



Лекция «Первичная обработка и анализ данных»

Михайлова Елена Георгиевна  
Графеева Наталья Генриховна

Санкт-Петербург  
2019

# Содержание

<b>1</b>	<b>Преобразование данных</b>	<b>2</b>
1.1	Описательная статистика . . . . .	2
1.2	Преобразование данных . . . . .	10
1.3	Нормировка данных . . . . .	15
1.4	Временные ряды . . . . .	24
1.5	Сглаживание временных рядов . . . . .	29
1.6	Определение трендов временных рядов . . . . .	32
1.7	Построение моделей для временных рядов с сезонными составляющими . . . . .	38

# 1 Преобразование данных

## 1.1 Описательная статистика

**Описательная статистика** — это первичная систематизация данных, полученных из различных источников.

Описательная статистика активно используется на этапе разведочного анализа данных, а в некоторых случаях вообще оказывается достаточной для полного анализа данных. Рассмотрим основные виды описательных статистик и их практическое применение.

**Центральная тенденция.** Измерением центральной тенденции (measure of central tendency) называют процесс выбора одного числа, которое наилучшим образом описывает все значения выбранной переменной из набора данных. Оно с одной стороны позволяет получить информацию о распределении значений переменной в сжатой форме, а с другой — ведет к потере информации по сравнению с распределением частот значений переменной. Различают следующие характеристики центральной тенденции:

- среднее значение;
- мода;
- медиана.

**Среднее.** Среднее значение определяют как сумму всех значений переменной, деленную на количество значений. Обозначается  $\bar{x}$  или **Mean**.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x$$

Среднее вычисляется только в числовых шкалах и в дихотомических данных с 0 и 1. Для каждого набора данных имеется только одно среднее.

Рассмотрим пример вычисления среднего для отметок студента. У студента в процессе его обучения в университете были получены следующие отметки:

5, 4, 2, 5, 4, 3, 3, 4, 5, 3, 5, 5, 5, 2, 5

Среднее будет вычислено как сумма всех значений деленная на количество, т.е.:

$$\text{Mean} = \frac{58}{14} \approx 4.14$$

Среднее может также вычисляться и для дихотомических данных. Если два значения переменной представляются 0 и 1, то среднее для таких данных

указывает долю единиц в выборке. Например, для следующих данных:

1, 0, 0, 0, 1, 1, 1, 0, 0, 0

40% значений выборки принимают значение, равное единице:

$$\text{Mean} = \frac{4}{10} = 0.4$$

**Мода.** Мода – это значение переменной, которое встречается чаще других. Обозначается **Mo**. Мода может быть определена на данных любой шкалы. Моде может соответствовать несколько значений. В этом случае говорят про мультимодальное распределение значений переменной. Если ни одно из значений переменной в наборе данных не повторяется, то говорят, что мода отсутствует. На рисунке вы можете видеть пример вычисления моды для отметок студента. Наиболее частая отметка – 5.

Отметка	Частота	$Mo = 5$
5	7	
4	3	
3	3	
2	1	

**Медиана.** Еще одна характеристика центральной тенденции – медиана. Медиана основывается на понятии вариационного ряда.

**Вариационный ряд** – это упорядоченные данные, расположенные в порядке возрастания значения переменной, либо в порядке убывания. Ряд называется вариационным потому, что содержит варианты значений признака. Рассмотрим пример построения вариационных рядов из отметок студентов (исходные данные представлены на рисунке). По нему построим два вариационных ряда. Первый ряд упорядочен по возрастанию, второй – по убыванию.

Теперь можем определить понятие медианы. Медиана (обозначается **Me**) это значение, соответствующее среднему элементу вариационного ряда. Понятие «средний элемент» отличается для четного и нечетного количества значений переменной. Для набора данных из  $n$  значений, если  $n$  нечетно, средний элемент имеет номер  $\frac{n+1}{2}$ , а для четного значения  $n$  медиана находится как среднее арифметическое двух соседних средних элементов с номерами  $\frac{n}{2}$  и  $\frac{n}{2} + 1$ .

Медиана может быть определена для числовых и порядковых данных. Для каждого набора данных имеется только одна медиана. Рассмотрим пример вычисления медианы для отметок студента. Возьмем упорядоченный по возрастанию вариационный ряд с отметками. В нем 14 элементов. Значит, медиана вычисляется как среднее значение 7 и 8 элементов и равна 4.5.

Пример (вычисление медианы для отметок студента)

2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5

14 элементов

Медиана вычисляется как среднее значение 7 и 8 элементов  $(4 + 5)/2$

$Me = 4.5$

Еще один пример – вычисление медианы для показателей силы морского ветра по шкале Бофорта. Есть следующие наблюдения о силе морского ветра по шкале Бофорта, которые вы видите на экране. Сформируем на их основе вариационный ряд по возрастанию значения переменной. Количество элементов – 13. Значит 7-й по порядку элемент и есть медиана.



Пример (вычисление медианы для показателей силы морского ветра по шкале Бофорта)

0, 2, 2, 1, 1, 3, 3, 1, 1, 0, 0, 1, 2

Вариационный ряд по возрастанию

0, 0, 0, 1, 1, 1, ① 1, 2, 2, 2, 3, 3

$Me = 1$  – тихий ветер

		
0 баллов штиль	4 балла умеренный ветер	8 баллов очень крепкий ветер
1 балл тихий ветер	5 баллов свежий ветер	9 баллов шторм
2 балла лёгкий ветер	6 баллов сильный ветер	10 баллов сильный шторм
3 балла слабый ветер	7 баллов крепкий ветер	11 баллов жестокий шторм
		12 баллов ураган 

Подведем итоги. Мы рассмотрели 3 характеристики центральной тенденции. В следующей таблице указано какие характеристики могут быть применимы к тем или иным шкалам. Какая из этих характеристик лучше? Какую

Характеристики центральной тенденции	Номинальные данные	Порядковые данные	Интервальные данные	Относительные данные
Мода	✓	✓	✓	✓
Медиана		✓	✓	✓
Среднее			✓	✓

из них применять, если есть возможность выбора? На первый взгляд может показаться, что среднее – наиболее емкая, широко известная и применяемая на практике характеристика. В плане известности, несомненно, да, но в плане

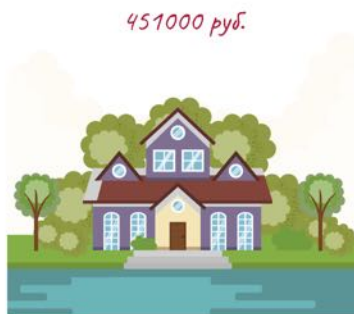
полезности применения – далеко не всегда. Приведем хорошо известный пример на эту тему.

В некоторой деревне проживает 50 жителей. Среди них 49 человек – сельские жители с месячным доходом 1 тысяча рублей, а один житель – зажиточный фермер с доходом 451 тыс. рублей. Вычислим средний доход жителей деревни. Он равен 10 тыс. рублей. Совершенно очевидно, что это число не отражает адекватно доход жителей деревни. В этом случае гораздо более рационально было бы использовать в качестве меры центральной тенденции моду или медиану (обе равны 1 тыс. рублей). Кроме того, в этом случае одного числа явно не хватает для описания доходов жителей в этой деревне.

*Пример*

*Доходы жителей*

1000 руб. – 49  
451000 руб. – 1



451000 руб.

✗  $Mean = 10000 \text{ руб.}$



1000 руб.

✓  $Mo = 1000 \text{ руб.}$

✓  $Me = 1000 \text{ руб.}$

## Размах и квартильный размах

**Мера центральной тенденции** – всего лишь одно число, которое используется для описания типового значения из изучаемой выборки. Оно не дает представления о том, насколько разнообразны данные в выборке. Именно поэтому было придумано понятие размаха и квартильного (или межквартильного) размаха.

**Размах (Range).** Размах – разность между наибольшим и наименьшим значениями набора данных. Для набора данных, представляющих отметки студента, который вы видите на экране, размах равен 3.

*Размах – разность между наибольшим и наименьшим значениями набора данных.*

$$R = x_{\max} - x_{\min}$$

*Пример (вычисление размаха для отметок студента)*

*5, 4, 2, 5, 4, 3, 3, 4, 5, 3, 5, 5, 2, 5*

$$R = 5 - 2 = 3$$



**Квартили (Quartile).** Следующая характеристика разброса данных в выборке – квартильный (или межквартильный) размах. Он основывается на понятии квартилей. Под квартилями понимаются значения  $Q_1, Q_2, Q_3$  которые делят вариационный ряд на четыре равные части.

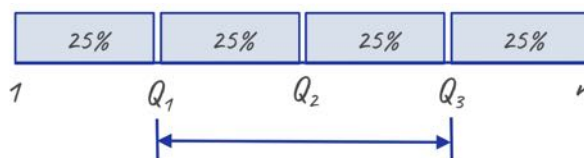
*Под квартилями понимаются значения  $Q_1, Q_2, Q_3$ , которые делят вариационный ряд на четыре равные части.*



Второй квартиль  $Q_2$  совпадает с медианой.  $Q_1$  – это медиана для значений, которые меньше  $Q_2$ .  $Q_3$  – это медиана для значений, которые больше  $Q_2$ . Существует несколько вариантов точного определения значения квартилей, которые могут немного отличаться. Например, при определении первого  $Q_1$  и третьего квартилей  $Q_3$  можно включать или, наоборот, исключать медиану (то есть слова меньше/больше понимать как строго больше/строго меньше или не строго больше/не строго меньше). Именно поэтому в большинстве современных инструментов есть две версии функции квартиль: с исключением и включением медианы при определении первого и третьего квартилей. Называются такие функции: QUARTILE.EXC и QUARTILE.INC.

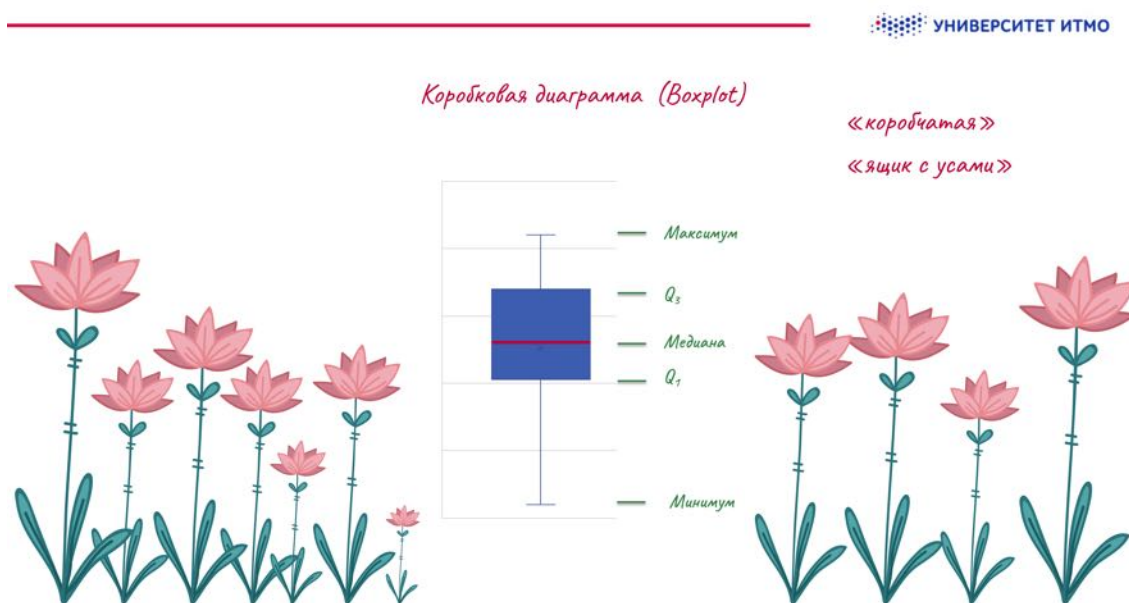
**Размах квартилей (Inter Quartile Range).** Размах квартилей – это разница между третьим и первым квартилем и вычисляется по формуле:

$$IQR = Q_3 - Q_1$$



В чем принципиальная разница между размахом и квартильным размахом? Размах – очень простая и «грубая» мера вариации, т.к. при вычислении размаха используются только наименьшее и наибольшее значения переменной. При вычислении квартильного размаха игнорируются только крайние значения, расположенные за пределами первого и третьего квартилей. Между третьим и первым квартилем оказываются 50% всех данных.

**Коробковая диаграмма (BoxPlot).** При проведении разведочного анализа очень полезной оказывается, так называемая, коробковая диаграмма. Она имеет вид, изображенный на рисунке, и может быть нарисована как



в горизонтальном, так и в вертикальном виде. На ней отображены минимум, максимум и три квартиля. Это позволяет очень емко и выразительно отобразить основные значения данных.

Квартили могут использоваться для определения, так называемых, выбросов, т.е. значений, которые слишком отличаются от остальных. Классифицируют два вида выбросов: **умеренные** и **экстремальные**. Умеренными



называют выбросы, которые удалены ниже первой квартили или выше третьей от  $1.5IQR$ , но не более, чем на  $3IQR$ . Экстремальные выбросы удалены ниже первой квартили или выше третьей более, чем на  $3IQR$ . Схема определения выбросов приведена на экране. Определение выбросов крайне важно на этапе подготовки данных и позволяет избавиться значений, достоверность которых сомнительна.

