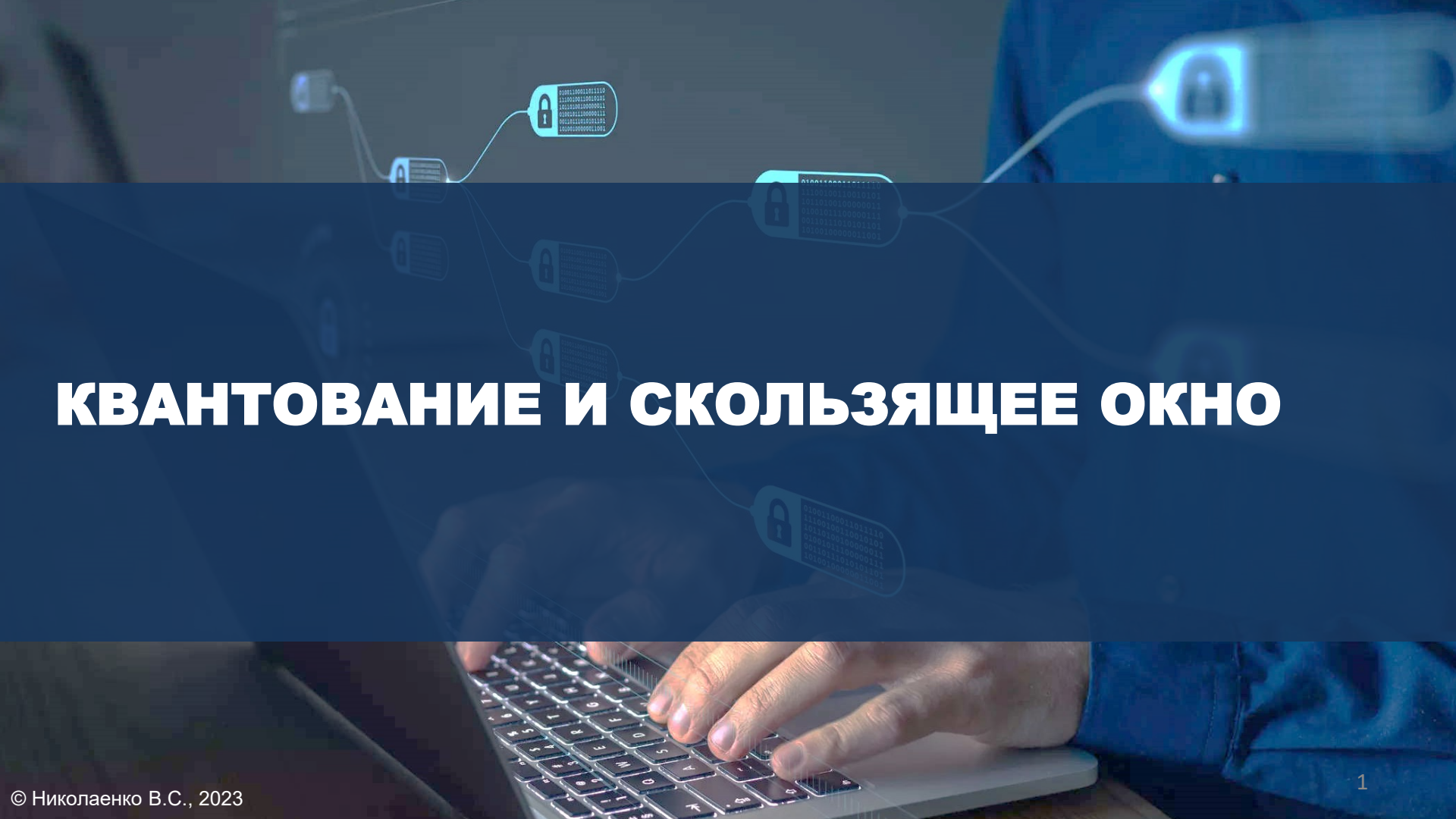


КВАНТОВАНИЕ И СКОЛЬЗЯЩЕЕ ОКНО

КВАНТОВАНИЕ

В основе **квантования** лежит процедура состоящая из двух шагов.

