АДЫГЕЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ADYGHE STATE UNIVERSITY

Математика для всех

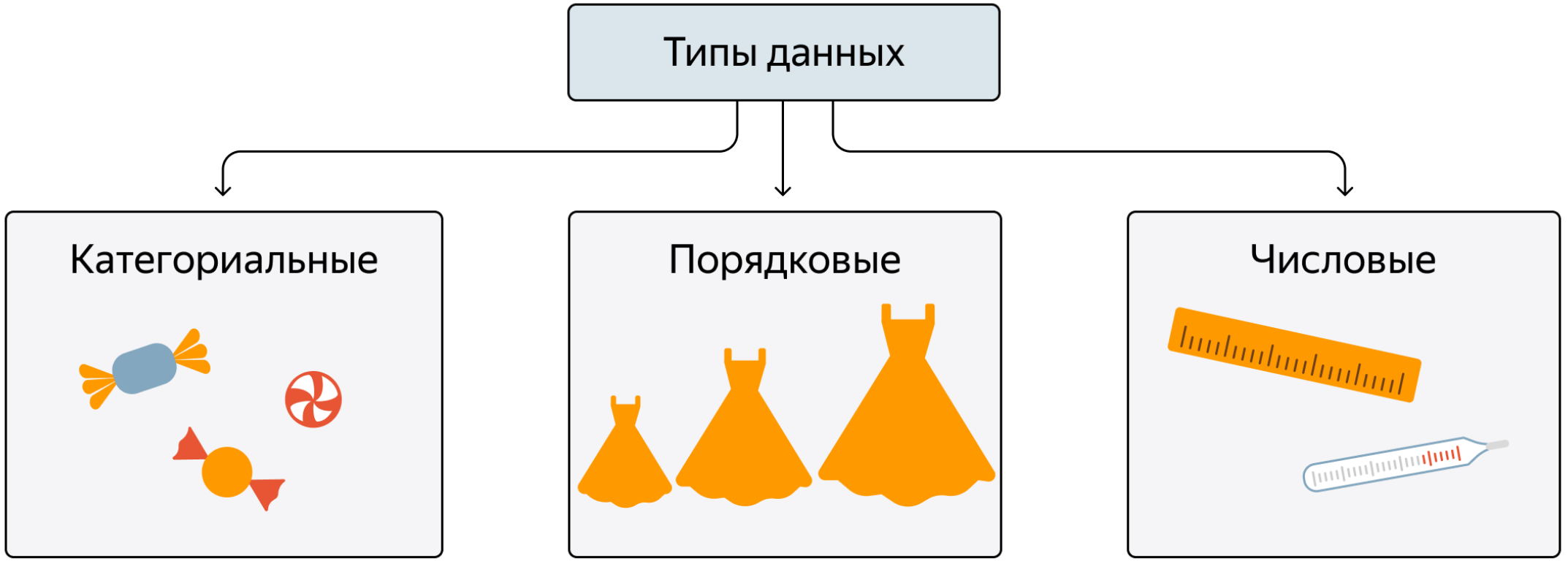
Семинар 5
Основы статистики

Карпенко Юрий Александрович

Институт точных наук и цифровых технологий
Кафедра алгебры и геометрии

Описательные статистики

Типы данных

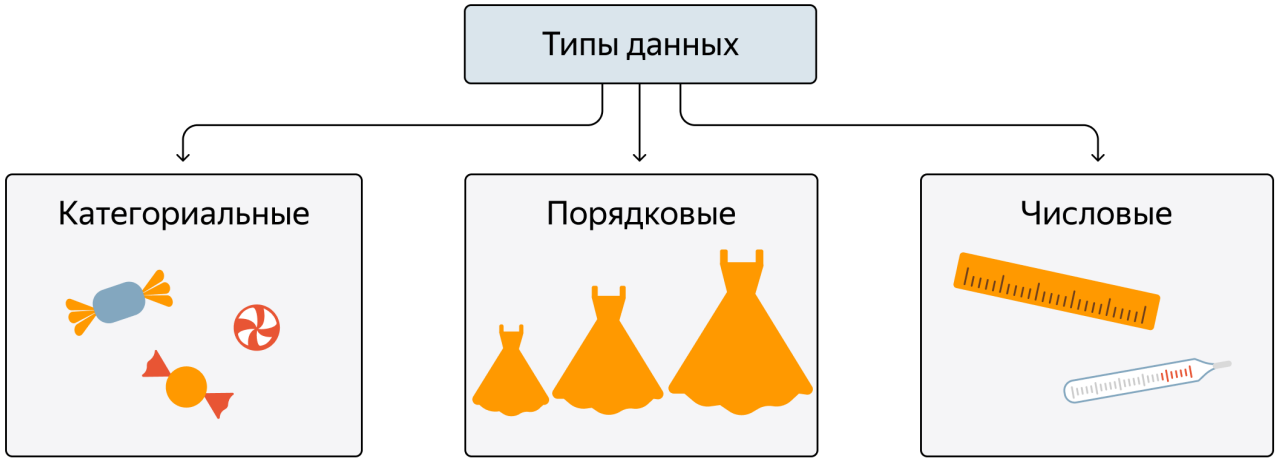


Категориальный тип данных

Категориальный тип описывает данные, которые могут принимать ограниченное количество уникальных значений. Например:

- цвет глаз — может быть серым, голубым, зелёным или карим, это и есть категории;
- тип кондитерской продукции — конфеты, пирожные, торты;
- конфеты — бывают с арахисом, вишней, лимоном.