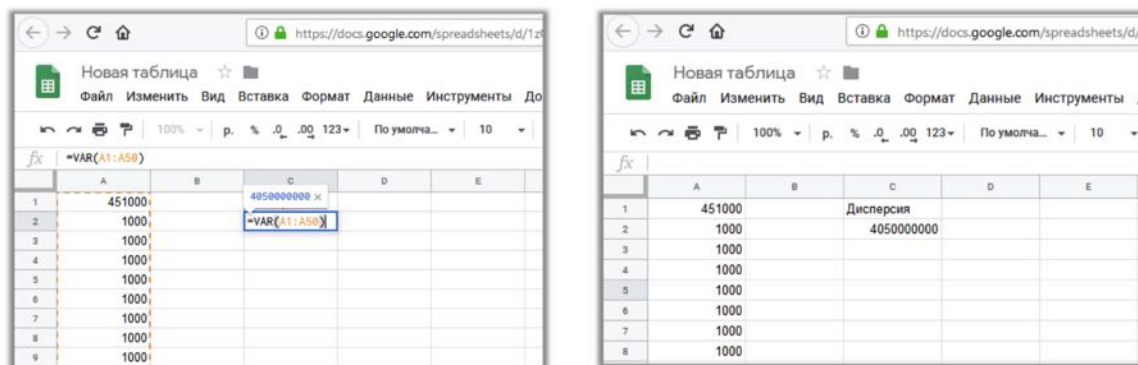


Кроме того, на основании этих данных рисуется, так называемая, коробочная диаграмма с расширением, на которой отображены выбросы. Рассчитывается она в два этапа: на первом определяются квартили и по ним – выбросы (они будут отображены на диаграмме в виде точек), а затем из данных исключаются выбросы и заново пересчитываются минимум, максимум и квартили и отображаются в виде обычной коробочной диаграммы. Пример такой коробочной диаграммы с расширением вы можете видеть на экране.

**Дисперсия.** Еще одно очень полезное статистическое понятие – дисперсия. Дисперсия для набора данных или выборки – это среднее арифметическое квадратов отклонений значений от их среднего:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Дисперсия для набора данных или выборки – это среднее арифметическое квадратов отклонений значений от их среднего.



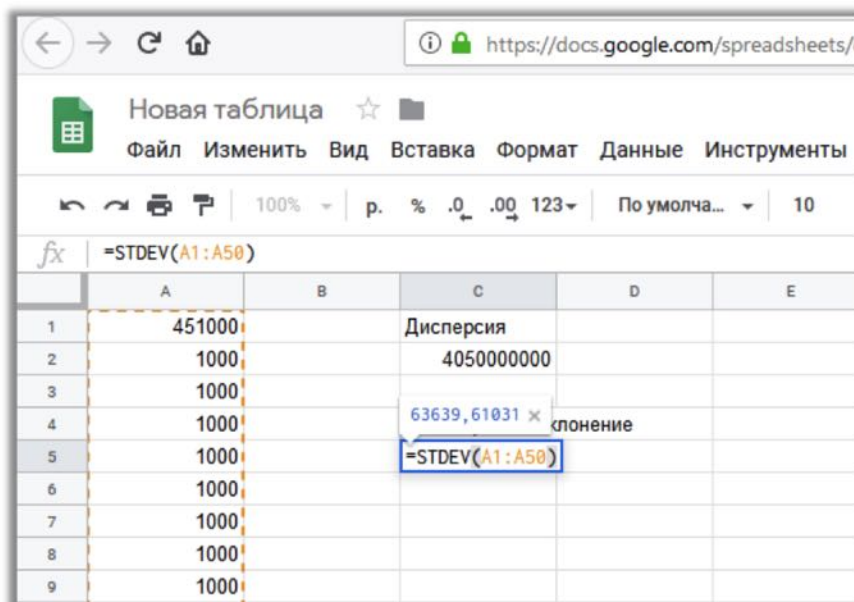
Значение дисперсии может быть вычислено явно по упомянутой выше формуле или с помощью любого подходящего инструментария, в котором

такая функция присутствует. В Google таблицах такая функция присутствует и называется VAR.

**Стандартное отклонение.** С понятием дисперсии тесно связана еще одна описательная статистика – стандартное отклонение. Стандартное отклонение – это квадратный корень из дисперсии выборки:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Стандартное отклонение – это квадратный корень из дисперсии выборки. Вычисляется по приведенной на экране формуле. Присутствует практически во всех инструментах. В Google таблицах ему соответствует функция STDEV.



The screenshot shows a Google Sheets interface with a spreadsheet titled "Новая таблица". The formula bar at the top displays `=STDEV(A1:A50)`. The spreadsheet has columns A through E and rows 1 through 9. Column A contains the values 451000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, and 1000. Column B is empty. Column C contains the text "Дисперсия" in row 1, the value 4050000000 in row 2, and a tooltip for the standard deviation value 63639,61031 in row 4. The tooltip also includes the text "клонение" and the formula `=STDEV(A1:A50)`. The formula bar at the bottom of the spreadsheet also displays `=STDEV(A1:A50)`.

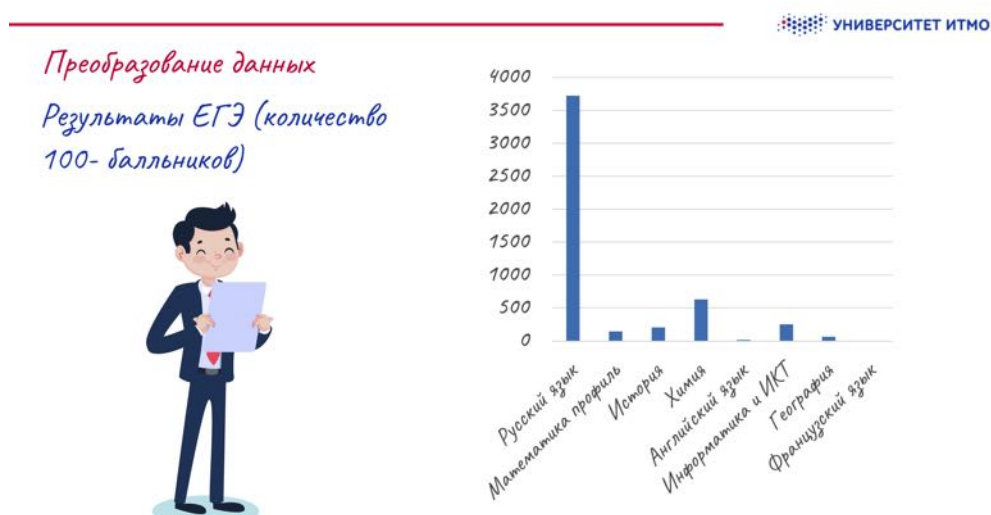
	A	B	C	D	E
1	451000		Дисперсия		
2	1000		4050000000		
3	1000				
4	1000		63639,61031 × клонение =STDEV(A1:A50)		
5	1000				
6	1000				
7	1000				
8	1000				
9	1000				

## 1.2 Преобразование данных

Преобразование данных – одна из распространенных процедур предварительной обработки данных, способная продемонстрировать характерные особенности, скрытые в данных и не видимые в их первоначальной форме.



Попытаемся аргументировать необходимость проведения преобразований на конкретном примере. В нашем распоряжении есть агрегированные данные о результатах ЕГЭ за 2018 год. Нет никакого сомнения, что они достоверны. Это данные из официальных релизов Рособрнадзора.

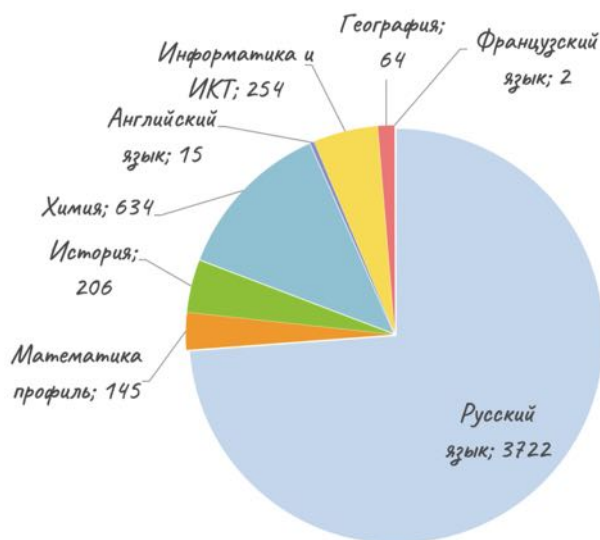


В таблице слишком много чисел, которые трудно охватить и сделать общие выводы. Попытаемся отобразить в виде простейших графиков количество 100-балльников по различным предметам. Данные по своему содержанию соответствуют агрегированным значениям из различных категорий.

Значит им подойдет визуализация в виде столбчатой или круговой диаграммы. Вот как выглядят наши данные на столбчатом графике.

Еще раз подчеркнем, что данные достоверны. Это агрегированные данные, которые ровно такие, как они есть. Но совершенно очевидно, что визуализировать их таким образом нельзя, так как некоторые значения (например, соответствующие категориям Английский язык, География и Французский язык) отображаются неадекватно – они почти не видны на графике.

Попробуем отобразить эти же данные в виде круговой диаграммы. Воз-



можно, стало немного лучше. География стала выглядеть более убедительно. Однако Английский и Французский языки, по-прежнему, практически не видны на диаграмме. Причина, по которой указанные значения не видны – большой разброс значений среди агрегированных данных. Что делать с такими данными и как их визуализировать? Возможное решение – преобразование данных таким образом, чтобы разброс значений уменьшился, а сами данные стали, как минимум соизмеримыми. Существует много различных методов преобразования. Рассмотрим наиболее традиционные и обсудим, как эта трансформация влияет на визуализацию.

## Распространенные преобразования

В таблице на ниже приведены наиболее распространенные способы преобразования и особенности их использования.

Так, например, логарифм натуральный или десятичный хорошо подходит для преобразования данных, сохраняет порядок среди значений, но не уместен при наличии нулевых значений в исходных данных. Преобразование квадратного корня также сохраняет порядок между значениями, уместно при нулевых значениях, но не уместно при наличии отрицательных. Преобразова-

<i>Преобразование</i>	<i>Не подходит для</i>
$\ln(x)$	нулевых значений
$\log_{10}(x)$	нулевых значений
$\sqrt{x}$	отрицательных значений
$x^2$	отрицательных значений
$1/x$	нулевых значений

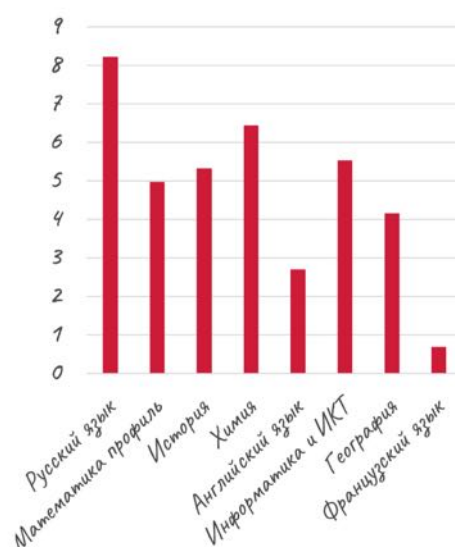
ние по формуле обратной дроби меняет порядок значений (что иногда может оказаться уместным), но не умеет работать с нулевыми значениями.

**Логарифмическое преобразование.** Рассмотрим, как применяются преобразования на примере логарифмического преобразования. Чтобы осуществить логарифмическое преобразование, необходимо вычислить логарифм каждого значения в наборе данных и использовать эти преобразованные данные вместо исходных. Логарифмические преобразования оказывают

<i>Предмет</i>	<i>Количество 100-балльников</i>	<i><math>\ln(\text{количество})</math> 100-балльников)</i>
Русский язык	3722	8.22
Математика профиль	145	4.98
История	206	5.33
Химия	634	6.45
Английский язык	15	2.71
Информатика и ИКТ	254	5.54
География	64	4.16
Французский язык	2	0.69

существенный эффект на форму распределения. На рисунке представлена столбчатая диаграмма о 100-балльниках ЕГЭ после применения натурального логарифмического преобразования.

$\ln(\text{количество } 100\text{-балльников})$

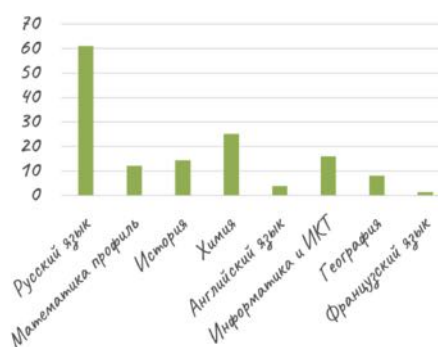


**Преобразование квадратного корня.** Преобразование квадратного корня оказывает более умеренный эффект на форму распределения.

Следующий график показывает столбчатую диаграмму о 100-балльниках ЕГЭ после применения преобразования квадратного корня.

