МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени
первого Президента России Б. Н. Ельцина»

МЕТОДЫ ПРОГНОЗИРОВАНИЯ ДАННЫХ

Лекция № 2

Основные характеристики моделей и временных рядов

Екатеринбург
2020

Содержание

Часть 1. Числовые характеристики

По определению временной ряд (ВР) есть только частная выборка случайной величины, описываемой некоторой функцией распределения. Если при этом опираться на модели ВР, которые описывались в лекции 1, то свойства этого распределения должны явно соотноситься со свойствами случайной составляющей $\xi(t)$.

Под числовыми характеристиками ВР мы будем понимать такие важные статистические характеристики, которые возможно рассчитать для любого выбранного временного ряда сразу же, без наличия каких-либо априорных знаний о его характере, а также без предварительной оценки функции распределения той случайной величины, что лежит в основе исходного анализируемого процесса. Стоит отметить, что, тем не менее, для расчета большинства из них придется изначально провести проверку временного ряда на принадлежность к тому или иному классу ВР. Самое важное – это на начальном этапе определить с помощью статистических проверок, является исследуемый ряд стационарным или нестационарным. Как мы увидим в дальнейшем, большинство оценок численных характеристик ряда опирается на понятие эргодичности, а такой характер случайных процессов не существует в рамках нестационарных временных рядов.

Из определения стационарных в широком смысле случайных процессов (а именно с ними мы будем работать до изучения адаптивных методов анализа ВР) вытекает, что **мат. ожидание** и **дисперсия** выбранного ряда не зависят от выбора начальной точки отсчета. Тогда эти важные *характеристики временного ряда* могут быть оценены на основе строгих теоретических определений этих величин для случайных последовательностей, с учетом их описания в дискретной форме в виде отсчетов временной сетки. Многие другие важные характеристики временных рядов опираются на так называемые начальные и центральные статистические моменты *n*-го порядка,

которые зависят от определения мат. ожидания и дисперсии. Именно поэтому с них мы и начнем.

По строгому определению, математическое ожидание любой детерминированной функции $q(x_1, x_2, \ldots, x_k)$ от k случайных аргументов, обозначается $M\left[q(x_1, x_2, \ldots, x_k)\right]$ и определяется следующим образом:

$$M\left[q(x_1, x_2, \ldots, x_k)\right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} q(x_1, x_2, \ldots, x_k) p(x_1, x_2, \ldots, x_k) dx_1 \ldots dx_k, \qquad (2.1)$$

где $p(x_1, x_2, \ldots, x_k)$ – k-мерная плотность вероятности. Операция $M[\ldots]$ означает *усреднение по ансамблю* реализаций и представляет собой интегрирование по диапазону возможных значений случайных величин с учетом вероятности получения данного значения.

Выражение (2.1) выглядит гораздо проще для одномерной плотности вероятности:

$$M\left[q(x)\right] = \int_{-\infty}^{\infty} q(x)\, p(x) dx. \qquad (2.2)$$

Математическое ожидание (2.1) является *начальным* моментом первого порядка. Для введения начальных моментов n-го порядка используется выражение вида:

$$M\left[X^n\right] = \int_{-\infty}^{\infty} x^n p(x) dx. \qquad (2.3)$$

Центральные моменты n-го порядка вводятся аналогичным образом:

$$M\left[\left(X - M\left[X\right]\right)^n\right] = \int_{-\infty}^{\infty} \left(x - M\left[X\right]\right)^n p(x) dx. \qquad (2.4)$$

Первый начальный момент (**мат. ожидание**) характеризует центр рассеяния данных, а второй центральный момент (**дисперсия**) – величину рассеяния данных.

Выражения (2.1) – (2.4) сами по себе уже достаточно применимы на практике, но проблема состоит в том, что ВР является только частной

выборкой случайной величины. Поэтому здесь необходимо использовать введенное в первой лекции понятие *эргодичности*, которое позволяет перейти от усреднения по реализациям к усреднению по одной достаточно длинной реализации (т.е. по времени). Тогда для определения мат. ожидания M_x эргодических случайных процессов можно использовать выражение (2.5), а для $D_x = \sigma_x^2$ дисперсии – (2.6):

$$M_x = \lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} x(t)\,dt = [x], \tag{2.5}$$

$$D_x = \sigma_x^2 = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \left(x(t) - M_x\right)^2 dt = \left[\left(x - [x]\right)^2\right]. \tag{2.6}$$

Здесь операция [..] уже является усреднением во времени, а не по ансамблю.

