	Сумма	69
	Счет	12

Рис. 16. Последовательность действий при использовании опции Описательная статистика в меню Анализ данных:
а — вкладка Описательная статистика с указанием данных из колонки D;
б — результаты расчета статистических характеристик программой *Excel*

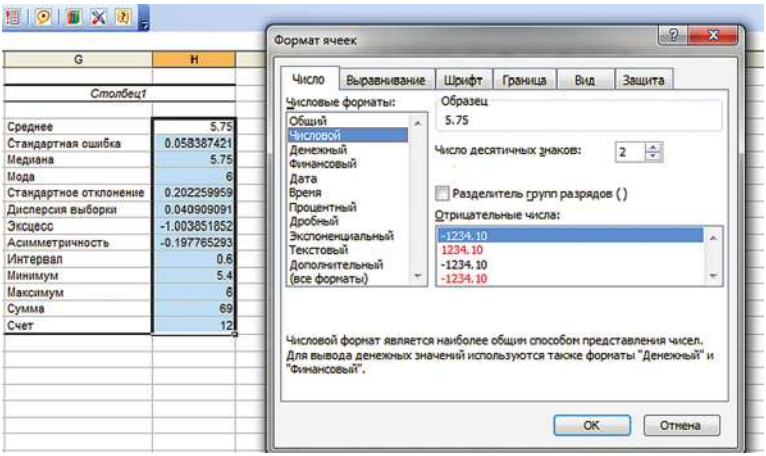


Рис. 17. Выбор формата ячеек в программе *Excel*

Таблица 2

Результаты расчета статистических характеристик блока данных программой *Excel*

	T, C	*P*, hPa	*E*, hPa	*F*, %	*V*, m/s	*Vx*, m/s	*Vy*, m/s	*d*, grad
Среднее	11,97	1007,15	7,52	53,68	2,25	1,59	1,59	212,17
Стандартная ошибка	0,18	1,34	0,13	0,81	0,07	0,05	0,05	3,30
Медиана	12,40	1012,30	6,41	53,00	1,70	1,20	1,20	220,00
Мода	15,70	1015,70	7,75	29,00	0,99	0,70	0,70	215,00
Стандартное отклонение	4,78	34,63	3,47	21,12	1,80	1,28	1,28	85,43
Дисперсия выборки	22,87	1199,54	12,03	446,10	3,26	1,63	1,63	7297,94
Эксцесс	−0,63	301,97	0,79	−1,14	3,31	3,31	3,31	22,27
Асимметричность	−0,04	−16,99	1,30	0,32	1,53	1,53	1,53	2,03
Интервал	21,70	627,70	14,44	72,00	13,01	9,20	9,20	1014,00
Минимум	1,70	389,10	2,32	23,00	0,00	0,00	0,00	1,00
Максимум	23,40	1016,80	16,76	95,00	13,01	9,20	9,20	1015,00
Сумма	8043,4	676807,9	5052,87	36072	1512	1069,1	1069,1	142579,8
Счет	672	672	672	672	672	672	672	672

3. *Построение гистограммы распределения временного ряда*. Открываем файл с временным рядом в программе *Excel*. В рассчитанных статистических характеристиках смотрим значения максимального и минимального значений метеовеличины, округляем полученные значения до целого и разбиваем разность на заданное количество интервалов с промежутком в 2 °C (0, 2, ..., +28, +30). Внести рассчитанные данные интервалов карманов в свободный столбец (на рис. 18*а*) столбец Е).

В командной строке *Excel* выбрать Сервис, Анализ данных. В открывшемся окне выбрать инструмент анализа — Гистограмма, нажать ОК (рис. 18*б*).

а)

	A	B	C	D	E	F
69	05.09.2012	0:01:21	13.30	0		
70	05.09.2012	0:16:21	13.30	2		
71	05.09.2012	0:31:21	13.40	4		
72	05.09.2012	0:46:21	13.60	6		
73	05.09.2012	1:01:21	13.80	8		
74	05.09.2012	1:16:21	13.60	10		
75	05.09.2012	1:31:21	13.60	12		
76	05.09.2012	1:46:22	13.40	14		
77	05.09.2012	2:01:22	13.00	16		
78	05.09.2012	2:16:22	13.00	18		
79	05.09.2012	2:31:22	13.00	20		
80	05.09.2012	2:46:22	13.20	22		
81	05.09.2012	3:01:22	13.10	24		
82	05.09.2012	3:16:22	13.30	26		
83	05.09.2012	3:31:22	13.30	28		
84	05.09.2012	3:46:22	13.40	30		
85	05.09.2012	4:01:22	13.40			
86	05.09.2012	4:16:22	13.50			
87	05.09.2012	4:31:22	13.50			

б)

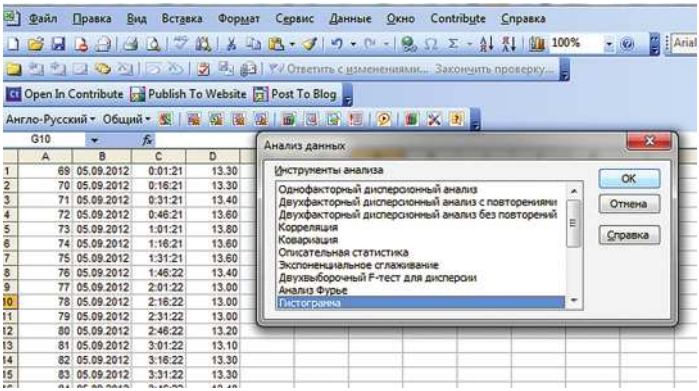


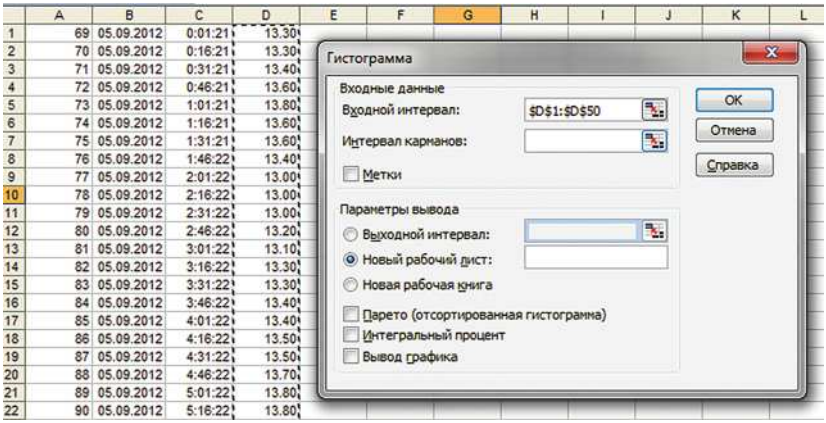
Рис. 18. Построение гистограммы распределения временного ряда в программе *Excel* — начальный этап:
а — пример заполнения рассчитанных интервалов карманов в колонку Е;
б — выбор инструмента анализа Гистограмма в программе *Excel*

В открывшемся окне ввести входной интервал — все значения выборки исследуемой метеорологической величины (рис. 19*а*).