Предмет	Количество 100-балльников	$\sqrt{\text{количество}}100\text{-балльников}$
Русский язык	3722	61.01
Математика профиль	145	12.04
История	206	14.35
Химия	634	25.18
Английский язык	15	3.87
Информатика и ИКТ	254	15.94
География	64	8.00
Французский язык	2	1.41

$\sqrt{\text{количество } 100\text{-балльников}}$

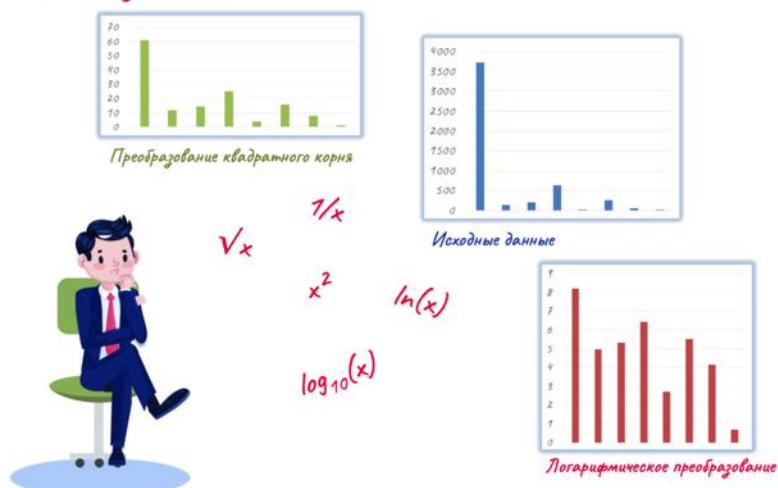


## Выбор подходящего преобразования

Возможных преобразований много. Как выбрать подходящее преобразование? Ответ на этот вопрос не очевиден, хотя формальные статистические методы для выбора преобразования существуют. Если не вникать в эти теории, то общая стратегия выбора преобразования заключается в том, чтобы четко определить цель преобразования (например, визуализация определенного типа, сохранение или, наоборот, разворот упорядоченности), а затем применить наиболее используемые преобразования, такие как логарифмы,

квадратный корень, квадрат, обратная дробь и выбрать лучший метод, исходя из цели и полученных результатов.

### Выбор подходящего преобразования





## Единицы измерения и обратные преобразования

Поскольку методы преобразования включают в себя применение к исходным данным математических функций, то необходимо обратить внимание на изменение единиц измерения данных. Например, при применении логарифмической функции к переменной численности 100-балльников, единицей измерения становится логарифм численности. Значит, при представлении дан-

	<i>Средняя отметка за период обучения</i>	<i>Количество грамот за участие в художественных конкурсах</i>	<i>Количество грамот за участие в интеллектуальных конкурсах</i>	<i>Количество грамот за участие в спортивных мероприятиях</i>
<i>Иван</i>	4.5	5	2	8
<i>Александра</i>	4.9	0	4	4
<i>Семен</i>	4	2	3	10
<i>Екатерина</i>	4.5	5	3	0
<i>Иннокентий</i>	4.2	2	6	9
<i>Анна</i>	5	2	5	0

ных на графиках и диаграммах надо явно указывать, какие именно преобразования были проведены, и в каких единицах измерения отображены данные. Если преобразованные данные использовались для вычисления статистик, то надо не забыть провести обратное преобразование, чтобы представить результат в начальных единицах измерения. Например, если было применено преобразование квадратного корня, необходимо совершить обратное преобразование, и возвести конечный результат в квадрат.

### 1.3 Нормировка данных

Нормировка данных – еще одна процедура возможной предварительной обработки данных. Назначение нормировки – обеспечение возможности для сравнения, агрегации и, возможно, визуализации значений нескольких переменных из различных шкал. Для некоторых алгоритмов машинного обучения (и не только) нормировка переменных является необходимым условием.

Попытаемся аргументировать целесообразность нормировки на очень простом конкретном примере. В распоряжении педагогического совета школы есть одна путевка в Артек и сведения об учащих, представленные на слайде.

Будем условно считать, что все грамоты приблизительно одного порядка и среди них нет каких-то невероятно выдающихся. Перед педсоветом стоит задача найти (а в дальнейшем опубликовать) формальный критерий, по



которому будет отобран претендент на единственную путевку в Артек. Как найти этот критерий? Задача состоит в том, чтобы каждому школьнику сопоставить одно число, которое будет представлять все его достижения, а затем на основании этих чисел составить рейтинг школьников и выявить первого претендента. Если бы все переменные и отметки измерялись в одинаковых шкалах и единицах измерения, можно было бы предложить сложить все значения, но это очень грубый подход, так как спортивные грамоты выдаются гораздо чаще других, и спортивные достижения тут же перекроют все остальные. Выход – нормировка значений переменных, а затем вычисление на их основе итогового критерия.

**Почему нужна нормировка показателей?** Обычно выраженность некоторого качества описывают числом. Переменная  $x$  меняется от некоторого минимального значения  $x_{min}$  (отражающего отсутствие качества) до некоторого максимального значения  $x_{max}$  (высшая степень проявления качества). Наличие критерия качества позволяет решать проблему сравнения двух объектов, но только по этому показателю. Но при этом надо помнить, в каких пределах меняется показатель. А диапазоны разброса значений и единицы измерений для разных переменных — самые разнообразные... Кроме того иногда необходимо оценивать, насколько близко конкретное значение к краям диапазона или к его середине. Если же речь идет о сравнении или агрегировании по различным показателям — дело обстоит совсем плохо. А ведь именно показатель качества интерпретируется как **степень выраженности** качества. А степени выраженности сравнивать и агрегировать можно и нужно! Но для этого показатели следует привести к одной шкале так, чтобы минимальное и максимальное значения для различных переменных совпадали. Такое преобразование и называется **нормировкой**. После этого преобразования можно сравнивать и агрегировать разнообразные показатели, полученные различными методиками.

## Классы числовых показателей

При всем разнообразии числовых характеристик объектов из них можно выделить два широких класса:

- **униполярные**, выражающие только степень наличия некоторого качества или количества (например, интенсивный цвет, очень хорошая отметка или количество чего-либо);
- **биполярные**, отражающие не только степень наличия качества, но и его «направленность».

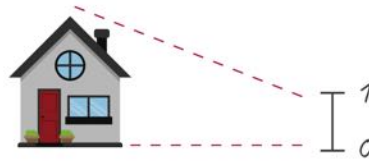
Методы нормировки различаются для этих классов. Рассмотрим последовательно некоторые из них.

## Нормировка униполярных показателей

Обычно униполярные показатели нормируются в диапазоне от 0 до 1. В качестве функции нормировки может выступать любая непрерывная возрастающая функция  $y = f(x)$  с минимальным значением – 0 и максимальным значением – 1.

*Функция нормировки для униполярного показателя*

$$\begin{aligned} y(x_{\min}) &= 0; & \frac{dy}{dx} &> 0 \\ y(x_{\max}) &= 1; \end{aligned}$$



Рассмотрим возможные варианты такой функции, безусловно, обладающих упомянутыми выше свойствами. Существует два возможных варианта нормировки, **экспоненциальная**

$$y(x) = 1 - \exp\left(1 - \frac{x}{x_{\min}}\right)$$

или **линейная**

$$y(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

Надо заметить, что линейная функция в силу своей простоты используется чаще других. Достоинством экспоненциальной функции считается то, что она равномернее распределяет исходные значения по диапазону от 0 до 1. И более того, небольшие модификации этой формулы с легкостью позволяют усилить эту равномерность распределения в конкретных случаях.

## Нормировка биполярных показателей

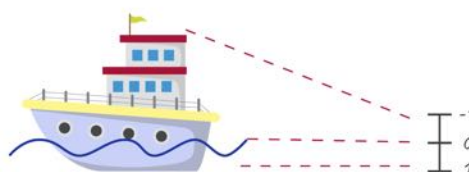
Биполярные показатели обычно нормируются в диапазоне от -1 до 1. В качестве функции нормировки может выступать любая непрерывная возрастающая функция  $y = f(x)$  с минимальным значением минус 1 и максимальным значением плюс 1. Такой линейной функции, основанной на минимальных и максимальных значениях, может быть:

$$y(x) = \frac{2x - (x_{\max} + x_{\min})}{x_{\max} - x_{\min}}.$$

Разумеется, есть и другие возможные варианты нормировки и некоторые из них не являются линейными и порою привязаны к специфике предметной области, в которой проводится нормировка показателей, тем не менее, в большинстве случаев такое преобразование является вполне достаточным для последующего анализа. Теперь, вооружившись полученными знаниями,

*Функция нормировки для биполярного показателя*

$$\begin{aligned} y(x_{\min}) &= -1; & \frac{dy}{dx} &> 0 \\ y(x_{\max}) &= 1; \end{aligned}$$



вернемся к нашему примеру со школьниками.

**Какие показатели у учащихся?**

	Униполярные показатели (отражают количественные успехи, представлены в виде положительных целых чисел)			
	Униполярный показатель (отражает качество успехов в обучении, измеряется от 1 до 5)	Количество грамот за участие в художественных конкурсах	Количество грамот за участие в интеллектуальных конкурсах	Количество грамот за участие в спортивных мероприятиях
Иван	4.5	5	2	8
Александра	4.9	0	4	4
Семен	4	2	3	10
Екатерина	4.5	5	3	0
Иннокентий	4.2	2	6	9
Анна	5	2	5	0

Рассмотрим все показатели нашего примера с учащимися. Все они – униполярные. Действительно, средняя отметка – униполярный показатель, который отражает однонаправленное качество успехов в обучении, как правило, измеряется от 1 до 5. Количество грамот за участие в художественных конкурсах – униполярный показатель (представлен в виде положительного целого). Количество грамот за спортивные достижения – униполярный показатель (представлен в виде положительного целого). Количество грамот за участие в интеллектуальных конкурсах – униполярный показатель (представлен в виде положительного целого). Это означает, что при нормировке данных мы можем воспользоваться любой из нормировок для униполярных

### Пример нормировки показателя

$x_{\min} = 0;$

$x_{\max} = 5;$

	Количество грамот за участие в художественных конкурсах	
	до нормировки	после нормировки
Иван	5	1.00
Александра	0	0.00
Семен	2	0.40
Екатерина	5	1.00
Иннокентий	2	0.40
Анна	2	0.40

показателей. Для простоты возьмем линейную нормировку. Рассмотрим, к примеру, как будет выполнена нормировка показателя «Количество грамот за участие в художественных конкурсах». Для начала определим минимальные и максимальные значения для этого показателя. Они равны соответственно – 0 и 5. Подставим эти значения в формулу линейной нормировки значения для показателя. Результаты вы можете видеть на экране.

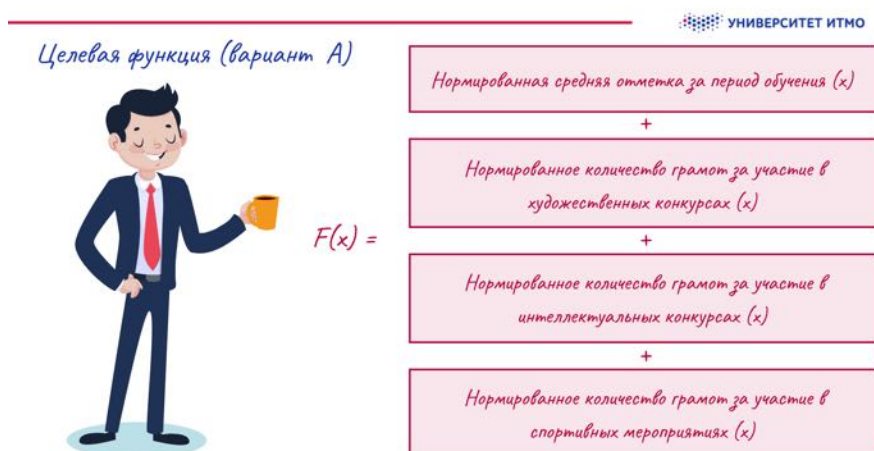
Воспользуемся формулой униполярного линейного преобразования для всех показателей и нормируем все исходные данные (причем нормировать надо каждый показатель по отдельности). Итак, у нас есть нормированные значения по каждому показателю. Что дальше?

### Показатели учащихся после нормировки

	Средняя отметка за период обучения	Количество грамот за участие в художественных конкурсах	Количество грамот за участие в интеллектуальных конкурсах	Количество грамот за участие в спортивных мероприятиях
Иван	0.50	1.00	0.00	0.80
Александра	0.90	0.00	0.50	0.40
Семен	0.00	0.40	0.25	1.00
Екатерина	0.50	1.00	0.25	0.00
Иннокентий	0.20	0.40	1.00	0.90
Анна	1.00	0.40	0.75	0.00

Теперь необходимо задать, так называемую, целевую функцию на основе нормированных значений (соответствующих ученикам). Что из себя представляет целевая функция? Это математическое выражение некоторого критерия качества объекта (процесса, решения). Целевая функция задается для того, чтобы вместо большого количества качественных параметров для каждого изучаемого объекта получить один, а затем на основе максимального (или минимального) значения функции определить объект, на котором достигается соответствующий экстремум. Так какое же значение функции нужно

использовать? Максимум? Минимум? Что именно – зависит от специфики поставленной задачи и вида самой функции.



Например, если эта функция отражает суммарные положительные качества ученика, то это, наверняка, максимум. А если это суммарная стоимость затрат для выполнения какой-то задачи, то логичнее использовать минимум. В нашем случае в качестве целевой функции можно использовать сумму нормированных значений, так как каждое из значений отражает какие-то положительные качественные характеристики ученика, а лучшим значением считать максимум такой функции. Ученик, для которого функция выдаст максимальное значение, будет считаться лучшим. В нашем распоряжении есть, как минимум, два возможных варианта, чтобы увидеть результат вычисления такой функции: мы можем добавить к таблице еще один столбец, в котором будет вычислена сумма нормированных показателей и найти учащегося с максимальным значением целевой функции или использовать замечательное средство для визуализации – гистограмму с накоплением. Этот тип диаграмм есть в большинстве популярных инструментов визуализации. Особенностью этой диаграммы является то, что она сама суммирует показатели и наша задача их всего лишь правильно задать и найти столбец с максимальным накопленным значением. Продемонстрируем оба варианта.