Стоит отметить, что все начальные моменты, в том числе и мат. ожидание, изменяются при прибавлении к случайной величине постоянного слагаемого. Центральные моменты, к которым относится дисперсия, зависят от выбранных единиц измерения. Чтобы устранить эти недостатки зачастую временные ряды изначально нормируют.

Любой временной ряд по определению представлен в дискретной форме и имеет ограниченный временной интервал. С учетом этих особенностей, начальные и центральные моменты не рассчитывают согласно (2.3) – (2.6), а *оценивают* дискретными приближениями.

Мат. ожидание можно оценить простым усреднением всех значений ВР:

$$M_x = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i\,, \tag{2.7}$$

где $x_i = x(t_i)$ – анализируемый временной ряд, N – количество наблюдений или размер выборки ВР.

Оценку дисперсии чаще всего производят по формуле:

$$D_x = \sigma^2 = \frac{1}{N} \sum_{i=0}^{N-1} \left(x_i - \bar{x}\right)^2\,, \tag{2.8}$$

которая является наименее чувствительной к ошибке округления.

Аналогичным образом дискретизируются формулы расчета начальных и центральных моментов *n*-го порядка. Среди статистических моментов высокого порядка следует выделить два важных нормированных момента: асимметрию *a* (третий центральный момент) и эксцесс *e* (четвертый центральный момент):

$$a = \frac{M\left[\left(x - \bar{x}\right)^3\right]}{\sigma^3}, \tag{2.9}$$

$$e = \frac{M\left[\left(x - \bar{x}\right)^4\right]}{\sigma^4}. \tag{2.10}$$

Асимметрия характеризует несимметричность распределения случайной величины, а эксцесс – степень выраженности хвостов распределения, то есть частоту появления удаленных от мат. ожидания значений. Нулевая асимметрия свидетельствует о симметричности распределения, лежащего в основе ВР. Если эксцесс какого-либо распределения превосходит таковой у гауссового (то есть более $3\sigma^4$), то это указывает на значительное число данных с большими амплитудами. Такие «хвосты» распределения случайной выборки считаются «толстыми», в сравнении с нормальным распределением.

Часть 2. Корреляционные характеристики

Подобно описанию центральных и начальных моментов из первой части, все корреляционные характеристики временных рядов изначально вводятся теоретически для некоторых случайных величин с заданной функцией плотности распределения $p(x)$.

Представим, что у нас есть две случайные величины X и Y, порождающие некоторые временные ряды. **Ковариацией** этих случайных величин X и Y называется интеграл:

$$R_{xy} = M\left[XY\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot y\cdot p\left(x,y\right)dxdy \qquad (2.11)$$

Из выражения (2.11) сразу видно, что, если средние значения этих случайных величин отличны от нуля, то значение ковариации может «уплыть» от ожидаемой величины. Поэтому на практике чаще используют центрированную версию ковариации, называемой **корреляцией** (2.12). При нулевых средних понятия ковариации и корреляции совпадают.

$$C_{xy} = M\left[\left(X - M\left[X\right]\right)\left(Y - M\left[Y\right]\right)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x-\overline{x}\right)\left(y-\overline{y}\right)p\left(x,y\right)dxdy \quad (2.12)$$

Корреляция и ковариация двух случайных величин показывает степень линейной зависимости этих величин друг от друга. Если случайные величины являются значениями одного и того же случайного процесса (что на самом деле и происходит на практике, так как значения выбранного ВР единственны на области его определения), то указанные функции (2.11) и (2.12) становятся **автоковариационной** (2.13) и **автокорреляционной** (2.14) функциями.

$$R_{xx} = M\left[X_1X_2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2p\left(x_1,x_2,\tau\right)dx_1dx_2\ , \qquad (2.13)$$

$$C_{xx} = M\left[\left(X_1 - M\left[X_1\right]\right)\left(X_2 - M\left[X_2\right]\right)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_1-\overline{x_1}\right)\left(x_2-\overline{x_2}\right)p\left(x_1,x_2,\tau\right)dx_1dx_2 \quad (2.14)$$

Эти статистические функции характеризуют меру статистической зависимости значений одной случайной величины, сдвинутых относительно

друг друга на интервал времени τ. Если вспомнить описание модели линейной авторегрессии (1.7) из лекции 1, то можно догадаться, что понятие автокорреляции будет играть существенную роль при построении оценок подобных моделей.