Особенность этого типа в том, что нельзя сравнить значения измерений или произвести любую математическую операцию над ними. Например, нельзя сказать, какой цвет глаз больше — зелёный или серый. Или вычесть из конфеты пирожное.



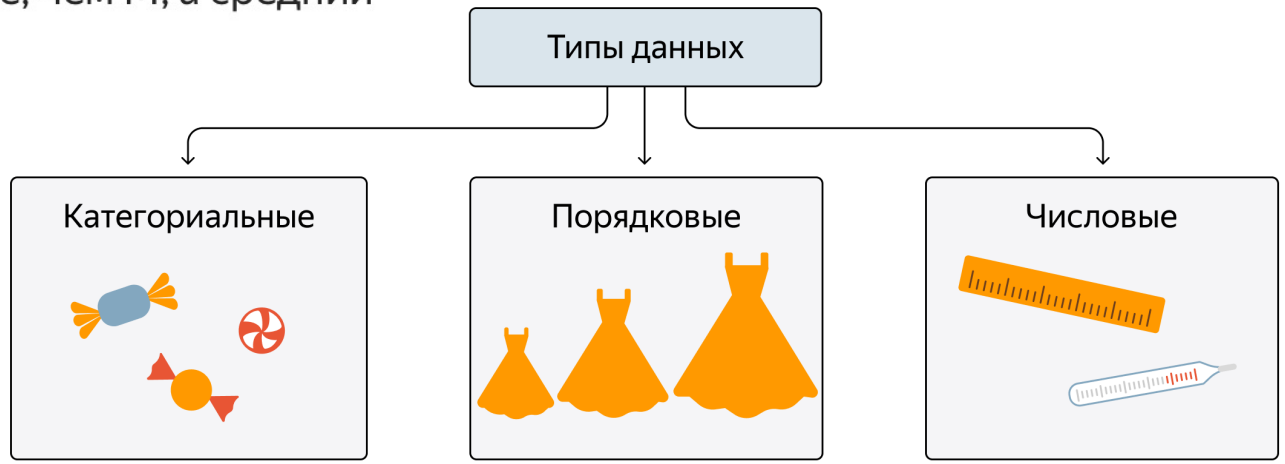
Порядковый тип данных

Порядковый тип похож на категориальный. Он тоже описывает данные, которые принимают ограниченное количество уникальных значений, но при этом для них есть естественный порядок.

Например:

- Размер одежды: XS, S, M, L, XL, XXL.
- Уровень образования: начальное, среднее, высшее.

Порядковые данные, как и категориальные, — это качественные характеристики, и к ним нельзя применить арифметические операции. Особенность этого типа в том, что значения можно сравнить: размер одежды L больше, чем M, а средний уровень образования выше начального.