Итак, на рисунке ниже представлена таблица, в которой в качестве последнего столбца добавлено значение целевой функции (вариант А – сумма всех нормированных показателей). Легко видеть, что максимальное значение целевой функции соответствует Иннокентию. Значит Иннокентий – победитель! Рассмотрим второй возможный вариант определения победителя – гистограмму с накоплением.

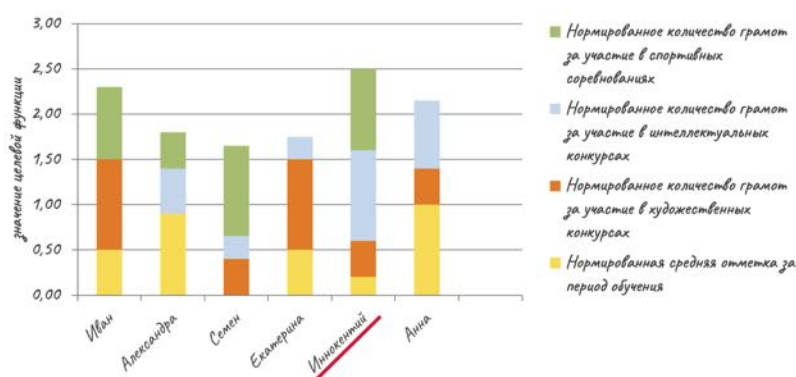
Построим гистограмму с накоплением на основе нормированных значений и увидим абсолютно аналогичный результат. Именно Иннокентию соответствует столбец с накопленным значением максимальной высоты. Все, о

Таблица с явно вычисленной целевой функцией

	Нормированная средняя отметка за период обучения	Нормированное количество грамот за участие в художественных конкурсах	Нормированное количество грамот за участие в интеллектуальных конкурсах	Нормированное количество грамот за участие в спортивных мероприятиях	Значение целевой функции (вариант А)
Иван	0.50	1.00	0.00	0.80	2.30
Александра	0.90	0.00	0.50	0.40	1.80
Семен	0.00	0.40	0.25	1.00	1.65
Екатерина	0.50	1.00	0.25	0.00	1.75
Иннокентий	0.20	0.40	1.00	0.90	2.50
Анна	1.00	0.40	0.75	0.00	2.15

чем нужно позаботиться в данном случае, это – правильно задать тип диаграммы и значения исходных данных.

Гистограмма с накоплением  
Значение целевой функции  
(вариант А)



На этом можно было бы остановиться, но оказалось, что педагогический совет настаивает, чтобы вдвое усилить значимость нормированного балла за успеваемость в школе, т.е. если все остальные нормированные значения войдут в целевую функцию с коэффициентом 1, то у нормированного среднего балла будет коэффициент значимости 2 (такие коэффициенты принято называть весовыми коэффициентами). Итоговая формула для целевой функции в этом случае отображена на рисунке ниже.



## Целевая функция (вариант В)


 $F(x) =$ 

Нормированная средняя отметка за период обучения (x) \* 2

+

Нормированное количество грамот за участие в художественных конкурсах (x)

+

Нормированное количество грамот за участие в интеллектуальных конкурсах (x)

+

Нормированное количество грамот за участие в спортивных мероприятиях (x)

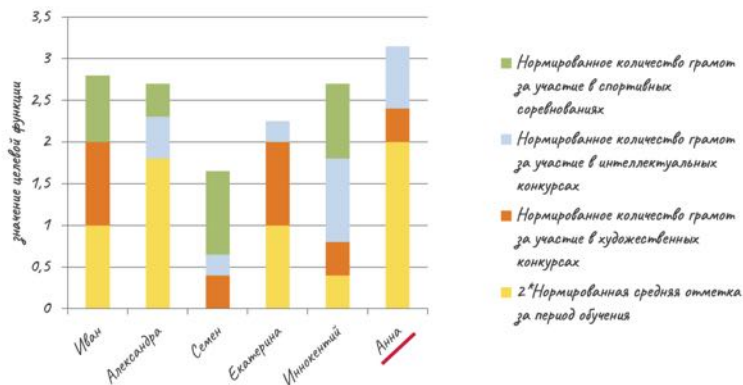
Попробуем определить победителя с новой целевой функцией. На рисунке представлена таблица, в которой в качестве целевой функции используется вариант В. Легко видеть, что максимальное значение целевой функции (2.833) соответствует Анне.

Таблица с явно вычисленной целевой функцией

	Нормированная средняя отметка за период обучения * 2	Нормированное количество грамот за участие в художественных конкурсах	Нормированное количество грамот за участие в интеллектуальных конкурсах	Нормированное количество грамот за участие в спортивных мероприятиях	Значение целевой функции (вариант В)
Иван	1.00	1.00	0.00	0.80	2.8
Александра	1.80	0.00	0.50	0.40	2.7
Семен	0.00	0.40	0.25	1.00	1.65
Екатерина	1.00	1.00	0.25	0.00	2.25
Иннокентий	0.40	0.40	1.00	0.90	2.7
Анна	2.00	0.40	0.75	0.00	3.15

Значит, в этом случае победителем является Анна! Аналогичный вариант мы можем увидеть и на гистограмме с накоплением.

Гистограмма с накоплением  
Значение целевой функции  
(вариант В)



Мы не можем изменить манеру накопления в гистограмме, но мы можем удвоить значения показателя за успеваемость в исходных данных для диаграммы и на их основе построить обычную диаграмму с накоплением. Как и ожидалось, победителем оказалась Анна.

Построение целевой функции на основе нормированных показателей, подбор подходящих весовых коэффициентов в общем случае весьма непростая задача, которая выходит за рамки нашей лекции. Однако к этой теме мы непременно вернемся в рамках курса по Машинному обучению.



## 1.4 Временные ряды

### Анализ временных рядов

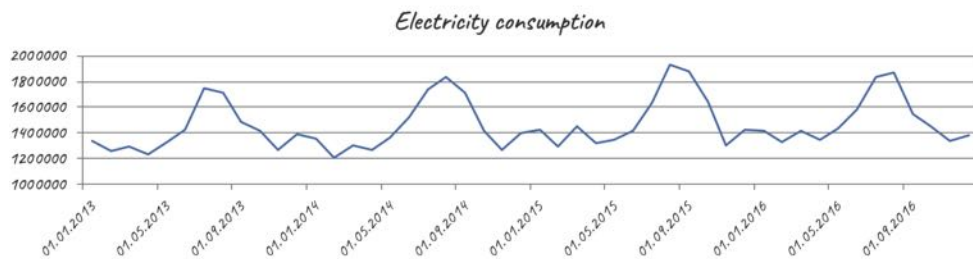
Временной ряд – последовательность наблюдений, упорядоченная по времени:

$$y_1, y_2, \dots, y_n$$

Где  $y_i$  – значения переменной в  $n$  равноотстоящих моментов времени:

$$t_1, t_2, \dots, t_n$$

Примерами временных рядов являются регулярно фиксируемые (каждый день, час, минуту и т.п.) данные о генерации электроэнергии, производстве продукции, продаже товаров, потреблению продукции, транспортных перевозках и т.п. Пример процесса, заданного временным рядом показан на рисунке.



Особенность анализа временных рядов заключается в том, что в отличие от других, независимых наблюдений, мы рассчитываем, что значения ряда в прошлом содержат сведения о его поведении в будущем.

В целом, задачи, с которыми сталкивается исследователь в процессе анализа временных рядов, можно разделить на два класса:

- анализ характерных **составляющих** временного ряда для понимания процессов, отображаемых рядом;
- **прогнозирование** поведения ряда в будущие периоды времени.

Рассмотрим последовательно эти задачи. Под составляющими временного ряда принято понимать следующее:

- **Тренд** – плавное, долгосрочное изменение уровня ряда.
- **Сезонность** – циклические изменения уровня ряда с постоянным периодом (например, ежемесячному потреблению электроэнергии соответствует период размером в 12 месяцев).

- **Цикл** – изменения уровня ряда с переменным периодом (например, экономические циклы, периоды солнечной активности 5-7 лет и т.п.).
- **Шум** – непрогнозируемая, случайная компонента ряда.

На рисунке приведено несколько примеров рядов с указанием визуально обнаруженных в них компонент.



Под прогнозированием ряда понимается построение (моделирование) такой функции  $f$ , которая на основе значений временного ряда  $y_1, y_2, \dots, y_t$  и дополнительного параметра  $h$  выдает прогнозное значение ряда для точек  $t + h$ :

$$f(y_1, y_2, \dots, y_t, h) = \hat{y}_{t+h},$$

где параметр  $h$  – это значение в интервале от 1 до  $H$ ,  $H$  – **горизонт прогнозирования**. В зависимости от значения горизонта прогнозирования модели делят на **краткосрочные** и **долгосрочные**. Однако временная градация прогнозов является условной и чаще всего зависит от особенностей временного ряда.

В процессе решения упомянутых выше задач исследователь сталкивается с рядом проблем, которые следует учитывать в процессе анализа. Перечислим некоторые из них.

- Не всегда просто выявить скрытые закономерности в истории ряда (например, оценить продолжительность периодов, подобрать подходящую аналитическую функцию для тренда и т.п.).
- Закономерности (если они, действительно, есть) могут быть искажены шумами, присутствующими в данных. Такие искажения особенно характерны для данных, получаемых со всякого рода датчиков. Именно

поэтому в анализе разработаны методы предварительной подготовки данных, ориентированные на удаление шумов с помощью специальных методов «сглаживания».

- Развитие динамики ряда в прошлом не гарантирует аналогичное поведение ряда в будущем, которое может значительно измениться под влиянием разного рода внешних факторов (например, динамика цен на нефть кардинально меняется при принятии решений об изменении квот на добычу нефти, смене правительств в нефтедобывающих странах и т.п.).

И, тем не менее, не смотря на указанные выше проблемы, строить модели временных рядов можно и нужно, так как даже с учетом упомянутых рисков они оказываются полезными для развития бизнеса в различных областях экономики. Хотя надо заметить, что в статистике существуют и более точные математические методы, позволяющие оценить, так называемый, **предсказательный интервал**, позволяющий определить диапазон, в котором предсказываемая величина окажется с вероятностью не меньше заданной.

#### *Аддитивная модель*

$$y_t = u_t + s_t + e_t$$

#### *Мультипликативная модель*

$$y_t = u_t \cdot s_t \cdot e_t$$

$u_t$  – трендовая составляющая,

$s_t$  – сезонная составляющая,

$e_t$  – случайная составляющая.

При построении моделей временного ряда принято различать два принципиально различающихся подхода: **аддитивный** и **мультипликативный**. Аддитивная модель предполагает прогнозирование ряда путем рекуррентного прибавления или вычитания некоторых приращений к известным значениям временного ряда. Мультипликативная модель предполагает прогнозирование ряда путем умножения известных членов ряда на некоторые коэффициенты. Например, при построении аддитивной модели было определено, что среднемесячное увеличение спроса на некоторый товар составляет 100 единиц. Тогда прогнозное значение спроса в следующем месяце определяется как предыдущее значение ряда плюс 100 единиц. В мультипликативной модели увеличение спроса могло бы быть определено как повышение спроса на 10 процентов. Тогда прогнозное значение спроса в следующем месяце вычислялось бы как предыдущее, умноженное на 1.1. Сезонная закономерность,

повторяющаяся в ряду с определенной периодичностью, также может быть определена аддитивным или мультипликативным образом.

В рамках данной лекции мы рассмотрим простейшие примеры моделирования временных рядов, однако прежде, чем мы перейдем к приемам построения моделей ряда, необходимо обсудить, как можно оценить качество построенной модели временного ряда и как можно сравнивать построенные модели между собой. Для такого рода сравнений принято использовать, так называемые, метрики качества. **Метрикой** называют функцию для определения расстояния между двумя элементами множества. Таких метрик существует достаточно много, в данном фрагменте лекции мы рассмотрим только некоторые из них.

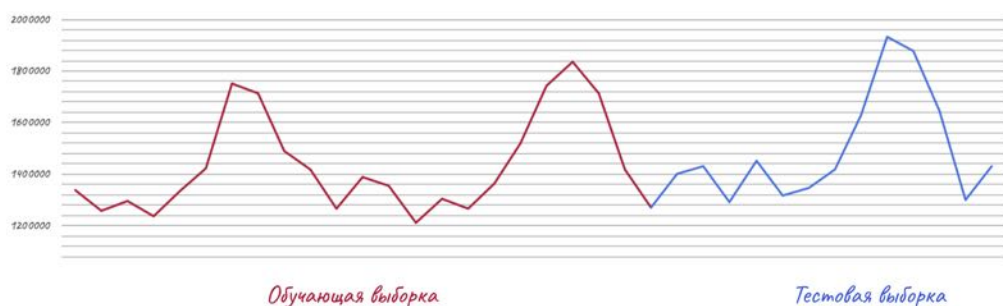
### Метрика

*Метрика – функция для определения расстояния между двумя элементами множества.*



### Метрики качества

Но прежде, чем определяться с выбором метрики, необходимо понять к каким данным они будут применяться. Для того, чтобы оценить качество прогноза, нам нужны не только предсказанные значения, но и реальные данные временного ряда. Причем строить модель для проверки качества нужно на одной части данных, а проверять – на второй. Данные, на которых строится модель, принято называть обучающей выборкой, а данные, на которых модель проверяется, тестовой выборкой. Как поделить временной ряд на обучающую и тестовую выборки? В общем случае, это вопрос может решаться многими способами, однако, в случае временного ряда самое простое и адекватное решение – разделить ряд на три части, а затем первые две использовать для построения модели как обучающую выборку, а оставшуюся часть – для проверки качества построенной модели, т.е. как тестовую.