Аналогично формулам из первой части, приведем выражение (2.14) к его практической дискретной форме, с учетом свойств эргодичности и стационарности исследуемого ВР, а также с учетом конечной длительности выборки ряда. Пусть имеется некоторый ряд $Y(t)$, из него выбираются две последовательности отсчетов $Y(t_1), Y(t_2), ..., Y(t_m)$ и $Y(t_{1+l}), Y(t_{2+l}), ..., Y(t_{m+l})$, сдвинутых относительно друг друга на l моментов времени, то есть с **лагом** l. Тогда **оценка автокорреляционной функции** будет описываться в общем виде:

$$C(l) = \left[\left(Y(t_i) - \bar{Y} \right) \left(Y(t_i + l) - \bar{Y} \right) \right], \qquad (2.15)$$

где операция [..] усреднения по времени используется аналогично выражению (2.7) ранее.

Отметим некоторые важные свойства автокорреляционной функции. Во-первых, для стационарного ВР величина (2.15) зависит только от лага l, при этом $C(-l) = C(l)$, то есть она является симметричной относительно оси OY, поэтому достаточно построить автокорреляционную функцию только для $l > 0$. Во-вторых, при $l = 0$, легко заметить, что выражение (2.15) превратится в определение дисперсии, то есть $C(0) = D$. В-третьих, при этом $|C(l)| \le C(0)$ для любого значения лага l. В-четвертых, что достаточно важно для анализа временных рядов, если исходный случайный процесс содержит некоторую периодическую составляющую, то автокорреляционная функция C будет тоже периодической с таким же периодом. Если же таких периодических компонент нет, то $\lim_{l \to \infty} C(l) = 0$.

На практике, для оценки количественной степени взаимосвязи между значениями одной и той же реализации случайного процесса (то есть отсчетов

одного и того же ВР) с учетом указанных выше свойств автокорреляционной функции $C(0)=D$ и $|C(l)|\le C(0)$ гораздо удобнее рассматривать нормированные аналоги автокорреляционной функции:

$$\rho(l)=\frac{C(l)}{C(0)}=\frac{C(l)}{D}=\frac{C(l)}{\sigma_x^2}, \qquad (2.16)$$

называемой **коэффициентом автокорреляции**.

Определим свойства этого коэффициента аналогично описанным выше для функции автокорреляции. Очевидно, что $\rho(-l)=\rho(l)$, $\rho(0)=1$, $|\rho(l)|\le 1$.

С учетом выражений (2.15) и (2.16) запишем окончательную **оценку выборочного коэффициента автокорреляции**:

$$r(l)=\frac{(N-l)\sum_{i=1}^{N-l}y_i y_{i+l}-\left(\sum_{i=1}^{N-l}y_i\right)\cdot\left(\sum_{i=1}^{N-l}y_{i+l}\right)}{\sqrt{(N-l)\sum_{i=1}^{N-l}y_i^2-\left(\sum_{i=1}^{N-l}y_i\right)^2}\cdot\sqrt{(N-l)\sum_{i=1}^{N-l}y_{i+l}^2-\left(\sum_{i=1}^{N-l}y_{i+l}\right)^2}}, \qquad (2.17)$$

где $y_i=y(t_i)$ – анализируемый временной ряд, N – количество наблюдений или размер выборки ВР, l – сдвиг или лаг автокорреляции. Заметим, что с увеличением l число наблюдений в расчетах сокращается, поэтому l не должно быть большим (обычно достаточно $l\le N/4$).

Часть 3. Анализ автокорреляционной функции

Отметим важные свойства автокорреляционной функции и коэффициента автокорреляции, которые пригодятся нам в дальнейшем для анализа ВР.

Во-первых, для стационарного ВР с увеличением лага l взаимосвязь отсчетов ослабевает, и абсолютные значения коэффициента автокорреляции $\rho(l)$ (2.16) *должны убывать*. Это очень важное свойство, которое как раз в той или иной форме используется в статистических тестах для проверки ВР на стационарность. При этом надо понимать, что **оценка** выборочного коэффициента $r(l)$ (2.17) может нарушать свойство монотонного убывания, особенно при небольших значениях $N–l$. Получается, что более короткие ряды с гораздо большей вероятностью будут являться нестационарными.

Во-вторых, по величине коэффициента автокорреляции можно судить о наличии линейной (или близкой к ней) тенденции развития ВР, так как от него зависят весовые коэффициенты при построении авторегрессионных моделей.