Аналогичным образом в ячейку Интервал карманов вводятся границы записанных данных. В пункте Выходной интервал указать ячейку, у которой будет построена таблица распределения частот. В этом же окне поставить галочку Вывод графика и нажать ОК (рис. 19*б*). В результате проделанной работы в документе появится построенная гистограмма (рис. 20*а*).

Далее необходимо дать гистограмме название, подписать оси с помощью Мастера диаграмм (рис. 20*б*) и проанализировать полученные результаты распределения метеорологической величины трех фрагментов ряда (7 + 7 + 14).

а)



б)

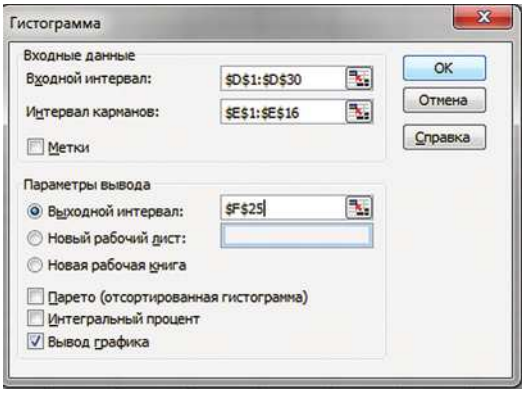
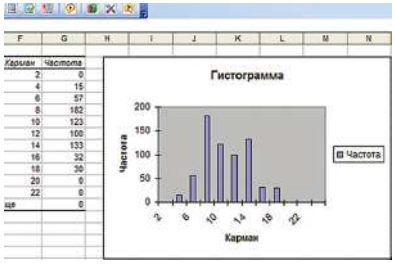


Рис. 19. Построение гистограммы распределения временного ряда в программе *Excel* — ввод данных:
а — выбор входного интервала в окне Гистограмма;
б — ввод данных для построения гистограммы

а)



б)

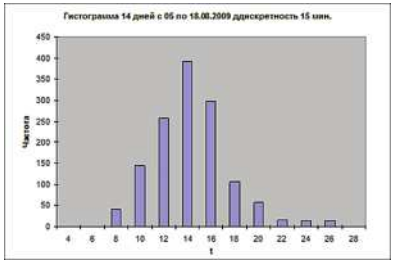


Рис. 20. Построение гистограммы распределения временного ряда в программе *Excel* — заключительный этап:
а — построенная гистограмма; *б* — законченная гистограмма

Лабораторная работа № 5
Сравнение временных рядов от АМС РГГМУ и ИРАМ за одинаковый период

Как известно, мегаполисы, к которым, в частности, относится и Санкт-Петербург, являются «островами тепла». Для того, чтобы проверить теорию «мегаполис — остров тепла» необходимы временные ряды, полученные в мегаполисе и за его пределами однотипными станциями. Для проведения такого исследования были выбраны АМС РГГМУ и АМС ИРАМ, как наиболее подходящие по всем параметрам. АМС РГГМУ расположена в Санкт-Петербурге, АМС ИРАМ — в поселке Воейково Ленинградской области. Расстояние между станциями — 20 км. В обоих пунктах расположены однотипные АМС («Погода»), укомплектованные датчиками фирмы *Vaisala*.

Цель работы:

1. Сравнить временные ряды метеовеличины АМС РГГМУ и АМС ИРАМ с дискретностью 15 мин и 3 ч.
2. Сравнить статистические характеристики временных рядов РГГМУ и ИРАМ с дискретностью 15 мин и 3 ч.
3. Построить и проанализировать гистограммы распределения временных рядов РГГМУ и ИРАМ с дискретностью 15 мин и 3 ч.

Порядок выполнения:

1. Получить данные временных рядов метеовеличины с АМС РГГМУ и АМС ИРАМ за одинаковый период с дискретностью 15 мин и 3 ч согласно варианту.

2. Для сравнения данных двух АМС построить график зависимости метеовеличины, с дискретностью 15 мин по времени. Оба временных ряда (РГГМУ и ИРАМ) должны находиться на одном графике (рис. 21). Провести анализ полученного графического изображения.

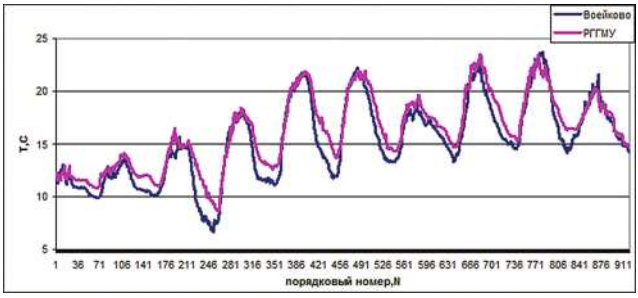


Рис. 21. Итоговый вид графика с данными РГГМУ и ИРАМ

Построить график зависимости временных рядов РГГМУ и ИРАМ с дискретностью 3 ч. Провести анализ полученного графического изображения.

3. Сравнить статистические характеристики временных рядов РГГМУ и ИРАМ с дискретностью 15 мин в пакете *Excel* (табл. 3). Проанализировать полученные результаты.

Сравнить статистические характеристики временных рядов РГГМУ и ИРАМ с дискретностью 3 ч в пакете *Excel*. Проанализировать полученные результаты.

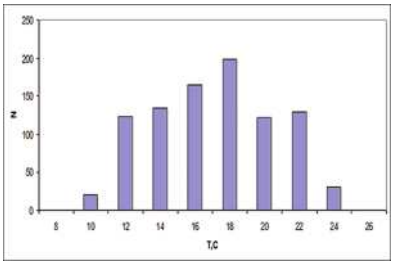
Таблица 3

Результаты расчета статистических характеристик временных рядов РГГМУ и ИРАМ с дискретностью 15 мин

	Санкт-Петербург РГГМУ май 2012	пос. Воейково ИРАМ май 2012
Среднее	12,29826389	10,25020833
Стандартная ошибка	0,135587319	0,120563258
Медиана	11,5	9,6
Мода	6,9	9,3
Стандартное отклонение	5,145176983	4,575053966
Дисперсия выборки	26,47284618	20,93111879
Эксцесс	−0,03179207	−0,357027564
Асимметричность	0,638219388	0,292806745
Интервал	24,9	22,8
Минимум	2,7	0,6
Максимум	27,6	23,4
Сумма	17709,5	14760,3
Счет	1440	1440

4. Построить гистограммы распределения временных рядов РГГМУ и ИРАМ с дискретностью 15 мин (рис. 22). Гистограммы распределения строить отдельно для временного ряда РГГМУ и данных ИРАМ (должно быть две гистограммы). Провести анализ полученных гистограмм.

а)



б)

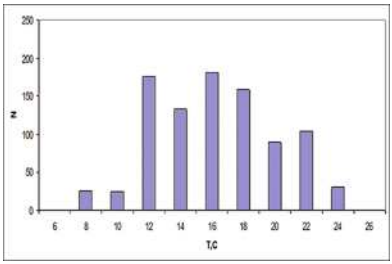


Рис. 22. Гистограммы распределения приземной температуры воздуха за 9 дней августа 2012 г. (с 10.08 по 20.08) в Санкт-Петербурге (*а*) и Воейково (*б*)

Построение гистограмм распределения временных рядов РГГМУ и ИРАМ с дискретностью 3 ч происходит аналогично. Провести анализ распределения.