Теперь можно обсудить и сами метрики. Большинство метрик качества основано на понятии **ошибки прогноза** (будем в дальнейшем ее обозначать как  $e_t$ ). Ошибка прогноза в момент времени  $t$  – это разность между предсказанным и реальным значением переменной в момент времени  $t$ , т.е.

$$e_t = \hat{y}_t - y_t,$$

где  $\hat{y}_t$  – предсказанное значение,  $y_t$  – реальное значение переменной.

Все приведенные далее метрики основаны на предположении, что если модель обеспечивает небольшие суммарные ошибки в прошлом, то она же обеспечит небольшие суммарные ошибки в будущем.

Первая из метрик (MAE) с названием **средняя абсолютная ошибка**, получается как результат деления суммы абсолютных значений ошибок прогноза на количество точек тестовой выборки:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

Вторая (MSE) – **среднеквадратичная ошибка**, вычисляется как сумма квадратов ошибок прогноза, деленная на количество точек тестовой выборки:

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

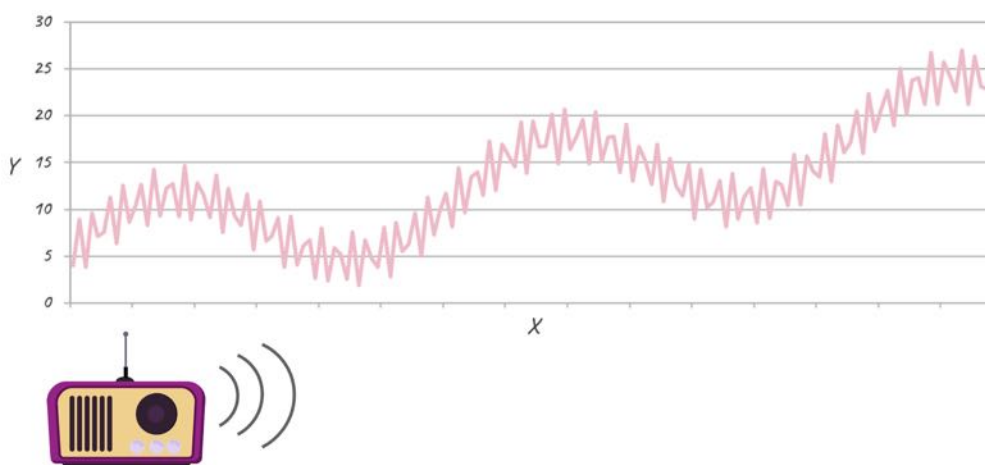
Третья (MAPE) – **средняя абсолютная процентная ошибка**, определяется как процентное соотношение суммы отношений ошибок прогноза и реальных значений временного ряда к количеству точек тестовой выборки. В приведенных формулах:  $e_t$  – ошибка прогноза,  $y_t$  – реальное значение переменной,  $n$  – количество точек тестовой выборки:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \cdot 100\%$$

Именно эти метрики мы и будем использовать в дальнейшем для оценки моделей временных рядов. При подборе подходящей модели будем ориентироваться на метрики с минимальными значениями.

## 1.5 Сглаживание временных рядов

Мы уже упоминали в предыдущем фрагменте лекции, что некоторые значения временных рядов активно сопровождаются шумами, т.е. случайными вариациями в той или иной форме. Как правило, это относится к рядам, которые формируются на основе показаний разного рода датчиков. Пример такого зашумленного ряда вы можете видеть на экране. Шумы мешают пониманию реальной структуры ряда и интерпретации процессов, образующих этот ряд, и поэтому существует множество методов, позволяющих избавиться от шумов. В этом фрагменте лекции мы рассмотрим несколько наиболее известных и применяемых методов, используемых на практике для удаления шумов.



Первый из методов – **метод скользящего среднего**. Суть метода сводится к тому, что для каждого значения переменной ряда формируется окно из соседних значений ряда (в идеале к значений до сглаживаемого значения и к после), а затем на основании этих соседей и самого исходного значения ряда вычисляется среднее арифметическое). Формулу, по которой происходит сглаживание:

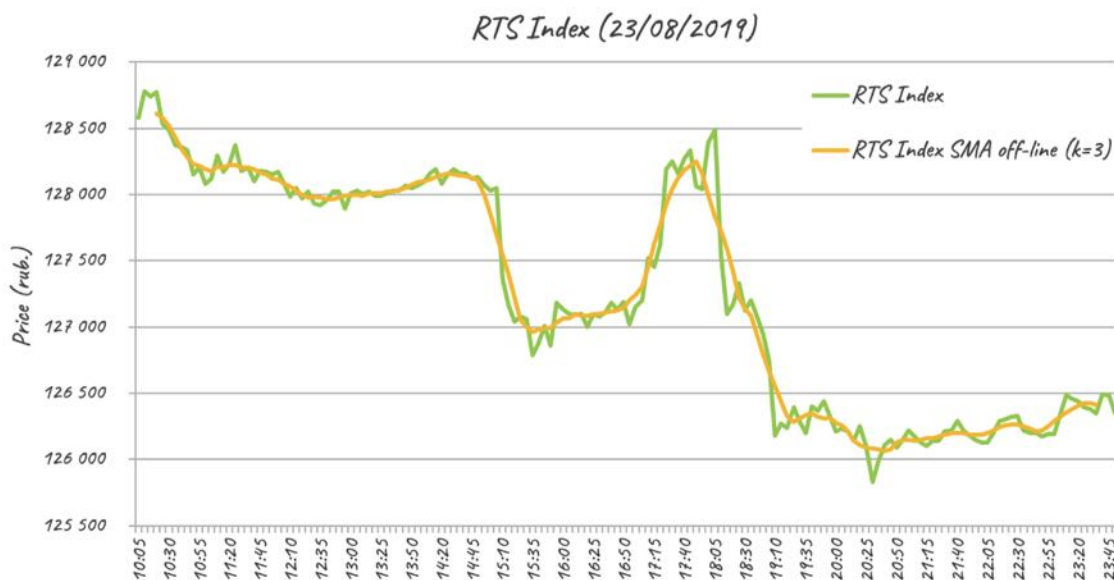
$$s_i = \frac{1}{2k+1} \sum_{j=-k}^k y_{i+j}$$

Здесь  $y_i$  – значение исходного ряда,  $s_i$  – значение сглаженного ряда,  $2k+1$  – ширина окна. От ширины окна зависит степень сглаживания. При задании большой ширины окна сглаживание будет грубым и, возможно, будет потеряна полезная информация о динамике ряда. При задании небольшого окна в 5-7 точек могут остаться шумы. Универсальные значения для ширины скользящего окна задать невозможно – они сильно зависят от предметной области и от целей усреднения в каждом конкретном случае.

Сглаживание активно используется на этапе технического анализа биржевых котировок и встроено во все инструменты, предназначенные для бир-



жевой аналитики. На экране приведен пример исходного и сглаженного графика котировок индекса RTS с применением метода скользящего среднего при значении  $k$  равном 3. Обратите внимание на то, что особенности расчета скользящего среднего не позволяют рассчитать сглаженное значение для  $k$  первых и  $k$  последних точек ряда. Приведенная выше формула расчета

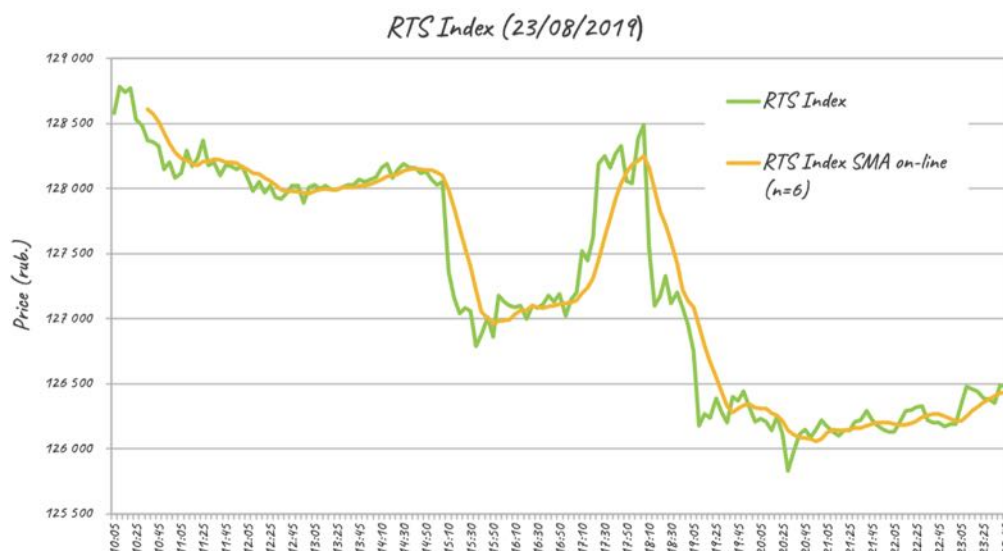


скользящего среднего подразумевает, что сглаживание происходит в режиме off-line (т.е. когда известны и предыдущие и следующие значения для каждой сглаживаемой точки ряда). Однако на практике иногда приходится сталкиваться с необходимостью сглаживания ряда в режиме on-line (когда известны только предыдущие значения и сама сглаживаемая точка ряда). Именно такое сглаживание применяется в биржевой деятельности при оперативном анализе котировок. Можно ли в режиме on-line применять метод скользящего среднего?

Можно. Только формула для расчета в этом случае будет указывать соседей, которые появились до сглаживаемого значения:

$$s_i = \frac{1}{n+1} \sum_{j=-n}^0 y_{i+j},$$

где  $y_i$  – значение исходного ряда,  $s_i$  – значение сглаженного ряда,  $n+1$  – ширина ряда. Для такого сглаживания характерно запаздывание в отображении сглаженного ряда, но, тем не менее, для некоторых задач этого метода сглаживания вполне достаточно. Для сглаживания в режиме on-line часто используется еще один популярный метод сглаживания. Это – **экспоненциальный метод сглаживания**. Он чрезвычайно прост в реализации, так как описывается рекуррентной формулой следующего вида.



Пусть  $Y = \{y_1, \dots, y_T\}$  – временной ряд. Экспоненциальное сглаживание ряда осуществляется по рекуррентной формуле:

$$s_t = \alpha y_t + (1 - \alpha)s_{t-1},$$

где  $y_t$  – значение исходного ряда в точке  $t$ ,  $s_{t-1}$  – значение сглаженного ряда в точке  $t - 1$ ,  $\alpha \in (0, 1)$  – коэффициент сглаживания. При этом начальное значение  $s_1$  определяется как первая точка ряда:

$$s_1 = y_1.$$

Выбор коэффициента сглаживания  $\alpha$  является решающим моментом при экспоненциальном сглаживании.

Текущее сглаженное значение складывается из предыдущего и некоторой доли ошибки предыдущего сглаживания. Величина этой ошибки, которая используется для корректировки, определяется коэффициентом сглаживания  $\alpha$ . Чем ближе значение  $\alpha$  к 1, тем большая часть расхождения сглаживания и реального значения считается закономерной и используется для вычисления очередного значения. Чем ближе значение  $\alpha$  к нулю, тем большая доля расхождения между сглаженным и реальным значением считается случайной и, соответственно, меньшая часть используется для вычисления очередного значения.

Формула для экспоненциального сглаживания может быть переписана в нереккуррентном виде:

$$s_t = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha)^2 y_{t-2} + \alpha(1 - \alpha)^3 y_{t-3} + \dots + \alpha(1 - \alpha)^{t-1} y_1$$

Из формулы на экране хорошо видно, что сглаженное значение представляет собой взвешенную сумму всех предыдущих значений ряда, причем коэффициенты уменьшаются по мере удаления значения ряда от текущего момента



времени. Так, например, если  $\alpha = 0.1$ , то формула приобретает следующий вид:

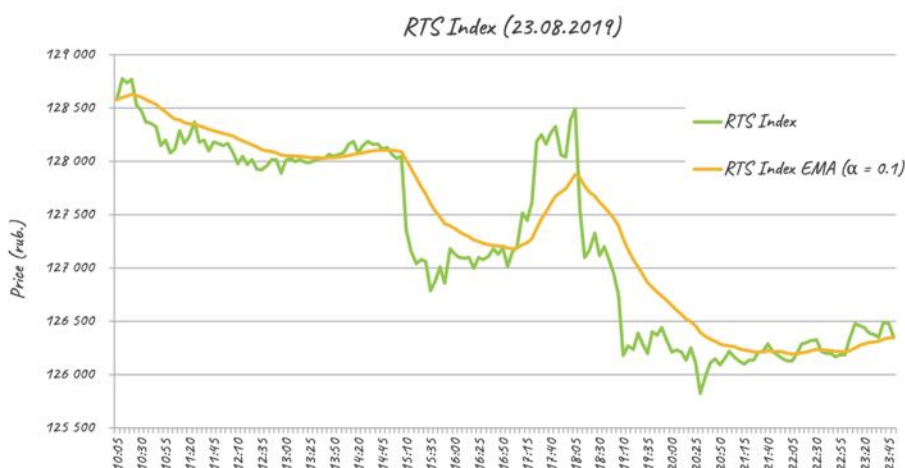
$$s = 0.1y_t + 0.09y_{t-1} + 0.081y_{t-2} + 0.0729y_{t-3} + \dots$$

Как правильно выбрать коэффициент сглаживания? Не существует четких формальных критериев выбора значения  $\alpha$ . На практике (по крайней мере в биржевой деятельности) чаще всего используются значения  $\alpha$ , лежащие в пределах от 0.1 до 0.3. Можно сказать, что значение коэффициента сглаживания отражает субъективное мнение исследователя относительно устойчивости изменения изучаемого показателя.