В-третьих, по знаку коэффициента автокорреляции *нельзя* делать вывод о возрастающем или убывающем тренде ВР. Между ними нет никакой связи, однако подобное заблуждение оказывается распространенным на практике.

Последовательность коэффициентов автокорреляции $\rho(0),\rho(1),\rho(2),...$ будем называть **автокорреляционной функцией ВР**, а график зависимости значений $\rho(l)$ от лага l – **коррелограммой ВР**. Анализируя эти две характеристики, с учетом всех вышеперечисленных свойств функции автокорреляции, можно выявлять структуру исследуемого ВР в соответствие с аддитивной моделью (1.3), то есть наличие в нем тренда, сезонных составляющих и периодических составляющих.

Перечислим самые важные выводы из всех перечисленных свойств и определений:

1) Если наиболее высоким оказался коэффициент автокорреляции $\rho(1)$, то исследуемый ВР содержит в основном **тренд**. По-другому говорят,

что *тренд покрывает наибольший процент дисперсии* исходного ВР. Простым языком, в анализируемом ряде лучше всего виден тренд, на его фоне периодика не так хорошо заметна. Но это не значит, что периодики там нет совсем. Просто без удаления из исходного ВР тренда (а это возможно для аддитивной модели) искать сезонные и циклические компоненты будет очень сложно.

2) Если наиболее высоким оказался коэффициент автокорреляции $\rho(l)$, то ряд содержит колебания с периодичностью $T = l \cdot \Delta t$.
3) Если ни один из коэффициентов $r(l)$ не является значимым, то здесь может быть два предположения: в ряде нет ни тренда, ни периодики, по характеру он близок к белому шуму; либо в нем есть нелинейный тренд, который скрывает от наблюдателя все остальные компоненты.

Часть 4. Пример использования автокорреляционной функции

Рассмотрим пример применения методики анализа ВР на основе построения его коррелограммы и оценки выборочной автокорреляционной функции. Пусть задан ВР, значения которого по столбцам приведены в таблице 2.1. Этот ряд также изображен на рисунке 2.1.

Таблица 2.1.

Ряд, содержащий значения выхода партии химического процесса

1-7	8-14	15-21	22-28	29-35	36-42	43-49	50-56	57-63	64-70
47	41	58	57	25	58	45	52	34	60
64	59	44	50	59	45	57	38	35	39
23	48	80	60	50	54	50	59	54	59
71	71	55	45	71	36	62	55	45	40
38	35	37	57	56	54	44	41	65	57
64	57	74	50	74	48	64	53	38	54
55	40	51	45	50	55	43	49	50	23