Лабораторная работа № 6
Оценка коэффициентов корреляции метеорологических величин, полученных с АМС РГГМУ

Цель работы:

1. Оценить коэффициент корреляции метеорологических величин временных рядов, полученных по данным АМС РГГМУ за первые 7 дней ряда, вторые 7 дней ряда и за 14 дней временного ряда.
2. Провести оценку коэффициента автокорреляции метеовеличины, согласно варианту, за те же временные периоды.

Порядок выполнения:

1. Сформировать временные ряды за указанные периоды согласно варианту.

2. Рассчитать статистические характеристики фрагментов временного ряда (7 + 7 + 14) аналогично лабораторной работе № 4.

3. С помощью пакета Анализ данных, Корреляция рассчитать коэффициент корреляции метеопараметров для каждого фрагмента временного ряда (7 + 7 + 14). Для этого необходимо выбрать функцию Корреляция в пункте меню Анализ данных. В качестве входного интервала ввести блок данных, включая строку с наименованием метеовеличины и ее значениями. Поставить галку в графе Метки в первой строке (рис. 23), выбрать выходной интервал и нажать ОК.

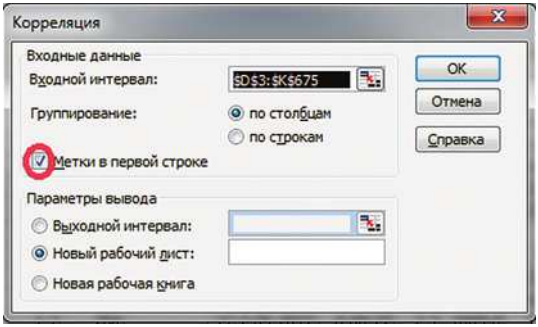


Рис. 23. Ввод данных для расчета коэффициента корреляции в программе *Excel*

В результате появляется таблица с результатом расчета коэффициента корреляции (табл. 4).

Таблица 4

Результат расчета коэффициента корреляции программой *Excel*

	T, C	*P*, hPa	*E*, hPa	*F*, %	*V*, m/s	*Vx*, m/s	*Vy*, m/s	*d*, grad
T, C	1							
P, hPa	−0,07119	1						
E, hPa	0,316003	−0,13765	1					
F, %	−0,38782	−0,08215	0,72925	1				
V, m/s	0,163391	0,072729	−0,20583	−0,29146	1			
Vx, m/s	0,163454	0,072731	−0,20584	−0,29151	0,999999	1		
Vy, m/s	0,163454	0,072731	−0,20584	−0,29151	0,999999	1	1	
d, grad	−0,06825	−0,39743	0,041747	0,090745	0,143429	0,143385	0,143385	1

Выделить столбцы, содержащие название метеовеличины и данные корреляции (рис. 24*а*), открыть Мастер диаграмм и построить гистограмму коэффициента корреляции между метеовеличинами (рис. 24*б*). Обработать гистограмму и проанализировать результаты расчета.

а)

	A	B	C
1		*T,C*	*P,hPa*
2	T,C	1,00	
3	P,hPa	-0,07	1,00
4	E,hPa	0,32	-0,14
5	F,%	-0,39	-0,08
6	V,m/s	0,16	0,07
7	Vx,m/s	0,16	0,07
8	Vy,m/s	0,16	0,07
9	d,grad	-0,07	-0,40

б)

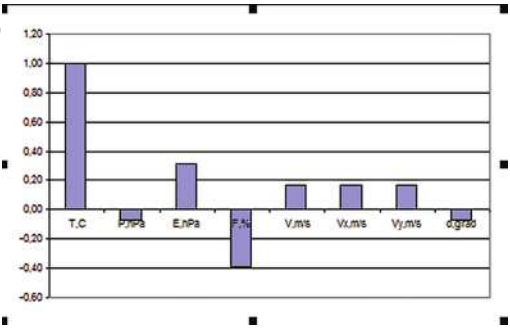


Рис. 24. Построение гистограммы,
иллюстрирующей корреляцию между величинами:
а — ввод данных; *б* — полученная гистограмма

4. Анализ и оценка коэффициента корреляции проводится согласно разделу 4.

5. Рассчитать коэффициент автокорреляции метеовеличины (согласно варианту) можно двумя способами:

- используя надстройку Автокорреляционная функция в программе *Excel*;
- с помощью дополнительного ввода данных.

Способ №1. Открыть меню Сервис – Автокорреляционная функция, в открывшемся окне (рис. 25) ввести входной интервал (значения метеовеличины), заполнить величину лага (например, 4) и количество переменных в использованной регрессии (например, 2) и нажать OK.

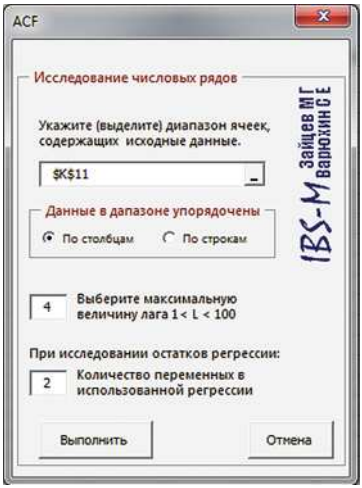


Рис. 25. Надстройка панели Автокорреляционная функция

В результате на новом листе откроется таблица с результатом расчета и построенной гистограммой (рис. 26).

	A	B	C	D	E	F	G
5		Исходный временной ряд	Статистика Дарбина-Ватсона (DW)		АКФ(...)	Ошибка АКФ	
6	1	6,0000	**0,004**	1	**0,720**	0,310	-0,310
7	2	7,5000	DW Up	2	**0,465**	0,638	-0,638
8	3	7,5000	1,540	3	**0,207**	0,703	-0,703
9	4	7,2000	DW Low	4	**-0,025**	0,715	-0,715
10	5	6,9000	1,100				
11	6	6,6000					
12	7	6,2000					
13	8	5,9000					

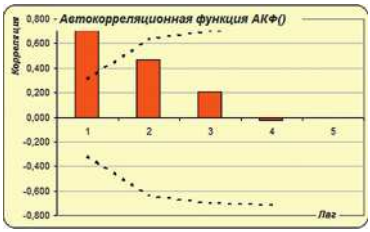


Рис. 26. Результат расчета автокорреляционной функции надстройкой *Excel*

Способ №2. Сдвинуть исходный ряд на один уровень (табл. 5), через функцию Анализ данных – Корреляция произвести расчет, оценить коэффициент автокорреляции согласно разделу 4.

Таблица 5

Подготовка к расчету коэффициента автокорреляции

N	y_t	y_{t-1}	N	y_t	y_{t-1}
1	6	-	6	6,6	6,9
2	7,5	6	7	6,2	6,6
3	7,5	7,5	8	5,9	6,2
4	7,2	7,5	9	5,7	5,9
5	6,9	7,2	10	5,7	5,7

Лабораторная работа № 7
Анализ временного ряда на наличие тренда

Цель работы:

1. Построить графики зависимости метеорологической величины от времени с линейным и полиномиальным трендом.
2. Провести анализ дисперсий временных рядов.
3. Провести анализ средних значений временных рядов.
4. Выполнить проверку наличия тренда для данных временных рядов.