На рисунке приведены примеры экспоненциального сглаживания индекса RTS с различными значениями коэффициента сглаживания. На первом графике коэффициент  $\alpha = 0.1$ . На втором графике коэффициент  $\alpha$  в 10 раз

*Пример*

*экспоненциальное сглаживание ( $\alpha = 0,1$ )*

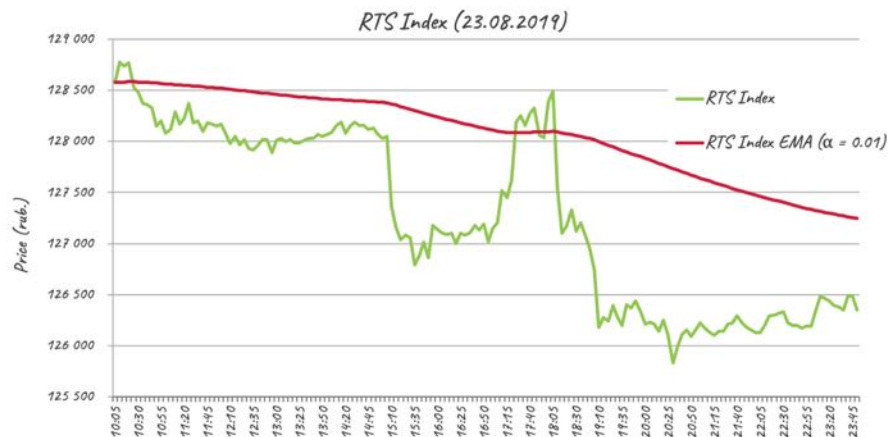


меньше – 0.01. Обратите внимание, как сильно повлиял выбор коэффициента сглаживания на результат.

## 1.6 Определение трендов временных рядов

В предыдущем фрагменте мы рассмотрели основные составляющие типичного временного ряда – тренд, сезонная составляющая и шумы. Мы уже обсудили, что в случае наличия шумов они могут быть удалены при предварительном анализе ряда с помощью специальных методов. И очистка от шума позволит лучше увидеть динамические тенденции ряда. Как выделить тренд и сезонную составляющую? Можно ли их определить аналитически (т.е. с помощью математической функции, зависящей от времени)? Что это дает? Если мы научимся описывать аналитически поведение временного ряда – мы сможем в дальнейшем прогнозировать поведение ряда. С точки зрения

## Пример экспоненциальное сглаживание ( $\alpha = 0,01$ )



УНИВЕРСИТЕТ ИТМО

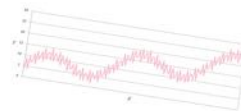
Определение трендов временных рядов  
Как определить тренд временного ряда?



$$y(x) = 1 - \exp\left(1 - \frac{x}{x_{\min}}\right)$$



$$s_i = \frac{1}{2k+1} \sum_{j=-k}^k y_{i+j}$$



реального бизнеса это значит, что мы можем планировать продажи автомобилей, количество пассажиров, потребителей различных услуг, посетителей ресторанов и т.п. А вот это – действительно интересно!

Итак, как определить тренд временного ряда? На практике для этого, как правило, используются следующие аналитические функции: линейная, полиномиальная, экспоненциальная, логарифмическая:

Разумеется, другие функции тоже возможны, но именно эти функции используются чаще других и встроены во многие существующие инструменты. Как определить, какая функция подходит в том или ином случае? Конечно, для этого существуют и формальные методы, но простейший способ – вспомнить, как выглядят графики соответствующих функций и по графику временного ряда подобрать подходящую по виду функцию. Приведем несколько примеров.

Однако для того, чтобы в дальнейшем моделировать поведение ряда,

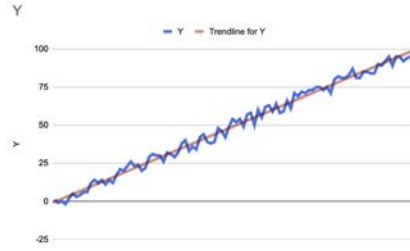
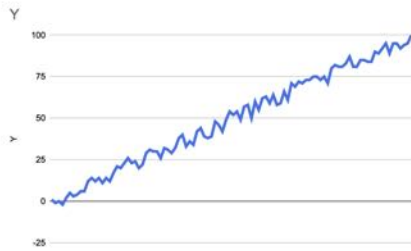
Линейная  $f(x) = a + bx$

Полиномиальная  $f(x) = a + b_1x + b_2x^2 + \dots + b_kx^k$

Экспоненциальная  $f(x) = ce^{a+bx}$

Логарифмическая  $f(x) = a \log_b x + c$

### Пример 1 (линейный тренд)



недостаточно выяснить тип функции, соответствующий линии тренда. Надо выяснить точные параметры функции. Как их узнать? Для каждого из упомянутого выше типа функций существуют подходящие математические методы, позволяющие определить эти параметры.

Так, например, для определения параметров линейного тренда можно воспользоваться, методом наименьших квадратов, который позволяет явно вычислить коэффициенты функции по формулам, приведенным на экране. Для определения параметров других трендов также существуют специальные математические методы. Но в рамках данной лекции мы не будем останавливаться на них подробно. Однако обратим внимание на то, что сами тренды и аналитические функции, которые им соответствуют, замечательно определяются простейшими инструментами.

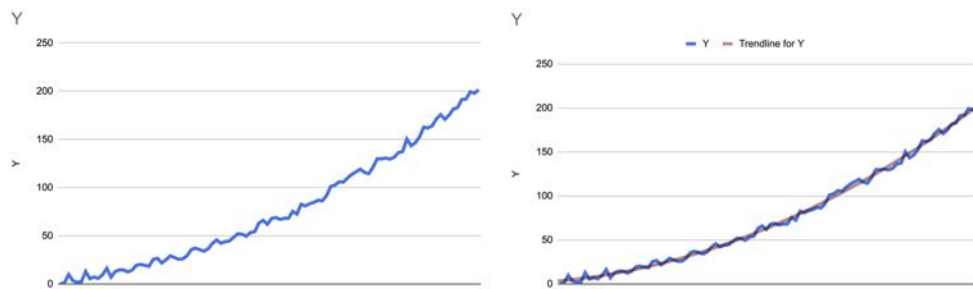
$$y = ax + b,$$

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2},$$
$$b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n},$$

где  $n$  — количество измерений,  $y_i$  — элемент временного ряда,  $x_i$  — время.

Прежде, чем приступить к непосредственному построению трендов, обратим внимание еще на одно обстоятельство. Как подобрать подходящую линию тренда, если вариантов может быть несколько, и вы не уверены, какой из них лучше? Есть ли формальные критерии, позволяющие выбрать тип

### Пример 2 (полиномиальный тренд)



### Пример 3 (логарифмический тренд)

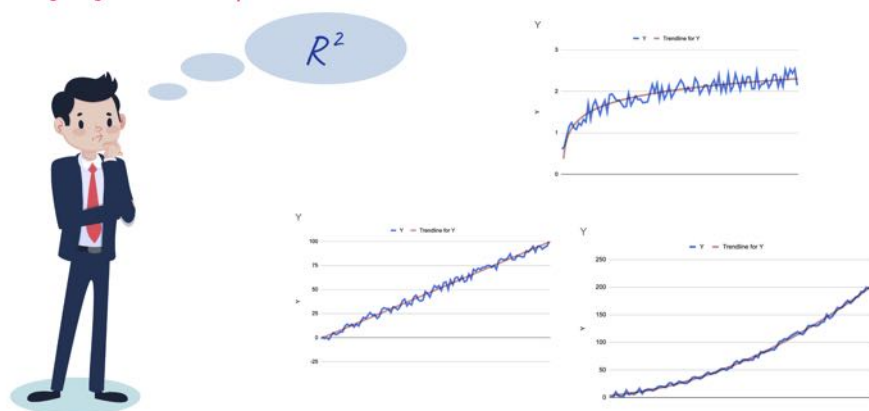


тренда? Оказывается, есть! Такой критерий называется – **коэффициент детерминации**. Обозначается –  $R^2$ :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - y_{avg})^2},$$

где  $y_i$  – значения временного ряда в момент времени  $i$ ,  $f_i$  – значение тренда в момент времени  $i$ ,  $y_{avg}$  – среднее значение элементов временного ряда.

*Как выбрать лучшую линию тренда?*



Коэффициент детерминации может использоваться для оценки качества подобранного уравнения тренда. Он принимает значения от 0 до 1. Для приемлемых моделей тренда предполагается, что коэффициент детерминации

должен быть хотя бы не меньше 0.5. Модели с коэффициентом детерминации выше 0.8 можно признать достаточно хорошими. Значение коэффициента детерминации  $R^2 = 1$  означает функциональную зависимость между переменными (т.е. между исходным временным рядом и трендом). Точная формула для расчета коэффициента детерминации представлена на слайде.

Коэффициент детерминации реализован во многих инструментах, и мы будем его использовать для оценки качества построенных линий трендов.

## Построение линии тренда

Кроме отображения уравнения тренда на диаграмме, в инструментах для обработки данных существуют функции, которые могут вычислять коэффициенты линейного тренда без прорисовки соответствующего графика. Это функции SLOPE(НАКЛОН) и INTERSECT(ОТРЕЗОК). Если уравнение для тренда задается как  $y(x) = ax + b$ , то функция SLOPE вычисляет коэффициент  $a$ , а INTERSECT –  $b$ . В некоторых обстоятельствах пользоваться этими функциями предпочтительнее, так как они выдают более точный результат, чем уравнение с округленными коэффициентами, которое мы видим на графике с трендом.

На рисунке приведен пример использования функции SLOPE в Google таблице: И функции INTERSECT в той же таблице. И продемонстрирован

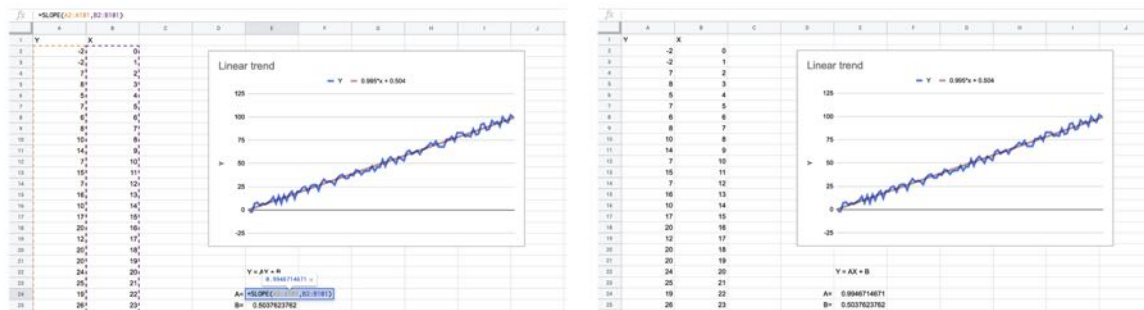


Рис. 1: Функция SLOPE

полученный результат. Сравните точность представления для коэффициента  $a$  в уравнении на графике и в результате применения функции. Итак, мы научились строить тренды и определять аналитические функции, которые им соответствуют. Как можно воспользоваться полученными знаниями в практических целях? Можно попробовать применить их для прогнозирования поведения несложных временных рядов, в которых, возможно, присутствуют шумы и тренды, но отсутствует сезонная составляющая. То есть мы можем смоделировать ряд на основе аналитической функции, соответствующей тренду. В нашем распоряжении есть подходящий ряд.