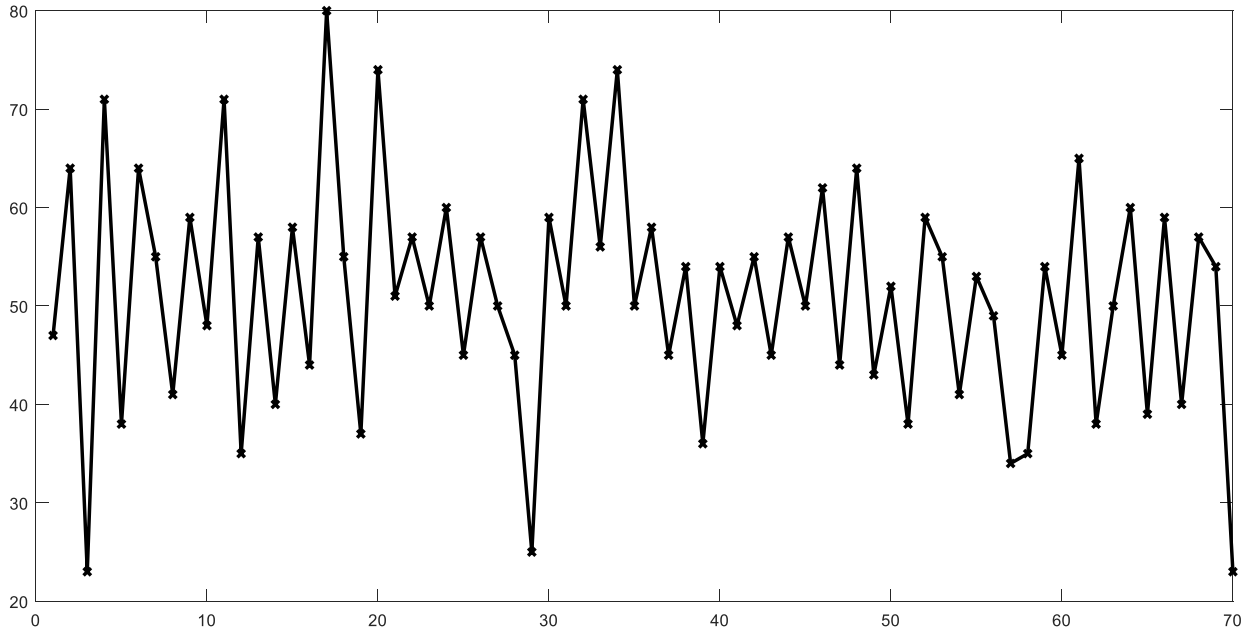

Рисунок 2.1 – Исходный временной ряд на основе табл. 2.1

Построим коррелограмму данного временного ряда. Для этого произведем расчет оценки выборочной автокорреляционной функции по формулам (2.16) и (2.17). Получившаяся выборочная автокорреляционная функция от лага представлена на рис. 2.2 ниже.

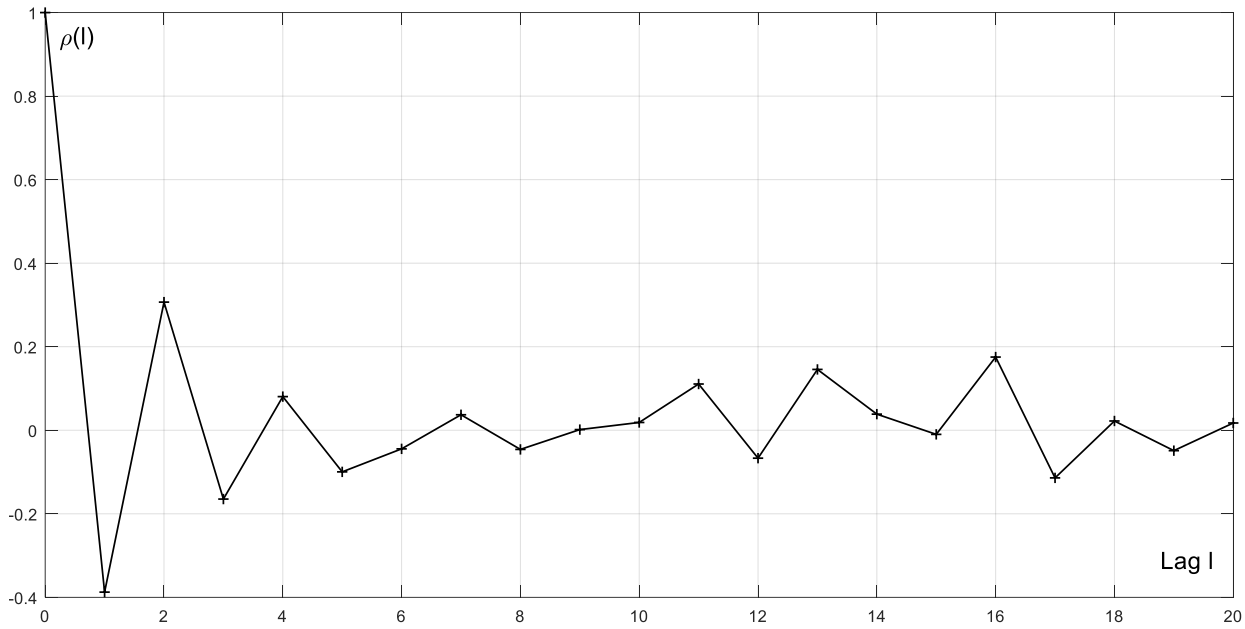

Рисунок 2.2 – Автокорреляционная функция ВР из рисунка 2.1

Из рисунка 2.2, то есть из коррелограммы, можно сделать несколько выводов. Во-первых, выборочная функция автокорреляции является знакопеременной с тенденцией к затуханию по мере роста лага/задержки. При этом затухание не является абсолютным, поэтому ряд нельзя однозначно отнести к стационарным ВР. Во-вторых, самым большим значением по модулю обладает $\rho(1)$, а значит в ряде есть тренд, который занимает значительную долю всей его структуры. Как видно из рисунка 2.3, примерные оценки линии тренда выдают его убывающий характер. В-третьих, после первого лага самым высоким значением обладает коэффициент $\rho(2)$. Это значит, что для каждой точки через одну наблюдается некоторая зависимость. В самом деле, если обратить внимание на ВР, то его значения зачастую «скачут» через одну точку: чередуются высокие значения с низкими, так

называемый «чередующийся» ряд. Знакопеременная автокорреляция также подтверждает это.