Диапазон значений, в пределах которого изменяется некоторая числовая величина (признак, показатель и так далее), разбивается на некоторое количество интервалов, каждому из которых присваивается определенный номер.

Эти интервалы называются **интервалами квантования**, а присвоенные им номера – **уровнями квантования**.

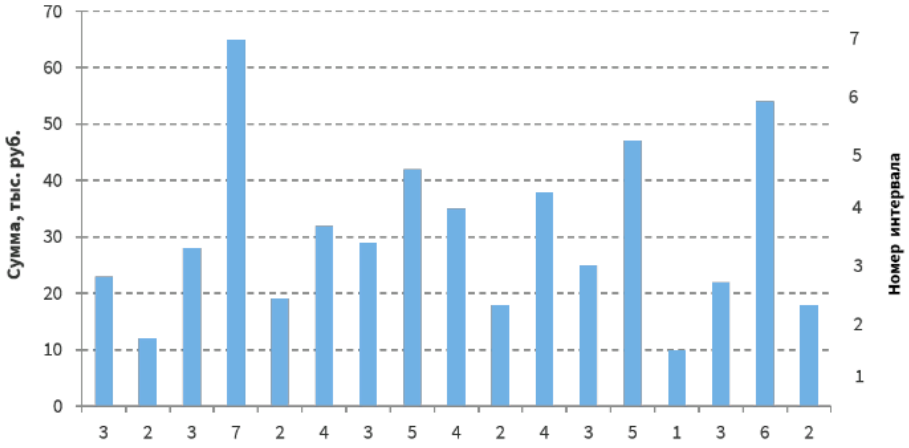
Каждое значение заменяется номером или меткой интервала квантования, в который попало данное значение.

КВАНТОВАНИЕ

Пусть наблюдаемый на рисунке ряд значений представляет собой суммы выданных кредитов. При этом минимальная сумма составляет **10 000 руб.**, а максимальная – **65 000 руб.**

Если диапазон значений ряда от **0** до **70 000 руб.**, то его можно разбить на 7 равных интервалов, взятых через **10 000**, по которым и будет проводиться **квантование**.

Для этого каждому интервалу будет присвоен порядковый номер, после чего все наблюдаемые значения будут заменены номерами интервалов квантования, в которые они попали. То есть вместо значения **23**, которое принадлежит третьему интервалу квантования, в результирующем наборе данных будет **3**, вместо значения **35** будет **4** и так далее.



КВАНТОВАНИЕ

Исходное значение	23	12	28	65	19	32	29	42	35	18	38	25	47	10	22	54	18
Квантованное значение	3	2	3	7	2	4	3	5	4	2	4	3	5	1	3	6	2

Итоговый результат преобразования представлен на слайде.

При квантовании необходимо определить, какую из границ интервала следует включить в этот интервал. Поскольку нижняя граница диапазона всегда принадлежит нижнему интервалу, то и для других интервалов можно условиться о включении нижней границы.

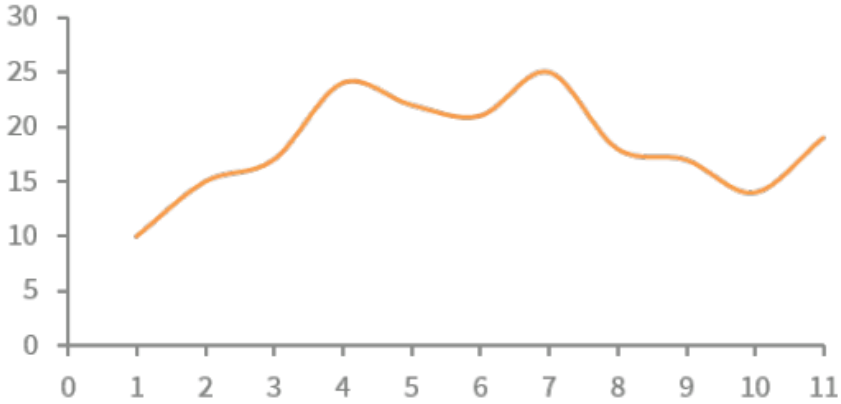
Единственным исключением является самый верхний интервал, который включает в себя как верхнюю, так и нижнюю границы.