Порядок выполнения:

1. Постройте графики зависимости метеорологической величины от времени (за первые 7, вторые 7 и 14 дней) и добавьте линии линейного тренда и полинома второй степени. При построении линии тренда не забудьте открыть вкладку Параметры (рис. 27) и поставить галку в пунктах Показывать уравнение на диаграмме и Поместить на диаграмму величину достоверности аппроксимации.

На графике исходный временной ряд, линейный и полиномиальный тренды оформите разным цветом (рис. 28). Проанализируйте полученные графики.

2. Для временных рядов за первые и вторые 7 дней рассчитайте эмпирическое значение критерия Фишера F^* с помощью функции FРАСПОБР (рис. 29) с вероятностью 0,05. Графы Степени_свободы 1 и 2 заполняются согласно выражению $N - 1$ (где N — количество значений временного ряда) для первых и вторых 7 дней соответственно. В нижней части окна Аргументы функции появляется рассчитанное программой значение F^*.

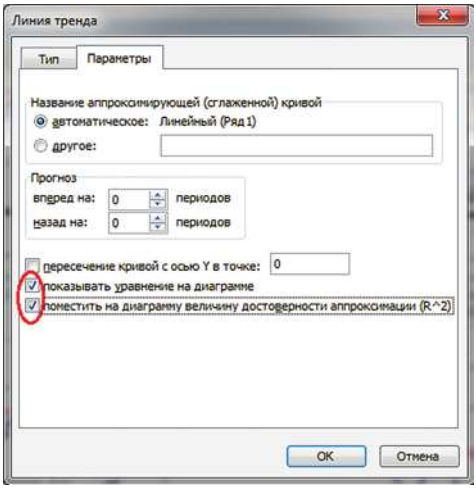


Рис. 27. Окно Линия тренда, вкладка — параметры

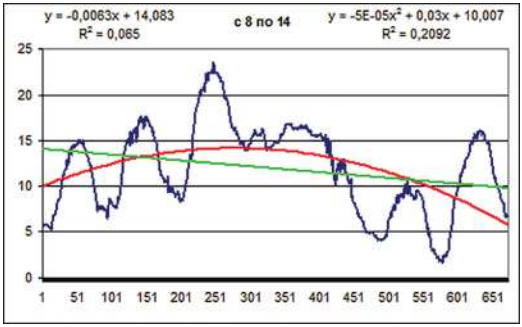


Рис. 28. Окончательный вид графика

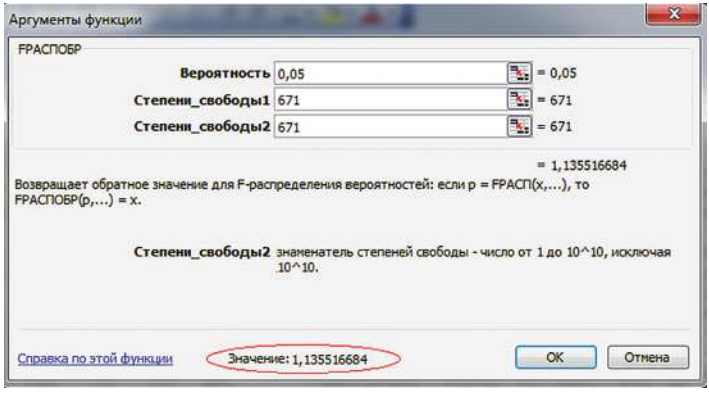


Рис. 29. Окно функции FРАСПОБР

3. Проверку гипотезы равенства дисперсий выполните в *Excel* с помощью *F*-теста (запустите Сервис – Анализ данных – Двухвыборочный *F*-тест для дисперсии). Заполните входные данные для выполнения *F*-теста, указав диапазон значений метеовеличины за первые и вторые 7 дней соответственно на уровне значимости 0,05. Проанализируйте результаты расчета (табл. 6) согласно разделу 2.

Таблица 6

Результат двухвыборочного *F*-теста для дисперсии

	Переменная 1	Переменная 2
Среднее	8,156696429	11,96934524
Дисперсия	10,92880755	22,87408124
Наблюдения	672	672
df	671	671
F	0,477781268	
$P(F<=f)$ одностороннее	0	
F критическое одностороннее	0,880656369	

Сравните эмпирическое (F^*) и критическое (в рамке табл. 6) значения критерия Фишера, проанализируйте результат.

4. Проверьте гипотезу о равенстве (однородности) дисперсий обеих частей ряда с помощью *F*-критерия Фишера (большее значение дисперсии разделите на меньшее). Проанализируйте результат.

5. Выполните проверку гипотезы равенства средних значений. С помощью функция СТЬЮДРАСПОБР найдите эмпирический критерий Стьюдента (t^*) согласно разделу 1.

Рассчитайте критическое значение критерия Стьюдента ($t_{кр}$) с помощью пакета *Excel* (Сервис – Анализ данных – Двухвыборочный *t*-тест с одинаковыми дисперсиями). Входные интервалы 1 и 2 содержат данные двух семидневных фрагментов соответственно. Результат расчета представлен в табл. 7.

Сравните эмпирическое (t^*) и критическое (в рамке табл. 7) значения критерия Стьюдента, проанализируйте результат согласно разделу 1.

Таблица 7

Результат двухвыборочного *t*-теста с одинаковыми дисперсиями

	Переменная 1	Переменная 2
Среднее	8,16	11,97
Дисперсия	10,93	22,87

	Переменная 1	Переменная 2
Наблюдения	672	672
Объединенная дисперсия	16,90	
Гипотетическая разность средних	0	
df	1342	
t-статистика	−16,9994367	
$P(T <= t)$ одностороннее	3,80195	
t критическое одностороннее	1,645989863	
$P(T <= t)$ двухстороннее	7,60391	
t критическое двухстороннее	1,961733219	

Лабораторная работа № 8
Регрессионный анализ

Цель работы:

1. Постройте поле корреляции и сформулируйте гипотезу о форме связи.
2. Рассчитайте параметры уравнения линейной регрессии $y = a + b \cdot x$.
3. Оцените тесноту связи с помощью показателей корреляции и детерминации.
4. Дайте сравнительную оценку силы связи фактора с результатом с помощью среднего (общего) коэффициента эластичности.
5. Оцените с помощью средней ошибки аппроксимации качество уравнений.
6. Оцените с помощью *t*-критерия Стьюдента и *F*-критерия Фишера статистическую надёжность результатов регрессионного моделирования.
7. Рассчитайте и проанализируйте параметры авторегрессии первого и второго порядка.
8. Постройте поле автокорреляции и сформулируйте гипотезу о форме связи.

Порядок выполнения:

1. В корреляционной матрице (см. лабораторную работу № 6, табл. 4) выбрать наибольший (по модулю) коэффициент корреляции для *t* (см. табл. 8). Подготовить исходные данные, выделив из них два соответствующих наибольшему коэффициенту ряда (в данном случае *t* и *f*) в отдельный массив.