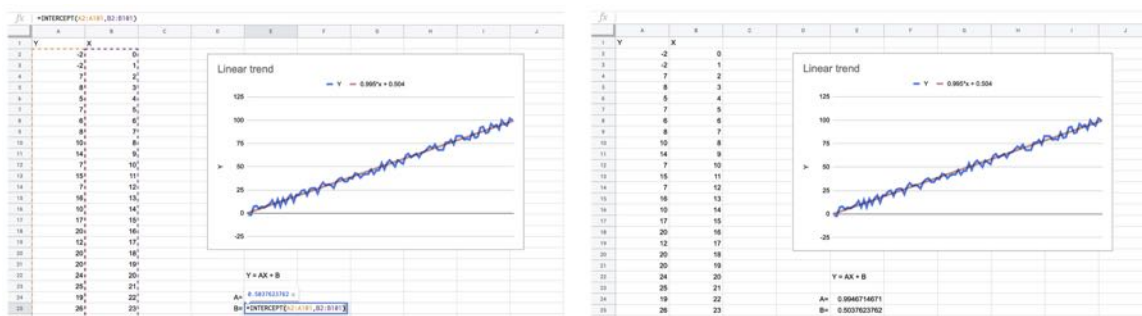
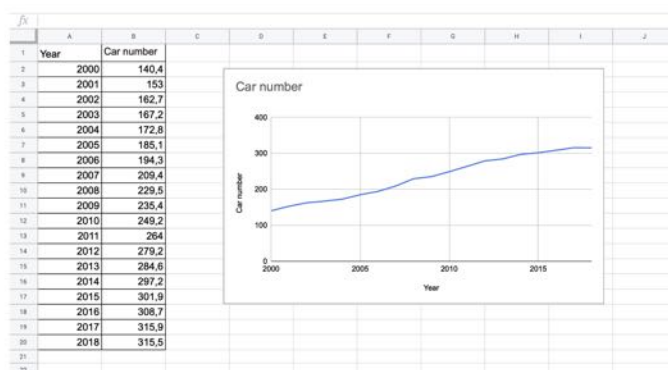


Рис. 2: Функция INTERSEPT

Это уже упоминавшийся ранее ряд с количеством автомобилей на 1000 человек в Центральном Федеральном Округе. Данные содержат сведения с 2000 по 2018 годы. Построим линейный график, соответствующий этому ряду,

*Количество автомобилей на 1000 человек в Центральном Федеральном Округе*



и убедимся, что в нем нет сезонности. Значит, у нас есть основания моделировать ряд аналитической функцией вида  $ax + b$ .

Попробуем проверить, насколько эффективно моделируют точки ряда с помощью такой функции. Для этого разделим ряд на обучающую и тестовую выборки, определим параметры тренда на обучающей выборке, выполним прогноз на тестовой выборке, а потом оценим качество прогноза с помощью метрики MAPE.

Обучающую выборку выделим желтым цветом, тестовую – голубым. Определим параметры линейного тренда с помощью функций SLOPE и INTERSEPT. Вычислим прогнозные значения на тестовой выборке (т.е. с 2013 по 2018 гг.) по формуле  $ax + b$ . Результат вы можете видеть на экране в столбце Forecast. Отметим, что этот же результат в Google таблицах можно было бы получить и иным способом, т.к. существует встроенная функция FORECAST.LINEAR, которая выдает прогнозные значения на основе модели линейного тренда указанного временного ряда. Далее можно оценить качество прогноза с помощью метрики MAPE. Для этого заведем в табли-



## Вычисление метрики MAPE

A	B	C	D	E	F	G
2000	140.4					
2001	153					
2002	162.7					
2003	167.2					
2004	172.8					
2005	185.1					
2006	194.3					
2007	209.4					
2008	229.5					
2009	235.4					
2010	249.2					
2011	264					
2012	279.2					
2013	284.6	283.10	0.01			
2014	297.2	294.50	0.01			
2015	301.9	305.91	0.01			
2016	308.7	317.32	0.03			
2017	315.9	328.72	0.04			
2018	315.5	340.13	0.08			

Y=AX+B  
A= 11,40714286  
B= -22679,48242

MAPE= 2.90%

це специальный столбец с заголовком Error и вычислим в нем слагаемые числителя для метрики MAPE. После этого остается сложить полученные в столбце Error значения и разделить на количество элементов тестовой выборки. Результат 2.9% показывает среднее отклонение от реальных значений и выглядит достаточно убедительно. В следующем фрагменте мы рассмотрим, как моделируются ряды с сезонными составляющими.

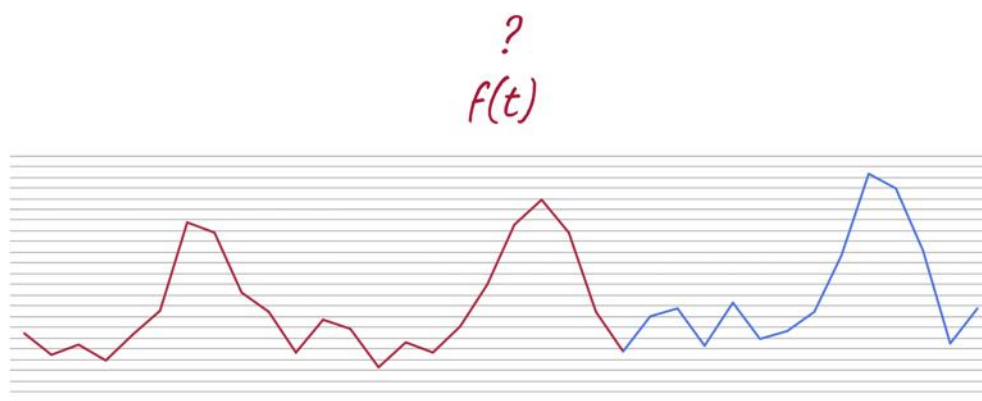
## 1.7 Построение моделей для временных рядов с сезонными составляющими

Как определить сезонную составляющую временного ряда и как в дальнейшем ее использовать для моделирования ряда? Именно этот вопрос и является предметом обсуждения данного фрагмента. Многие временные ря-



ды экономического происхождения содержат сезонную составляющую в силу

особенностей бизнес-процессов. Например, общая картина потребления электроэнергии периодически повторяется и зависит от месяца, посещаемость ресторанов также повторяется и зависит от дня недели, количество пассажиров в общественном транспорте зависит от времени суток и т.п. Эти процессы изначально позволяют определить продолжительность повторяющегося периода, который так важен для моделирования временных рядов с сезонной составляющей. В приведенных примерах длина периода равна 12 месяцам для потребления электроэнергии, 7 дням для ресторанов и 24 часам для общественного транспорта. А как следует поступать с рядами, с менее очевидными сезонными составляющими? Как определить длину повторяющегося периода и его сезонные компоненты? В общем случае для определения длины периода существуют достаточно сложные математические теории, однако мы упростим себе задачу и в рамках данного фрагмента будем предполагать, что длина периода нам точно известна, и, кроме того, исследуемый ряд уже избавлен от шумов. Итак, рассмотрим, как можно определять сезонные составляющие для периодического временного ряда и строить модели для его прогнозирования. Пусть исходные данные – временной ряд  $f(t)$ .



- Во-первых, рекомендуется построить аналитическое выражение для трендовой составляющей. Полученный ряд (т.е. тренд) для определенности назовем  $d(t)$ .
- Во-вторых, нужно выделить сезонную составляющую. Для этого потребуется вычесть трендовую составляющую из исходного временного ряда. То есть, построить ряд  $r(t) = f(t) - d(t)$ .
- И, наконец, можно строить модель для прогнозирования ряда на основе трендовой и сезонной составляющих.

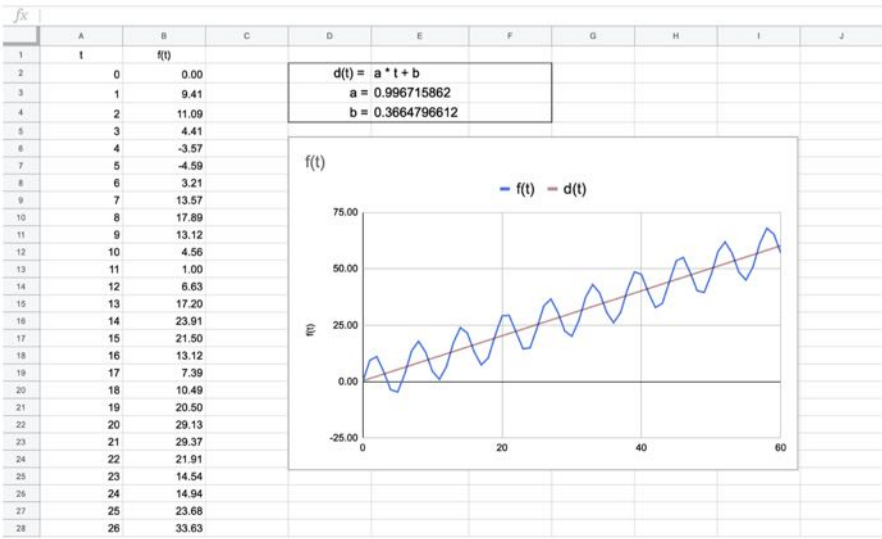
Такая схема построения моделей является достаточно типичной для моделей с сезонной составляющей. Однако сами модели могут сильно отличаться



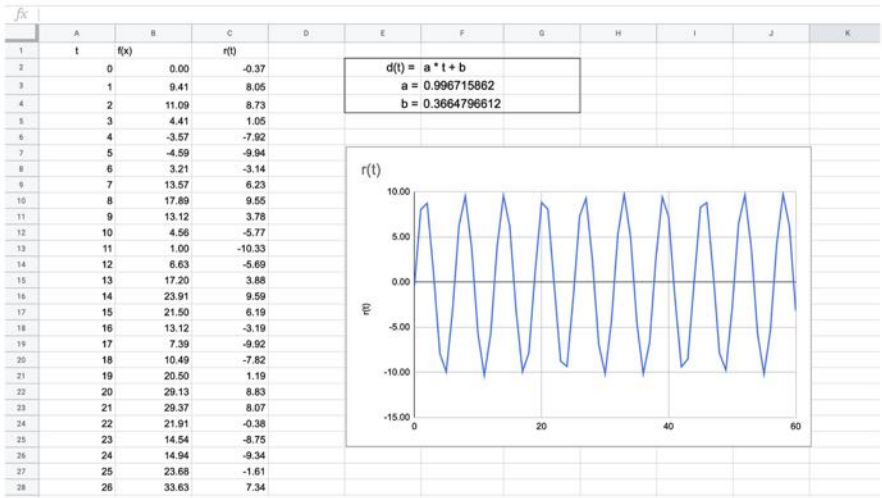
ся в зависимости от особенностей сезонной составляющей. Продемонстрируем эти особенности на двух модельных примерах.

**Ряд с сезонной составляющей постоянного размаха**

На рисунке представлен исходный ряд с сезонной составляющей постоянного размаха. Кроме исходных данных построен график, соответствующий этим данным, линия тренда и параметры для линейного уравнения тренда. Периодичность такого ряда равна 6. Так как в нашем распоряжении есть па-



раметры линейного тренда, нетрудно построить и сезонную составляющую ряда, которая получается простым вычитанием из исходного ряда трендовой составляющей. Результат в виде таблицы и графика вы можете видеть на рисунке. Именно на этом графике отчетливо видно, почему мы назвали



этот ряд рядом с сезонной составляющей постоянного размаха. Локальные

минимумы сезонной составляющей (точно так же, как и максимумы) лежат практически на одной линии параллельной оси времени.

Для ряда такого вида характерно, что сезонные компоненты для момента времени  $t$  такие же, как в момент времени  $t - n$ , где  $n$  – длина периода. А это дает возможность написать очень простую формулу для прогнозирования такого ряда на краткосрочный период (т.е. на 1 период вперед). Пусть последнее значение ряда определено в момент времени  $t$ . Тогда формула для краткосрочного прогнозирования в момент времени  $t + k$  может состоять из двух слагаемых: значение трендовой составляющей в момент времени  $t + k$  и сезонной составляющей в момент времени  $t + k - n$ .

**Модель для краткосрочного прогнозирования ряда с сезонной компонентой постоянного размаха:**

$$\text{forecast}(t + k) = d(t + k) + r(t + k - n),$$

где  $t$  – момент времени, в который известна последняя точка исходного ряда,  $t + k$  – момент, для которого осуществляется прогнозирование  $k \leq n$ ,  $n$  – длина периода,  $d(t + k)$  – тренд в точке  $t + k$ ,  $r(t + k - n)$  – сезонная составляющая в момент  $t + k - n$ .

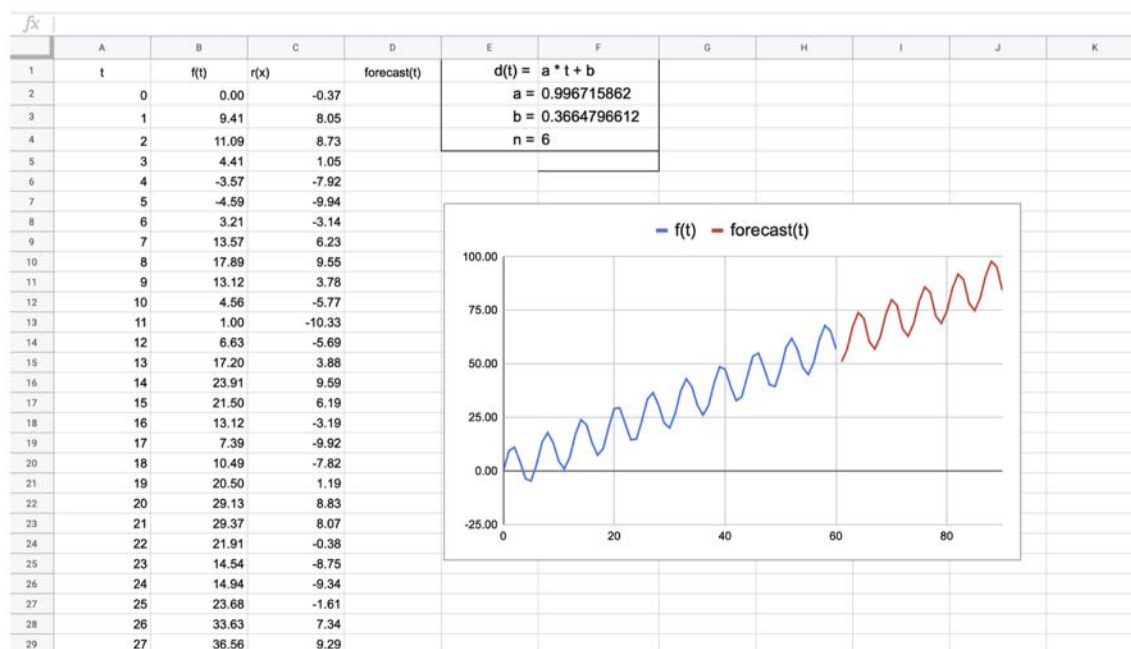
Что должно измениться в модели для долгосрочного прогнозирования? Трудность состоит в том, что можно учитывать сезонную составляющую только на основе реально существующих значений ряда! То есть до момента времени  $t$ . Но и с этим можно справиться! Можно использовать в формуле последний известный период, даже если мы прогнозируем на несколько периодов вперед. Для этого в формуле, которую мы использовали для краткосрочного прогнозирования, достаточно в сезонной составляющей использовать не само значение  $k$ , которое указывает как далеко мы отошли от известных значений ряда, а его значение по модулю  $n$ , где  $n$  – длина периода.