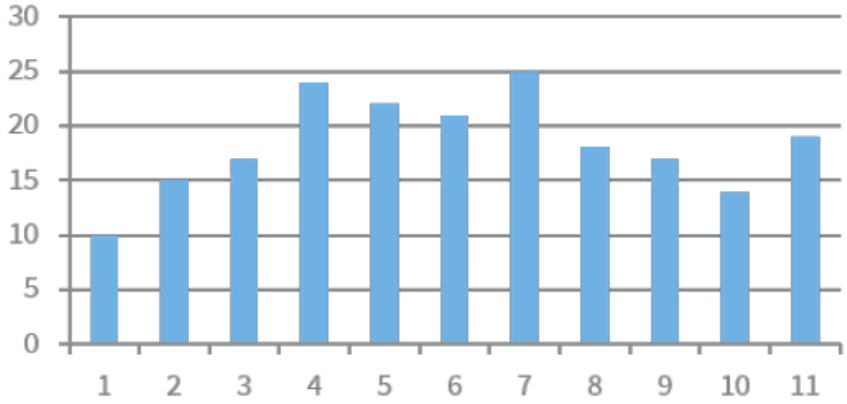
КВАНТОВАНИЕ

Рассмотрим подробнее где и как обычно используется квантование.

1. Квантование широко используется во всех областях, где возникает необходимость в обработке, передаче и хранения данных.
2. Квантование – неотъемлемая часть процесса преобразования **аналоговых** (то есть непрерывных по времени и амплитуде) сигналов в **цифровые** (то есть дискретные по времени и квантованные по амплитуде).
3. Квантование позволяет представлять и хранить данные в более компактном и защищенном от искажений виде.
4. Процесс дискретизации заключается в представлении непрерывной функции в виде набора отдельных значений, взятых в определенные моменты времени – **отсчеты**.
5. В результате квантования значения отсчетов преобразуется в номера интервалов квантования, в которые эти значения попали.

КВАНТОВАНИЕ





ЦЕЛИ КВАНТОВАНИЯ

В **бизнес-аналитике** квантование способствует достижению следующих целей:

- изменяется вид данных (из непрерывных они могут быть преобразованы в дискретные);
- сокращается размерность данных (уменьшается число разнообразных значений признака).

Например, если для анализа клиентов банка, получающих кредит, интерес представляют не **отдельные клиенты и суммы кредитов**, а **группы**, **объединяющие клиентов по интервалам сумм**, то в результате квантования можно получить более удобный для анализа ряд данных.

Уменьшение разнообразия значений признаков в некоторых случаях позволяет сделать работу моделей более эффективной. Действительно, если с точки зрения анализа нет разницы между суммами кредита **15** и **17 тыс.**, то нет смысла рассматривать эти величины отдельно.

КВАНТОВАНИЕ

В некоторых случаях представляет интерес использование в качестве результатов квантования не номеров интервалов, а других значений, связанных с них:

- **Нижняя граница интервалов**. Вместо значения, которое попало в интервал, устанавливается значение нижней границы.
- **Верхняя граница интервалов**. Вместо значения, которое попало в интервал, устанавливается значение верхней границы.
- **Среднее арифметическое интервала**. Вместо значения, которое попало в интервал, устанавливается его срединное значение. Границы и середину интервала удобно применять в тех случаях, когда квантованный ряд значений должен сохранять количественное выражение исходных данных. Однако в этом случае результирующее поле по-прежнему останется непрерывным и не сможет быть использовано в качестве выходного в классификационной модели. Преимуществом использования данного метода будет сокращение разнообразия значений признака.
- **Метка интервала**. Пользователь может задать произвольное значение, которое обозначат интервал, например, наименование категории, к которой будет относится объект классификации.