Исходя из того, что между значениями *t* и *f* наблюдается зависимость, можно сделать предположение, что связь между признаками прямая и её можно описать уравнением прямой.

С помощью пакета *Excel* постройте поле корреляции. Выделите область ячеек, содержащую данные. Затем выберете Вставка – Точечная диаграмма – Точечная с маркерами.

Таблица 8

Фрагмент корреляционной матрицы

	t, °C
t, °C	1
p, hPa	−0,07119
e, hPa	0,316003
f, %	−0,38782
V, m/s	0,163391
Vx, m/s	0,163454
Vy, m/s	0,163454
d, grad	−0,06825

а) *б*)

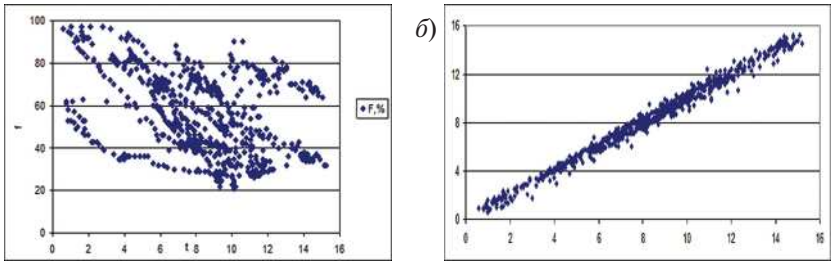


Рис. 30. Пример построения поля корреляции:
а — связь между переменными является случайной;
б — связь между переменными– функциональная

Затем следует проанализировать полученное поле корреляции на наличие корреляционной связи между переменными, рассматривая положение точек на графике. Если расположение точек близко к круговому (рис. 30*а*), то одному значению первой переменной соответствует любое значение второй переменной, связь между переменными является случайной, а коэффициент корреляции близок к нулю. Если эллипс вырожден в прямую линию (рис. 30*б*), то одному значению первой переменной соответствует одно значение второй переменной, связь между переменными функциональная и коэффициент корреляции близок к единице. В остальных случаях одному значению первой переменной соответствует некоторое значение второй с определенной вероятностью. Связь является стохастической и коэффициент корреляции находится между нулем и единицей.

2. Для расчёта параметров уравнения линейной регрессии используйте встроенную статистическую функцию пакета *Excel*: ЛИНЕЙН.

Откройте существующий файл, содержащий анализируемые данные и выделите в нем пустые ячейки (5 строк, 2 столбца) для вывода

результатов регрессионной статистики. Активизируйте Мастер функций, выберете ЛИНЕЙН. В открывшемся окне заполните аргументы функции следующим образом:

- диапазон, содержащий данные результативного признака (например, *f*);
- диапазон, содержащий данные факторного признака (например, *t*);
- Константа = 1 (если Константа = 0, то свободный член равен 0);
- Статистика = 1 (если Статистика = 0, то выводятся только оценки параметров уравнения).

Результат расчета появится в левой верхней ячейке выделенной области (первый элемент итоговой таблицы). Чтобы раскрыть всю таблицу, нажмите на клавишу <F2>, а затем на комбинацию клавиш <Ctrl> <Shift> <Enter> (одновременно). В итоговой таблице (табл. 9*а*) дополнительная регрессионная статистика выводится соответственно табл. 9*б*.

Таблица 9

Результат расчета функции «ЛИНЕЙН»: числовые значения (*а*), расшифровка полученных данных (*б*)

а)

−2,19307	73,8644
0,201366	1,772068
0,150408	17,24382
118,6134	670
35269,61	199224

б)

Значение коэффициента *b*	Значение коэффициента *a*
Среднеквадратическое отклонение *b*	Среднеквадратическое отклонение *a*
Коэффициент детерминации R^2	Среднеквадратическое отклонение *y*
F-статистика	Число степеней свободы
Регрессионная сумма квадратов $\sum(\hat{y}_x - \bar{y})^2$	Остаточная сумма квадратов $\sum(y - \hat{y}_x)^2$

Проанализируйте полученные данные следующим образом:

- коэффициент детерминации (R^2) означает, что 15 % вариации относительной влажности (*y*) объясняется вариацией *t*, а 85 % — действием других факторов, не включённых в модель.
- по полученному значению коэффициента детерминации рассчитать коэффициент корреляции: $r_{xy} = \sqrt{R^2}$. Сравнить полученное значение со значением коэффициента корреляции в лабораторной работе № 3. Сделать вывод о связи переменных.

3. С помощью среднего (общего) коэффициента эластичности определить силу влияния фактора на результат:

$$\overline{Э} = f'(x)\frac{\overline{x}}{\overline{y}}, \qquad (2)$$

где $\bar{Э}$ — средний коэффициент эластичности; $f'(x)$ — значение коэффициента b при расчете регрессионной статистики; $\bar{x}$ — среднее значение t исходного массива данных; $\bar{y}$ — среднее значение f исходного массива данных.

Средние значения величин t и f найдите, выделив данные (поочередно), с помощью функции Автосумма – Среднее. Результат расчета средних значений представлен в табл. 10.

Таблица 10

Фрагмент таблицы с исходными данными и расчетом средних значений величин *t* и *f*

664	7 05 2012	21:45:04	6	68
665	7 05 2012	22:00:04	6	68
666	7 05 2012	22:15:04	5,8	69
667	7 05 2012	22:30:04	5,8	70
668	7 05 2012	22:44:59	5,9	69
669	7 05 2012	23:00:04	5,9	70
670	7 05 2012	23:15:03	6,1	69
671	7 05 2012	23:30:07	5,9	71
672	7 05 2012	23:45:07	5,7	71
		Среднее	8,16	56

В результате расчета средний коэффициент эластичности получился равным 0,32 %. Таким образом, при изменении t на 1 % от своего среднего значения f изменится в среднем на 0,32 %.

4. С помощью инструмента анализа данных Регрессия можно получить:
- результаты регрессионной статистики;
- результаты дисперсионного анализа;
- результаты доверительных интервалов;
- остатки и графики подбора линии регрессии;
- остатки и нормальную вероятность.

В главном меню последовательно выберите: Сервис – Анализ данных – Регрессия. Заполните диалоговое окно ввода данных и параметров вывода (рис. 31).

Результаты регрессионного анализа для данных представлены на рис. 32.

5. Оцените с помощью средней ошибки аппроксимации качество уравнений. Для этого воспользуйтесь результатами регрессионного анализа Вывод остатка (в нижней части рис. 32). Чтобы раскрыть всю таблицу, нажмите на клавишу <F2>, а затем на комбинацию клавиш <Ctrl><Shift><Enter> (одновременно).