**Модель для долгосрочного прогнозирования временного ряда с сезонной компонентой постоянного размаха**

$$\text{forecast}(t + k) = d(t + k) + r(t + k \bmod n - n),$$

где  $t$  – момент времени, в который известна последняя точка исходного ряда,  $t + k$  – момент, для которого осуществляется прогнозирование,  $n$  – длина периода,  $d(t + k)$  – тренд в точке  $t + k$ ,  $r(t + k \bmod n - n)$  – сезонная составляющая в момент  $t + k - n$ .

На рисунке можно видеть рассчитанный по этой формуле долгосрочный прогноз ряда (на 5 периодов вперед). Исходный ряд изображен синей линией, а прогнозные значения – красной. Результат выглядит достаточно убедительно. Однако, при этом надо понимать, что долгосрочный прогноз всегда хуже краткосрочного, так как чем ближе мы находимся к исходным данным, тем



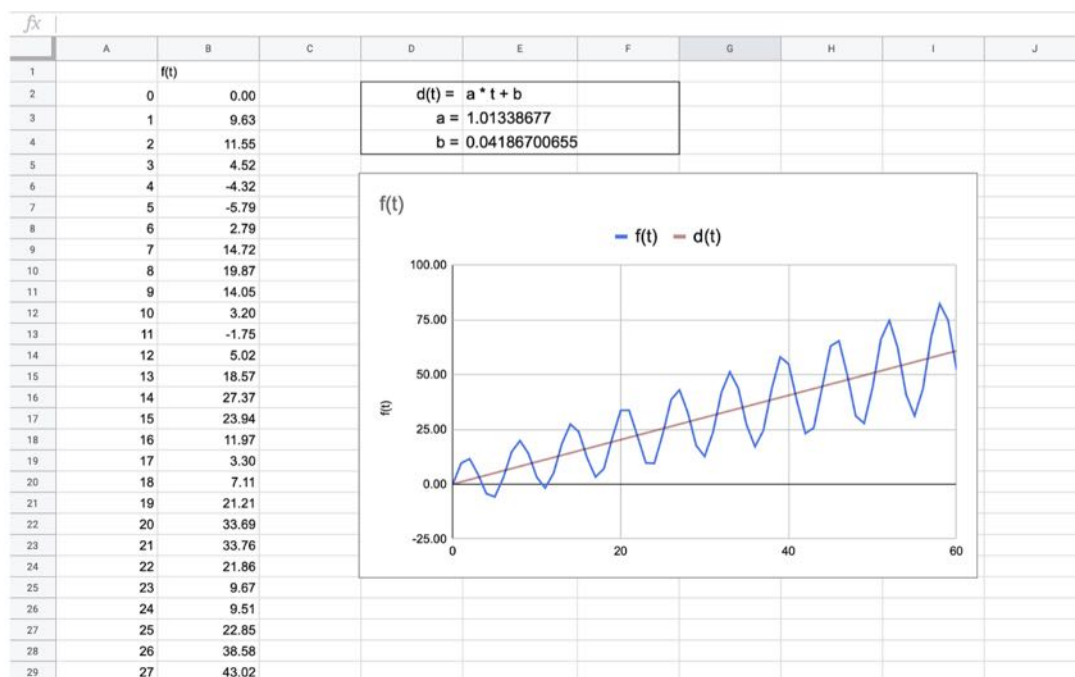
более основательны наши предположения о поведении ряда. Качество прогноза, разумеется, можно и нужно проверять с помощью метрик, которые мы уже обсуждали в данной лекции.

## Ряд с сезонной компонентой растущего размаха

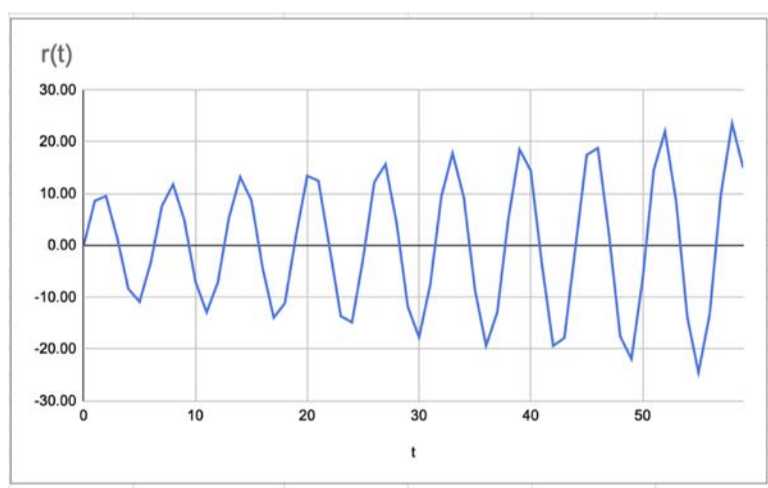
На рисунке представлен еще один временной ряд. Его отличительная особенность – наличие сезонной составляющей возрастающего размаха. Проведен первичный анализ данных. Построен график, соответствующий исходным данным, линия тренда и явно вычислены параметры для уравнения линейного тренда. Вычтем из исходного ряда трендовую составляющую. Визуализируем полученный результат. После визуализации растущий размах сезонных компонент ряда становится отчетливо виден.

На графике хорошо видно, что размах меняется со временем, причем достаточно равномерно. Модель краткосрочного прогнозирования, которая была использована для ряда с постоянным размахом, может быть использована и в этом случае, так как она все равно будет учитывать динамику изменения размаха, но с некоторым отставанием (будет запаздывать на один период).

Поэтому лучше все-таки формулу для краткосрочного прогнозирования второго ряда модифицировать и написать так, чтобы в ней учитывалось приращение сезонных компонент за 1 период. Более всего нас интересует поведение именно сезонных компонент, т.к. с трендовой составляющей мы можем справиться при помощи аналитического формулы. Как мы уже упоминали ранее, у нас есть два возможных подхода: аддитивный и мультипликатив-



ный. В аддитивном случае мы должны указать, на какое приращение изме-



няется компонента, а в мультипликативном – во сколько раз. Выберем второй, мультипликативный вариант. Остается совсем немного – выяснить, так называемый, коэффициент сезонности. Это та величина, на которую будут умножаться значения сезонных компонент последнего известного периода. Есть разные способы для вычисления такого коэффициента, но, пожалуй, простейший – сравнить размах последнего и предпоследнего периодов. Соотношение этих размахов и можно определить как коэффициент сезонности. Как вычислить размах в последнем и предпоследнем периодах? Он может быть определен как разница между максимальным и минимальным значениями точек периода. На экране приведен пример вычисления коэффициента сезонности в предположении, что длина периода – 6, а последняя известная точка сезонной составляющей содержится в ячейке C62.

fx =(MAX(C57:C62)-MIN(C57:C62))/(MAX(C51:C56)-MIN(C51:C56))							
	A	B	C	D	E	F	G
1	t	f(t)	r(t)	forecast(t)	d(t) = a * t + b		
2	0	0.00	-0.04		a = 1.01338677		
3	1	9.63	8.57		b = 0.0418670065		
4	2	11.55	9.48		n = 6		
5	3	4.52	1.44		t = 60		
6	4	-4.32	-8.42		i_season = 1.09		
7	5	-5.79	-10.90				

После того, как коэффициент сезонности определен, нет никаких проблем в получении модели для краткосрочного прогнозирования. Формула приведена на экране. Это практически тоже, что и для ряда с постоянным сезонным размахом, только сезонная компонента из последнего известного периода умножена на коэффициент сезонности.

**Модель для краткосрочного прогнозирования ряда с сезонной компонентой возрастающего размаха**  $t$  – момент времени, в который известна последняя точка исходного ряда,  $t + k$  – момент, для которого осуществляется прогнозирование  $k \leq n$ ,  $n$  – длина периода,  $d(t + k)$  – тренд в точке  $t + k$ ,  $r(t + k - n)$  – сезонная составляющая в момент  $t + k - n$ ,  $i_{season}$  – коэффициент сезонности.

Что меняется в случае долгосрочного прогнозирования? Необходимо учитывать то приращение, которое увеличивает размах сезонных компонент, причем учитывать его необходимо на несколько периодов вперед. Если коэффициент сезонности хотя бы приблизительно можно считать одинаковым для всех периодов, то для построения модели можно использовать следующую формулу, которая приведена на экране. Обратите внимание на выражение  $k \text{ div } n + 1$  ( $k$  деленное нацело на  $n$  плюс 1). Оно отражает относительный номер периода, для которого ведется прогноз. При этом нумерация таких периодов начинается после момента времени  $t$ .

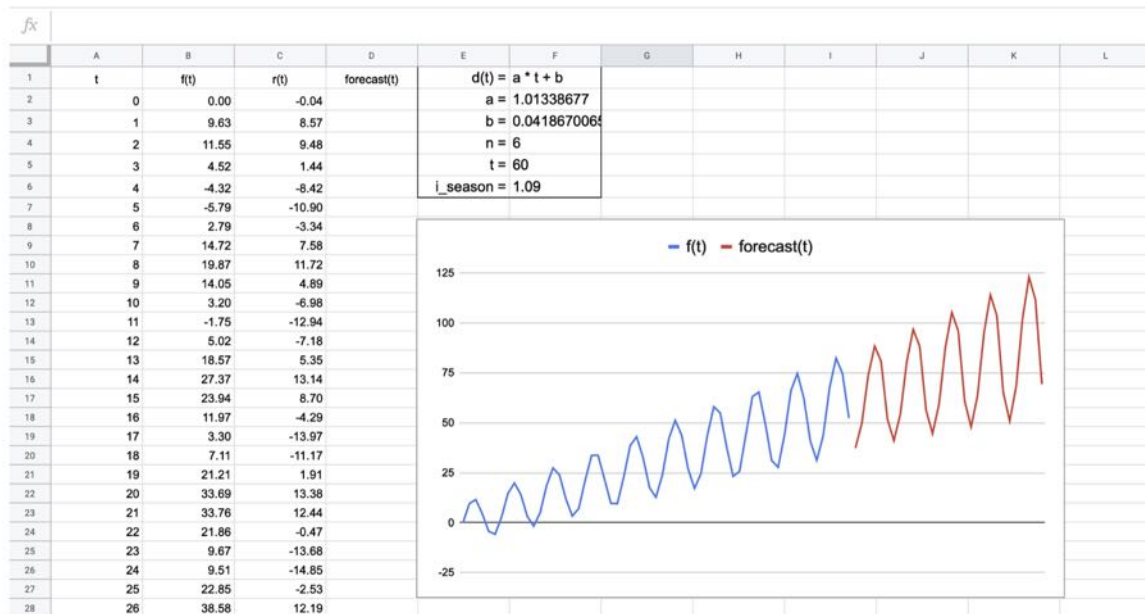
**Модель для долгосрочного прогнозирования с сезонной компонентой возрастающего размаха**

$$forecast(t + k) = d(t + k) + r(t + k \bmod n - n) \cdot i_{season}^{k \text{ div } n + 1},$$

где  $t$  – момент времени, в который известна последняя точка исходного ряда,  $t + k$  – момент, для которого осуществляется прогнозирование,  $n$  – длина периода,  $d(t + k)$  – тренд в точке  $t + k$ ,  $r(t + k \bmod n - n)$  – сезонная составляющая в момент  $t + k - n$ ,  $i_{season}$  – коэффициент сезонности. На следующем рисунке можно видеть рассчитанный по этой формуле долгосрочный прогноз ряда с возрастающим размахом. Исходный ряд на графике изображен синей линией, а прогнозные значения – красной.

Разумеется, в случае если мы имеем дело с реальными рядами, нужно оценить качество прогноза и убедиться, что мы выбрали лучшую модель

из возможных. Но мы не будем это делать в данном фрагменте, так как надеемся, что достаточно подробно говорили об оценке качества модели в предыдущем фрагменте лекции. Итак, мы рассмотрели несколько простей-



ших способов моделирования значения рядов с сезонной компонентой. Разумеется, точность прогноза уменьшается по мере удаления горизонта прогноза от исторических данных. Тем не менее, даже эти модели могут оказаться полезными при прогнозировании поведения реальных экономических рядов.