РЕЗУЛЬТАТ КВАНТОВАНИЯ

Использование меток интервалов дает возможность сделать результаты квантования более наглядными и сразу определить метки классов, если целью квантования является разбиение признака по категориям. Так, если цель квантования поля **Сумма кредита** – разделить всех клиентов на категории в зависимости от взятой ими суммы, то можно использовать соответствующие метки, как это демонстрирует таблица.

На практике этот набор данных можно использовать как обучающую выборку для построения классификационной модели, где в качестве целевого поля будет использоваться **Категория клиента**.

Срок кредита	Возраст	Пол	Образование	Сумма кредита	Категория клиента
6	37	Жен	Специальное	7 000	Категория 1
6	38	Муж	Среднее	7 500	Категория 1
12	60	Муж	Высшее	14 500	Категория 2
6	28	Муж	Специальное	15 000	Категория 2
12	59	Жен	Специальное	32 000	Категория 4
6	25	Жен	Специальное	11 500	Категория 1
6	57	Муж	Специальное	5 000	Категория 1
30	29	Муж	Высшее	61 500	Категория 7
12	37	Муж	Специальное	13 500	Категория 2
18	36	Муж	Специальное	25 000	Категория 3
24	68	Муж	Высшее	25 500	Категория 3
6	20	Жен	Высшее	9 500	Категория 1

ВЫБОР ЧИСЛА ИНТЕРВАЛОВ КВАНТОВАНИЯ

В квантовании важно правильно выбрать **число интервалов**.

Так как в результате квантования осуществляется переход от точных данных к некоторой интервальной оценке, неизбежна потеря информации. Фактически ряд значений, полученных в результате квантования, просто выражает отношения между исходными значениями признака.

То, что два значения расположены в двух соседних интервалах квантования, не позволяет точно определить, насколько одно из них больше или меньше другого. Можно сказать только, что они не различаются больше, чем на две **ширины интервала**.

Ширина интервала представляет собой разницу между верхней и нижней границами интервала. Следовательно, **чем больше интервалов используется при квантовании, тем точнее представление исходных значений данных**. При уменьшении ширины интервала в пределе мы получим исходный набор значений. Увеличение интервала, напротив, огрубляет описание данных и в пределе дает один интервал для всего диапазона значений, которые меняются на метку интервала (например, 0).

ВЫБОР ЧИСЛА ИНТЕРВАЛОВ КВАНТОВАНИЯ

Иными словами, меняя число интервалов квантования, можно перейти от точного воспроизведения исходных значений данных к полной потере информации об изменчивости значений признака.

На практике выбрать количество интервалов квантования можно исходя из следующих соображений:

1. Если квантование выполняется для преобразования непрерывных данных в дискретные, то число интервалов будет определяться числом уникальных значений (меток, категорий), которые используются при решении задачи анализа.
2. Необходимо учитывать требуемую точность описания данных. Например, может быть поставлено условие, что количество интервалов квантования должно быть таким, чтобы ширина интервала не превышала 10% от полного диапазона изменения исходных значений.
3. Иногда может потребоваться проведение экспериментов, чтобы определить лучшие параметры квантования с сточки зрения решения конкретной задачи анализа.