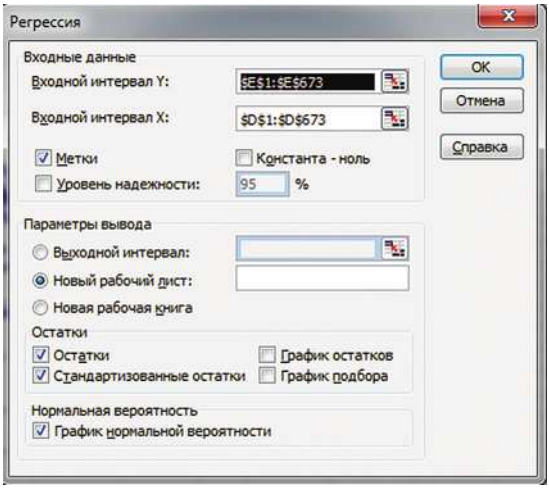


Рис. 31. Окно ввода параметров Регрессия

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*рессионная статистика*								
4	Множеств	0,387824							
5	R-квадрат	0,150408							
6	Нормиров	0,149139							
7	Стандартн	17,24382							
8	Наблюден	672							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*ачимость F*			
12	Регрессия	1	35269,61	35269,61	118,6134	1,52E-25			
13	Остаток	670	199224	297,3493					
14	Итого	671	234493,6						
15									
16		*Коэффициенндартная ошатистиt-Значениижние 95%ерхние 95%ижние 95,0рхние 95,0*							
17	Y-пересеч	73,8644	1,772068	41,6826	3E-188	70,38493	77,34388	70,38493	77,34388
18	T,C	-2,19307	0,201366	-10,891	1,52E-25	-2,58845	-1,79769	-2,58845	-1,79769
19									
20									
21									
22	ВЫВОД ОСТАТКА					ВЫВОД ВЕРОЯТНОСТИ			

Рис. 32. Результат применения инструмента Регрессия

Составьте новую таблицу (рис. 33*а*). В столбец А копированием внесите ряд зависимой величины. В столбец «В» — значения рассчитанных программой остатков. Столбец С содержит значения относительной ошибки аппроксимации:

$$A_i = \left| \frac{\left(y - \hat{y}_x\right)}{y} \right| 100, \qquad (3)$$

где A_i — относительная ошибка аппроксимации; $y - \hat{y}_x$ — остаток; y — значение зависимой переменной.

Чтобы произвести расчет в пакете *Excel*, необходимо выделить ячейку C2, вставить функцию ABS, которая возвращает модуль числа. В открывшимся окне функции ABS выполнить деление остатка на значение зависимой величины (B2/A2) и умножить на 100. Нажать *Enter* и протянуть введенную формулу вниз, для расчета всех 672 значений.

а)

ABS =ABS(B2/A2)*100

регрессия 2 *

	A	B	C	D	E
1	F,%	Остатки	A		
2	57	-3,70598	=ABS(B2/A2)*100		
3	49	-8,41637	17,17627		
4	51	-6,41637	12,58112		
5	54	-4,07429	7,544988		
6	58	-0,73221	1,262439		
7	61	1,609864	2,639121		
8	64	3,732636	5,832243		
9	66	5,074714	7,688961		

б)

регрессия 2 *

	A	B	C	D
666	68	7,294021	10,7265	
667	69	7,855407	11,38465	
668	70	8,855407	12,65058	
669	69	8,074714	11,70248	
670	70	9,074714	12,96388	
671	69	8,513328	12,33816	
672	71	10,07471	14,18974	
673	71	9,6361	13,57197	
674		Итого	20751,71	
675		Среднее	30,88053	
676				

Рис. 33. Расчёт средней ошибки аппроксимации:
а — ввод формулы; *б* — результат расчета среднего значения

Для выполнения расчета средней ошибки аппроксимации используется следующее выражение:

$$\overline{A} = \frac{1}{n}\sum A_i. \qquad (4)$$

В нижней части таблицы с помощью функций *Excel* Автосумма и Автосумма – Среднее найти соответствующие значения (рис. 33*б*). Значение средней ошибки аппроксимации менее 15 % свидетельствует о хорошо подобранной модели уравнения. В примере значение средней ошибки аппроксимации более 15 % (30,88), что говорит о плохо подобранной модели уравнения регрессии.

6. Из таблицы с регрессионной статистикой выпишите табличное значение *F*-критерия Фишера (рис. 32). Рассчитайте $F_{кр}$ согласно методике, изложенной в разделе 2. Сравните значения $F_{кр}$ и $F_{табл}$. Если $F_{табл} > F_{кр}$, то можно сделать вывод о значимости уравнения регрессии.

7. Оценку статистической значимости параметров регрессии провести с помощью *t*-статистики Стьюдента и путём расчёта доверительного интервала каждого из показателей. В таблице регрессионного анализа

(рис. 32) указаны фактические значения t-статистики ($t_{табл}$). Критическое значение критерия Стьюдента рассчитываем согласно разделу 1. Сравнить значения $t_{табл}$ и $t_{кр}$.

t-критерий для коэффициента корреляции рассчитать по формуле:

$$t_r = \sqrt{F}. \quad (5)$$

Сравнить значения t_r и $t_{кр}$. Если фактические значения t-статистики превосходят критическое значение, то нулевая гипотеза отклоняется, то есть параметры регрессии и коэффициент корреляции не случайно отличаются от нуля, а статистически значимы. Если фактические значения — меньше критического значения, то нулевая гипотеза принимается, параметры регрессии незначимы. Полученные оценки уравнения регрессии позволяют использовать его для прогнозирования.

8. Подготовьте массив данных для расчета авторегрессии первого порядка аналогично таблице 5 лабораторной работы № 6. Обратите внимание, что временной ряд, который будет использоваться для авторегрессионного анализа, сократился на одну позицию (рис. 34).

	A	B	C	D	E	F	G	H
1	N	date	time	T,C	y(t-1)	Y		
2	1	1 05 2012	0:00:30	6				
3	2	1 05 2012	0:15:30	7,5	6	=0,051799+0,993598*E3		
4	3	1 05 2012	0:30:30	7,5	7,5	7,503784		
5	4	1 05 2012	0:45:31	7,2	7,5	7,503784		
6	5	1 05 2012	1:00:30	6,9	7,2	7,205705		
7	6	1 05 2012	1:15:31	6,6	6,9	6,907625		
8	7	1 05 2012	1:30:30	6,2	6,6	6,609546		
9	8	1 05 2012	1:45:30	5,9	6,2	6,212107		
10	9	1 05 2012	2:00:30	5,7	5,9	5,914027		
11	10	1 05 2012	2:15:30	5,7	5,7	5,715308		

Рис. 34. Формирование массива данных для авторегрессионного анализа

9. Постройте и проанализируйте поле автокорреляции согласно пункту 1 выполнения данной работы.

10. Выберите Сервис – Анализ данных – Регрессия и заполните диалоговое окно (рис. 31). Не ставьте галочки у пунктов Остатки, Стандартизированные остатки и График нормальной вероятности. Результаты расчетов (рис. 35) пакет анализа выдает нам на новом листе (если в настройках не было указано иначе). Обратите внимание на значение коэффициента корреляции ($r^2 = 0{,}99$).