В целом, на примере этого ряда можно сделать вывод, что метод анализа ВР, опирающийся *только* на оценку выборочной автокорреляционной функции, не является хорошим и эффективным решением. В самом деле, многие сделанные выводы о характере ВР можно сделать просто глядя на сам исходный ряд. В этом и есть главный недостаток метода автокорреляционной функции – он не сможет показать того, что скрыто внутри исследуемого временного ряда от наблюдателя. Но в этом же и его преимущество – все, что является очевидным при визуализации исходного ряда, будет подтверждено его оценкой выборочной автокорреляционной функции. А все ложные системные особенности ряда, которые показались человеку при анализе ВР «на глаз», будут таким подходом отсекаться.

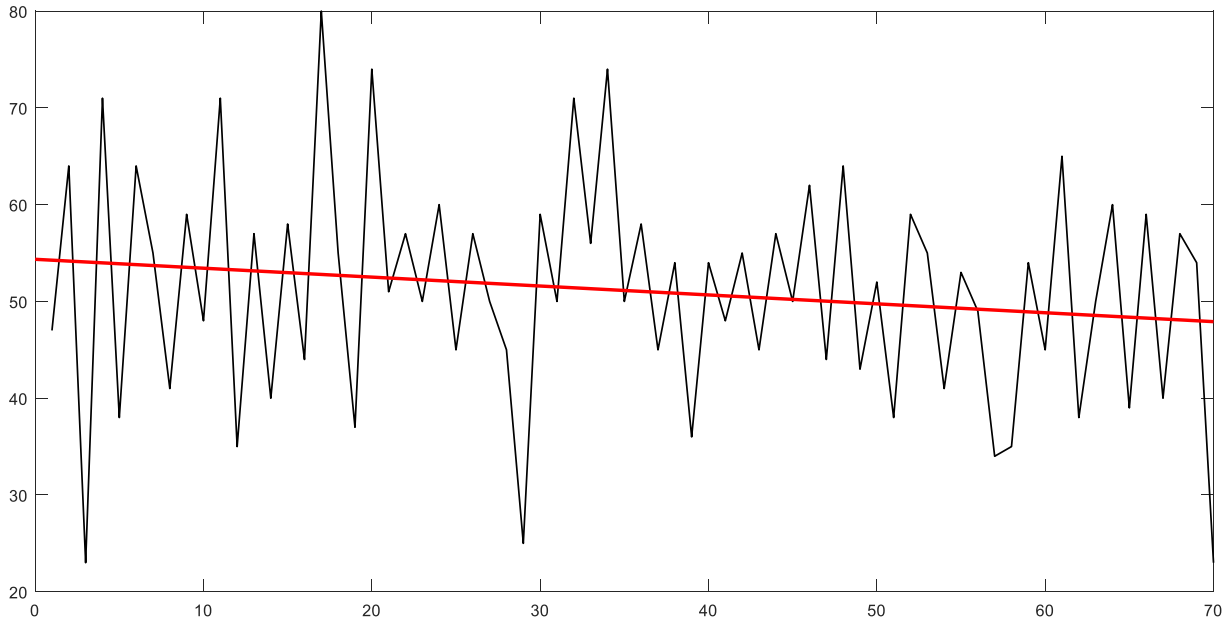

Рисунок 2.3 – Исходный ВР и его оценочный линейный тренд

Как мы увидим в дальнейшем, оценка выборочной автокорреляции относится к классу непараметрических методов, которые могут применяться

только для анализа *стационарных* ВР, так как все определения операций усреднения формулируются с учетом свойства эргодичности. Более того, на практике даже для стационарных ВР циклы могут иметь период, плавающий возле некоторого среднего значения. Оценка выборочной автокорреляции строится только от фиксированных значений задержки/лага, и зависимостей периода от момента времени мы здесь никак не получим. Для получения, так называемых, **частотно-временных характеристик** ряда, то есть зависимостей периода/частоты от времени, нам потребуется новый класс адаптивных методов, не требующих от исходного анализируемого ВР наличия стационарности даже в широком смысле этого определения.