ВЫБОР ЧИСЛА ИНТЕРВАЛОВ КВАНТОВАНИЯ

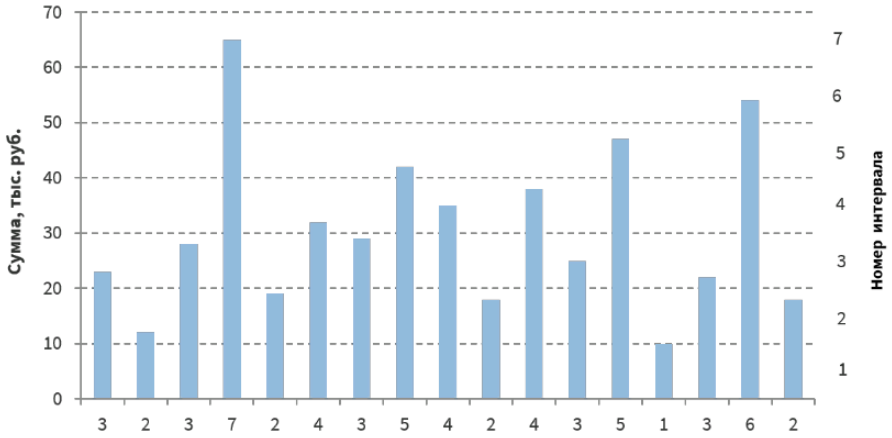
Кроме выбора числа интервалов, при выполнении операции квантования требуется выбрать ее **метод**. Выбор метода квантования зависит от характера данных. Различают два основных метода квантования.

МЕТОД КВАНТОВАНИЯ: РАВНОМЕРНОЕ (ОДНОРОДНОЕ) КВАНТОВАНИЕ

При **равномерном квантовании** диапазон изменения значений признака разделяется на интервалы одинаковой ширины. Количество интервалов в таком случае может задаваться **явно**, тогда ширина интервала рассчитывается как отношение разницы между верхней и нижней границами диапазона к заданному количеству, либо **неявно**, с помощью задания ширины интервала.

Во втором случае количество рассчитывается как отношение разницы между верхней и нижней границами диапазона к заданной ширине.

На рисунке представлен пример равномерного квантования, где диапазон значений поделен на **7** интервалов, ширина каждого из которых составляет **10 тыс. руб.**



МЕТОД КВАНТОВАНИЯ: НЕРАВНОМЕРНОЕ (НЕОДНОРОДНОЕ) КВАНТОВАНИЕ

При **неравномерном квантовании** ширина интервалов может быть различной. Здесь также есть несколько способов распределения значений признака. Например, при **плиточном** квантовании ширина интервалов выбирается таким образом, чтобы в каждый из них попало примерно одинаковое количество значений.

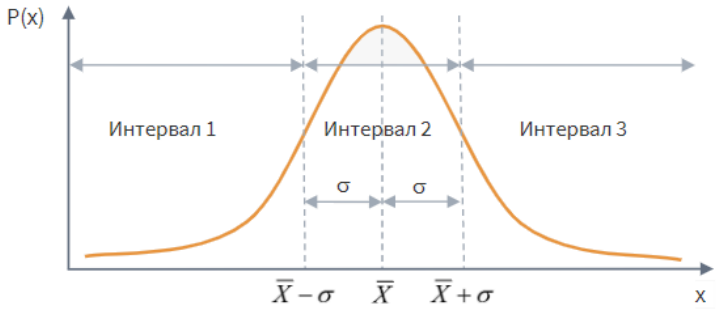
МЕТОД КВАНТОВАНИЯ: КОЭФФИЦИЕНТЫ СКО

Еще один способ неравномерного квантования – **коэффициенты СКО** (среднеквадратического отклонения).

При его использовании формируются интервалы с шириной, кратной среднеквадратическому отклонению значений признака.

Границы интервалов будут рассчитываться на основе вычисленных математического ожидания $\bar{X}$ и среднеквадратического отклонения σ, например, если используется одно среднеквадратическое отклонение, то формируется три интервала:

$$x < \bar{X} - \sigma, \bar{X} - \sigma \le x \le \bar{X} + \sigma, x > \bar{X} + \sigma$$



ПРИМЕР РАВНОМЕРНОГО КВАНТОВАНИЯ

Равномерное квантование используется, если данные равномерно распределены по всему диапазону их изменения, то есть в результате квантования не будет интервалов, в которых значения почти отсутствуют или заполнены очень плотно.