Из выделенных на рис. 35 данных соберите модель, подставляя в уравнение общего вида рассчитанные коэффициенты:

$$y_t = 0{,}051799 + 0{,}993598 \cdot y_{t-1}. \qquad (6)$$

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,993504							
5	R-квадрат	0,987049							
6	Нормированный R-квадрат	0,98703							
7	Стандартная ошибка	0,376653							
8	Наблюдения	671							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	7233,66208	7233,66208	50988,7752	0			
13	Остаток	669	94,90951512	0,141867736					
14	Итого	670	7328,571595						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	0,051799	0,038739638	1,337109464	0,181641369	-0,024266772	0,127865046	-0,024266772	0,127865046
18	Переменная X 1	0,993598	0,004400208	225,8069423	0	0,984957676	1,002237437	0,984957676	1,002237437
19									

Рис. 35. Результат расчета для авторегрессионного анализа

Расчет величины y_t введите в столбец F (рис. 34), затем протяните до конца массива данных. Введите в ячейку G3 массива данных формулу для нахождения значения отклонения $\varepsilon = y(t) - y_t$ и протяните до конца массива данных. С помощью функции Автосумма – Среднее найдите среднее отклонение.

11. Постройте на одном графике исходный и полученный по уравнению регрессии временные ряды (рис. 36). Проанализируйте полученные данные.

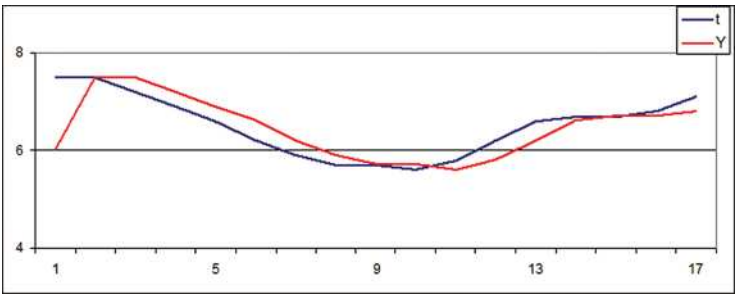


Рис. 36. Сравнение исходного и рассчитанного рядов с помощью графика

12. Аналогично пункту 8 подготовьте массив данных для расчета авторегрессии второго порядка. Эта модель отличается от первой тем, что включает в себя еще один влияющий фактор y_{i-2}, т. е. показывается зависимость от того каким было t не только один период назад, но и от того каким t было два периода назад. Порой это позволяет выявить большую взаимосвязь и соответственно построить более точный прогноз. Теперь

временной ряд, который будет использоваться для авторегрессионного анализа, сократился на две позиции.

Аналогично пункту 10 получите результаты расчета авторегрессии. Обратите внимание, что при заполнении Входной интервал X окна Регрессия (рис. 37) необходимо вводить данные столбцов E и F (y_{i-1} и y_{i-2}).

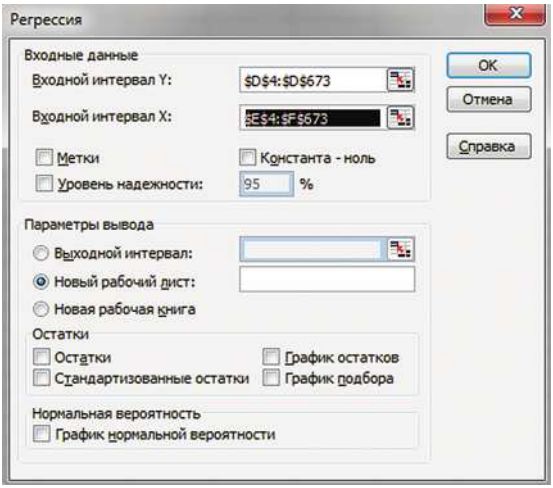


Рис. 37. Заполнение окна «Регрессия» для авторегрессии второго порядка

Соберите и рассчитайте модель авторегрессии второго порядка по формуле:

$$y_t = a_0 + a_1 \cdot y_{t-1} + a_2 \cdot y_{t-2}, \qquad (7)$$

где a_0 — значение «Y-пересечение» таблицы регрессионной статистики; a_1 и a_2 — значения переменных X_1 и X_2 соответственно.

Введите значения y_t и ε в таблицу массива данных, найдите среднее отклонение.

Обратите внимание на значение коэффициента корреляции (r^2), сравните его со значением, полученным при расчете авторегрессии первого порядка. Если показатель r^2 ниже чем у модели первого порядка и среднее отклонение больше чем у модели первого порядка, то модель первого порядка более точная.

13. Сравните реальные данные с рассчитанными с помощью графика.

ЗАКЛЮЧЕНИЕ

В заключение отметим, что в данном учебном пособии рассмотрены лишь наиболее простые (базовые) подходы к статистическому анализу временных рядов метеорологических величин. Однако авторы полагают, что овладение практическими навыками использования базовых методов будет способствовать изучению и использованию огромного арсенала методов и подходов, которыми обладает современная статистика, для исследования временных рядов метеорологических величин различной природы и длительности.

Статистический анализ, несмотря на наличие в нем большого числа количественных характеристик и методов их определения, это все еще во многом искусство, прекрасная область для раскрытия творческого потенциала исследователя. И пусть овладение простыми базовыми методами статистического анализа временных рядов метеорологических величин будет тем шагом, который в дальнейшем будет способствовать овладению этим искусством.

Если факторы, влиявшие на их формирование в прошлом и влияющие в настоящем, будут действовать и в будущем, то анализ временных рядов представляет собой эффективное средство прогнозирования и управления. Однако критики классических методов, основанных на анализе временных рядов, утверждают, что эти методы слишком наивны и примитивны. Иначе говоря, математическая модель, учитывающая факторы, действовавшие в прошлом, не должна механически экстраполировать тренды в будущее без учета физических моделей и экспертных оценок. Поэтому в последние годы специалисты разрабатывали сложные компьютерные модели прогнозирования, основанные на использовании временных рядов. Исходя из этого, методы анализа временных рядов представляют собой превосходный инструмент прогнозирования (как краткосрочного, так и долгосрочного), если они применяются правильно, в сочетании с другими методами прогнозирования.

ЛИТЕРАТУРА

1. *Васильев А.В., Мельникова И.Н.* Методы прикладного анализа натурных измерений в окружающей среде. — СПб.: Балт. гос. техн. ун-т., 2009. — 369 с.
2. *Бекряев В.И.* Практикум по основам теории эксперимента. — СПб.: РГГМУ, 2003 — 79 с.
3. *Гордеева С.М.* Практикум по дисциплине «Статистическая обработка гидрометеорологической информации». — СПб.: РГГМУ, 2010. — 74 с.
4. *Дьяконов В.П.* Справочник по алгоритмам и программам на языке бейсик для персональных ЭВМ. — М.: «Наука», 1987. — 246 с.
5. *Левин Д.М., Стефан Д., Кребиль Т.С., Беренсон М.Л.* Статистика для менеджеров с использованием Microsoft Excel. 4-е изд. — М.: Издательский дом «Вильямс», 2004. — 1312 с.
6. *Малинин В.Н.* Статистические методы анализа гидрометеорологической информации. — СПб.: РГГМУ, 2008. — 407 с.
7. *Мишулина О.А.* Статистический анализ и обработка временных рядов. – М.: МИФИ, 2004. — 180 с.
8. STATISTICA. Искусство анализа данных на компьютере. 2-е издание. — М.: Питер, 2003. — 631 с.
9. Ефременко Д.С. — Электронный ресурс: [http://aiismeteo.rshu.ru].
10. Ефременко Д.С. — Электронный ресурс: [http://www.fier867.0fees.net/iram/div.html].
11. Чукин В.В. — Электронный ресурс: [http://meteolab.rshu.ru:8080].

ПРИЛОЖЕНИЯ

Приложение 1

Примеры статистического анализа временных рядов, полученных с помощью автоматических метеорологических станций

Приведены примеры итоговых таблиц статистической обработки временных рядов, полученных с помощью автоматических метеорологических станций.

Пример 1. Результаты расчета доверительного интервала математического ожидания временных рядов приземной температуры в мае 2012 г.

Длительность ряда	14 дней	7 дней	1 день
Период / Критерий	с 3 по 17	с 3 по 10	3 мая
$t_{кр}$	1,96	1,96	1,98
Δx	1,37	0,81	0,25
$(\bar{x} - \Delta x) < X_{ср} < (\bar{x} + \Delta x)$	$10{,}93 < X_{ср} < 13{,}68$	$8{,}96 < X_{ср} < 10{,}59$	$5{,}99 < X_{ср} < 6{,}49$

Пример 2. Результаты расчета доверительного интервала для дисперсии D временных рядов приземной температуры.

Длительность ряда	14 дней	7 дней	1 день
Период / Критерий	с 3 по 17	с 3 по 10	3 мая
χ^2	1,64	1,64	1,64
χ_1^2	1492,23	732,37	118,75
χ_2^2	1310,01	611,90	73,52
ΔD_1	0,97	0,92	0,81
ΔD_2	1,10	1,10	1,31
$D \cdot \Delta D_1 < D < D \cdot \Delta D_2$	$25{,}6 < D < 29{,}1$	$9{,}8 < D < 11{,}8$	$0{,}99 < D < 1{,}6$

Пример 3. Значения коэффициента автокорреляции временных рядов приземной температуры воздуха.

Шаг временного сдвига ($\Delta t = 15$ мин)	Коэффициент автокорреляции		
	1 день $N = 85$	7 дней $N = 660$	14 дней $N = 1430$
1	0,96	0,99	1
2	0,92	0,98	0,99
3	0,89	0,97	0,99
4	0,86	0,96	0,98
5	0,81	0,95	0,98
6	0,77	0,93	0,97
7	0,73	0,91	0,96
8	0,69	0,89	0,95
9	0,64	0,87	0,94
10	0,61	0,84	0,93

Пример 4. Значения межрядовых коэффициентов корреляции: температуры воздуха T и относительной влажности f; температуры воздуха T и атмосферного давления P.

Шаг временного сдвига ($\Delta t = 15$ мин)	Коэффициент корреляции T и f		
	1 день $N = 85$	7 дней $N = 660$	14 дней $N = 1430$
0	−0,90	−0,61	−0,46
1	−0,91	−0,61	−0,45
2	−0,92	−0,60	−0,45
3	−0,90	−0,59	−0,44
4	−0,88	−0,58	−0,43
5	−0,86	−0,57	−0,42
6	−0,82	−0,55	−0,41
7	−0,78	−0,54	−0,40
8	−0,74	−0,52	−0,39
9	−0,70	−0,50	−0,38
10	−0,66	−0,48	−0,36

Шаг временного сдвига ($\Delta t = 15$ мин)	Коэффициент корреляции T и P		
	1 день $N = 85$	7 дней $N = 660$	14 дней $N = 1430$
0	−0,37	0,31	0,08
1	−0,36	0,30	0,08
2	−0,37	0,30	0,07
3	−0,36	0,30	0,06
4	−0,36	0,29	0,05

Шаг временного сдвига ($\Delta t = 15$ мин)	Коэффициент корреляции T и P		
	1 день $N = 85$	7 дней $N = 660$	14 дней $N = 1430$
5	−0,36	0,29	0,05
6	−0,36	0,29	0,04
7	−0,35	0,29	0,03
8	−0,34	0,29	0,03
9	−0,34	0,28	0,02
10	−0,33	0,28	0,01

Приложение 2

Таблица коэффициентов Стьюдента

$N - 1$	$\alpha = 0,95$	$\alpha = 0,99$	$N - 1$	$\alpha = 0,95$	$\alpha = 0,99$
3	3,182	5,841	16	2,12	2,921
4	2,776	4,604	18	2,101	2,878
5	2,571	4,032	20	2,086	2,845
6	2,447	3,707	22	2,074	2,819
7	2,365	3,499	24	2,064	2,797
8	2,306	3,355	26	2,056	2,779
10	2,228	3,169	28	2,048	2,763
12	2,179	3,055	30	2,043	2,75
14	2,145	2,977	∞	1,96	2,576

Приложение 3

Таблица значений *F*-критерия Фишера
(при уровне значимости α = 0,05)

k_1 / k_2	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Приложение 4

Квантиль χ^2-распределения
($k = N - 1$ и $q = 1 - \alpha$, где α — доверительная вероятность)

k	$1 - q/2$			$q/2$		
	0,99	0,95	0,9	0,1	0,05	0,01
1	0,0000157	0,000393	0,0158	2,706	3,841	6,635
2	0,0201	0,103	0,211	4,605	5,991	9,21
3	0,115	0,352	0,584	6,251	7.815	11,345
4	0,297	0,711	1,064	7,779	9,488	13,277
5	0,554	1,145	1,61	9,236	11,07	15,086
6	0,872	1,635	2,204	10,645	12,592	16,812
7	1,239	2,167	2,833	12,017	14,067	18,475
8	1,646	2,733	3,49	13,362	15,507	20,09
9	2,088	3,325	4,168	14,684	16,919	21,666
10	2,558	3,94	4,865	15,987	18,307	23,309
12	3,571	5,226	6,304	18,549	21,026	26,217
14	4,66	6,571	7,79	21,064	23,685	29,141
16	5,812	7,962	9,312	23,542	26,296	32
18	7.015	9,39	10,865	25,989	28,869	34,805
20	8,26	10,851	12,444	28,412	31,41	37,566
30	14,953	18,493	20,599	40,256	43,773	50,892

СОДЕРЖАНИЕ

CONTENTS

УЧЕБНОЕ ИЗДАНИЕ

Карина Левановна Восканян
Анатолий Дмитриевич Кузнецов
Ольга Станиславовна Сероухова

АВТОМАТИЧЕСКИЕ МЕТЕОРОЛОГИЧЕСКИЕ СТАНЦИИ

Часть 2. Цифровая обработка данных автоматических метеорологических станций

Практикум

Редактор: О.С. Крайнова
Компьютерная верстка: Ю.И. Климов

ЛР № 020309 от 30.12.96

Подписано в печать 01.09.15. Формат 60×90 1/16. Гарнитура Newton.
Печать цифровая. Усл. печ. л. 6,25. Тираж 300 экз. Заказ № 454.
РГГМУ, 195196, Санкт-Петербург, Малоохтинский пр. 98.
Отпечатано в ЦОП РГГМУ