В противном случае лучшие результаты дадут неравномерное квантование. Рассмотрим это на примере.

Пусть дан набор значений {2, 4, 3, 1, 4, 22, 24, 23, 21, 24} диапазон изменения его значений 23.

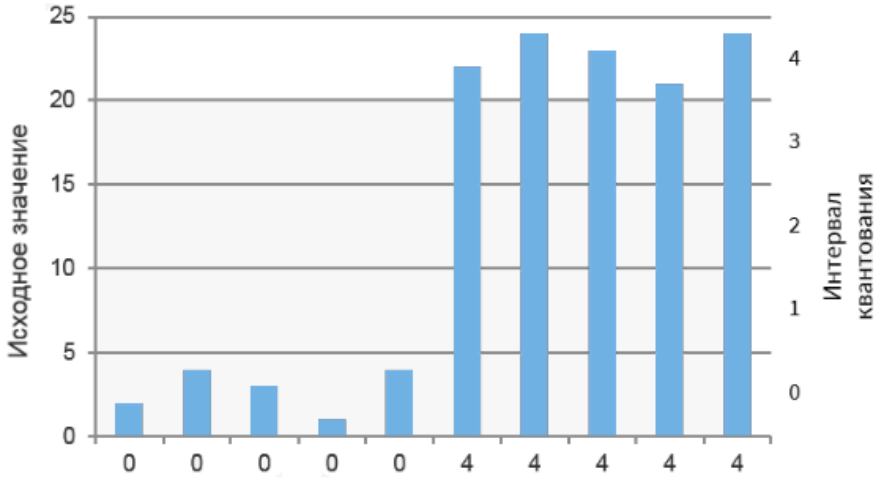
Разделим диапазон значений на 5 равных интервалов квантования, каждый из которых будет содержать 5 значений.

Интервал с номером 0 будет содержать значения от 0 до 4, с номером 1 – от 5 до9, с номером 2 – от 10 до 14 и так далее.

На рисунке видно, что значения в указанном диапазоне распределены **неравномерно**: в нулевом интервале квантования расположено 5 значений, а остальные значения – в диапазоне 20-24, который соответствует интервалу квантования с номером 4.

При этом в диапазон 5-20, которому соответствуют интервалы квантования с номерами 1, 2 и 3, не попало ни одного значения.

Следовательно, квантованные значения также распределяются неравномерно: интервалы 0 и 4 будут заполнены очень плотно, в то время как интервалы 1, 2 и 3 окажутся пустыми.



ПРИМЕР РАВНОМЕРНОГО КВАНТОВАНИЯ

Если квантование проводится для преобразования непрерывных данных в набор категорий, это приведет к тому, что все объекты выборки окажутся отнесенными всего к двум категориям – 0 и 4.

Исходное значение	Квантованное значение
2	0
4	0
3	0
1	0
4	0
22	4
24	4
23	4
21	4
24	4

ПРИМЕР НЕРАВНОМЕРНОГО КВАНТОВАНИЯ

Преодолеть данную проблему позволяет **неравномерное квантование**, когда используются интервалы разной ширины так, чтобы в каждый из них попало примерно одинаковое количество значений. Такое квантование называют **плиточным**.

ПРИМЕР НЕРАВНОМЕРНОГО КВАНТОВАНИЯ

Если его применить к приведенному примеру приведенному на рисунке, то получим следующие результаты. Как видно из правой таблицы, заполнение интервалов квантования получилось более равномерным, чем при обычном равноинтервальном квантовании.

Выбор интервалов в плиточном квантовании

№	Диапазон значений	Метка
0	От 0 до 2	0
1	От 3 до 4	1
2	От 5 до 21	2
3	От 22 до 23	3
4	От 23 до 23	4
5	От 24	5

1

Результат квантования

Исходное значение	Квантованное значение
2	0
4	1
3	1
1	0
4	1
22	3
24	5
23	4
21	2
24	5

2

ПРЕОБРАЗОВАНИЕ УПОРЯДОЧЕННЫХ ДАННЫХ

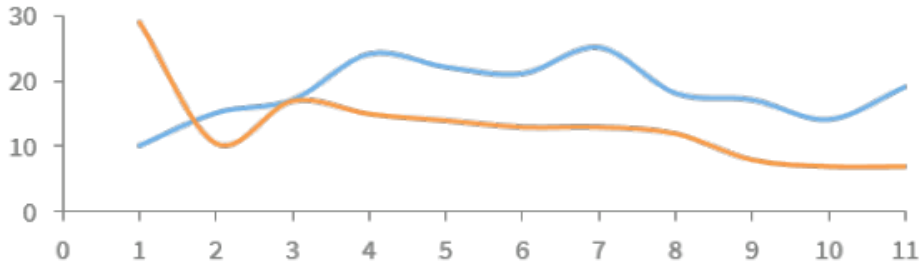
Следующая тема – это операции по **преобразованию упорядоченных данных**.

Многие аналитические задачи, например, анализ продаж, связаны с обработкой данных, которые зависят от времени. Такие данные называют **упорядоченными**, или **временными рядами**.

В процессе обработки временных рядов требуется специальная подготовка данных, чтобы оптимизировать их представление для всех возможных интервалов даты и времени. Это необходимо для решения определенных аналитических задач, в частности:

- прогнозирование;
- классификация состояний объектов;
- выявление закономерностей, объясняющих динамику бизнес-процессов.

ПРЕОБРАЗОВАНИЕ УПОРЯДОЧЕННЫХ ДАННЫХ



Временной ряд состоит из последовательности наблюдений за состоянием параметров (признаков) исследуемых объектов или процессов. Если наблюдения содержат один признак, то ряд является одномерным, а если два или более – многомерным.

Поскольку значения временного ряда определены только в фиксированные моменты времени, так называемые **отсчеты**, последовательность его значений может быть представлена в следующем виде:

$$X = \{x_1, x_2, \dots x_n\}$$

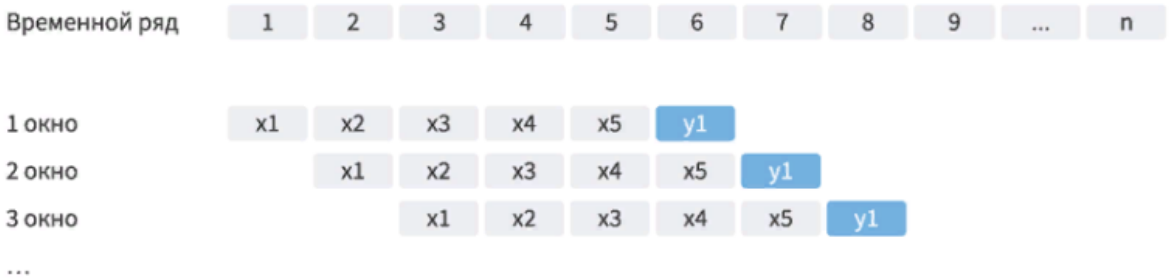
ПРЕОБРАЗОВАНИЕ УПОРЯДОЧЕННЫХ ДАННЫХ

Целью преобразований временных рядов является не изменение их содержания, а представление информации таким образом, чтобы обеспечивалась максимальная эффективность решения определенной задачи анализа.

Можно выделить два основных типа преобразования, которые наиболее часто используются при подготовке временных рядов к анализу:

- **Преобразование даты и времени**. Преобразование даты и времени заключается в приведении даты и времени к виду, наиболее удобному для визуального анализа и обработки временного ряда. При этом результаты преобразования даты не обязательно являются значениями типа **Дата/Время** и могут обрабатываться как обычные числа и строки.
- **Скользящее окно**. Скользящее окно применяется при решении задач прогнозирования и классификации состояний бизнес-объектов, чтобы преобразовывать последовательность значений ряда в таблицу, которую можно использовать для построения моделей или какой-либо другой обработки. С первым типом мы уже познакомились в первой лекции, поэтому далее подробно рассмотрим только **скользящее окно**.

СКОЛЬЗЯЩЕЕ ОКНО



Скользящее окно широко применяется при обработке временных рядов, например, чтобы построить модель прогноза временного ряда, или когда требуется при обработке набора данных сравнивать соседние значения – лучше, чтобы они располагались в столбцах, а не в строках.

Целью прогнозирования значений временного ряда является предсказание значения **x(n+1)**, на основе предыдущих значений признака. Решение задачи прогнозирования возможно только в том случае, если значения временного ряда связаны между собой.

СКОЛЬЗЯЩЕЕ ОКНО

Например, пусть при регистрации заявок, поступивших от клиентов, среди прочего фиксируются дата подачи заявки и адрес клиента. Из базы данных системы регистрации заявок можно извлечь временной ряд с последовательностью номеров квартир, указанных в адресах.

Очевидно, что пытаться предсказать номер квартиры следующего клиента на основе знания номеров квартир клиентов, чьи заявки были зарегистрированы ранее, бессмысленно.

СКОЛЬЗЯЩЕЕ ОКНО

Скользящее окно оперирует следующей терминологией анализа и прогнозирования временных рядов:

- **Интервал прогноза**. Временной интервал, на котором будет осуществляться прогнозирование (день, неделя, месяц, квартал, год).
- **Горизонт прогноза**. На какое количество интервалов (дней, недель и др.) вперед мы ходим получить прогноз.
- **Глубина истории**. Количество значений интервалов прогноза в прошлом, которое мы будем использовать для предсказания значений интервалов в будущем.

СКОЛЬЗЯЩЕЕ ОКНО

Пусть имеется ряд данных, содержащий 11 наблюдений:

$$X = \{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}.$$

Если глубину истории задать равной 5, то с помощью скользящего окна можно преобразовать исходный ряд данных в табличную форму, состоящую из 6 записей, которая может быть использована для дальнейших вычислений.

Таким образом, основная задача скользящего окна – преобразование ряда данных в таблицу, где каждая запись представляет собой наблюдение, сформированное из некоторого интервала ряда.

№	x_{n-5}	x_{n-4}	x_{n-3}	x_{n-2}	x_{n-1}	x_n	x_{n+1}	x_{n+2}
1	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
2	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
3	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
4	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
5	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
6	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}

СКОЛЬЗЯЩЕЕ ОКНО

Дата начала месяца	Число клиентов
01.01.2017	1 540
01.02.2017	960
01.03.2017	1 150
01.04.2014	1 230
01.05.2017	1 056
01.06.2017	995

Дата начала месяца	X_{n-2}	X_{n-1}	X_n	X_{n+1}
				1 540
01.01.2017			1 540	960
01.02.2017		1 540	960	1 150
01.03.2017	1 540	960	1 150	1 230
01.04.2014	960	1 150	1 230	1 056
01.05.2017	1 150	1 230	1 056	995
01.06.2017	1 230	1 056	995	
	1 056	995		
	995			

Дата начала месяца	X_{n-2}	X_{n-1}	X_n	X_{n+1}
01.03.2017	1 540	960	1 150	1 230
01.04.2014	960	1 150	1 230	1 056
01.05.2017	1 150	1 230	1 056	995

Рассмотрим еще пример. Пусть имеется ряд наблюдений, который отражает количество клиентов, обслуженных за месяц (таблица 1).

Задача состоит в том, чтобы построить модель прогноза числа клиентов на будущий месяц. При этом глубина погружения задается равной 2, горизонт прогнозирования – 1, то есть на основе двух предыдущих месяцев прогнозируется следующий.

Выборка, полученная в результате обработки данных скользящим окном, будет содержать в начале и в конце неполные записи, количество которых будет равно глубине погружения (таблица 2).

Неполные записи можно исключить из рассмотрения. Если это сделать, то результирующая выборка будет иметь вид, представленный в таблице 3.