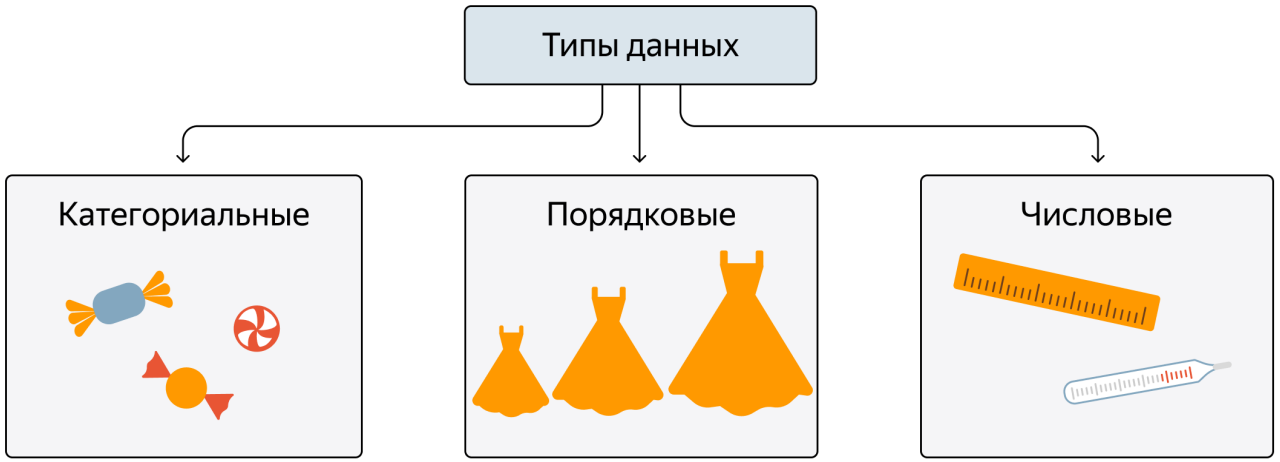


Числовой тип данных

Числовой тип описывает данные, которые можно измерить. Например:

- количество работников в компании,
- объём продаж за месяц,
- температура в цехе,
- процент сахара в ирисках.

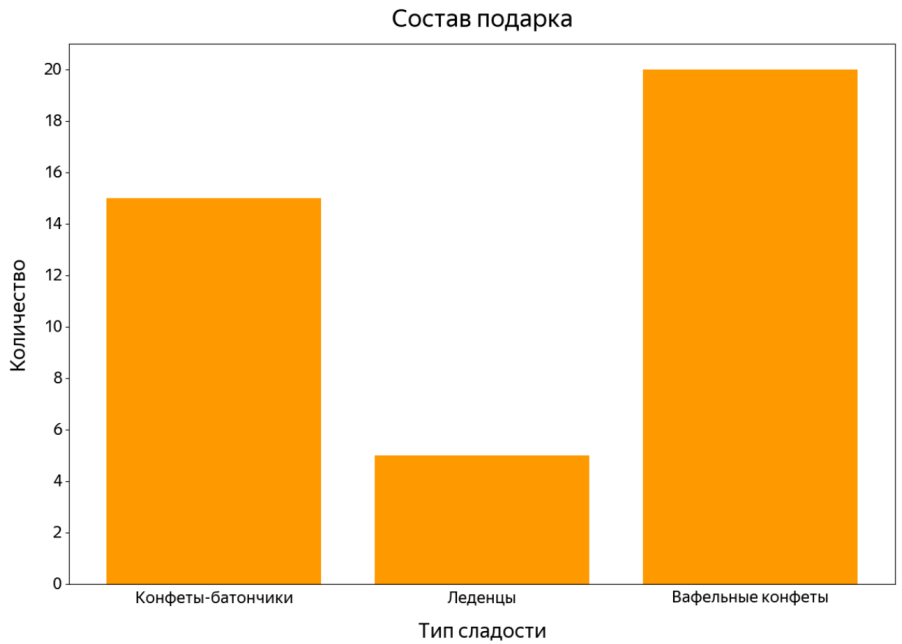
Можно и сравнивать числовые данные, и выполнять над ними арифметические операции.

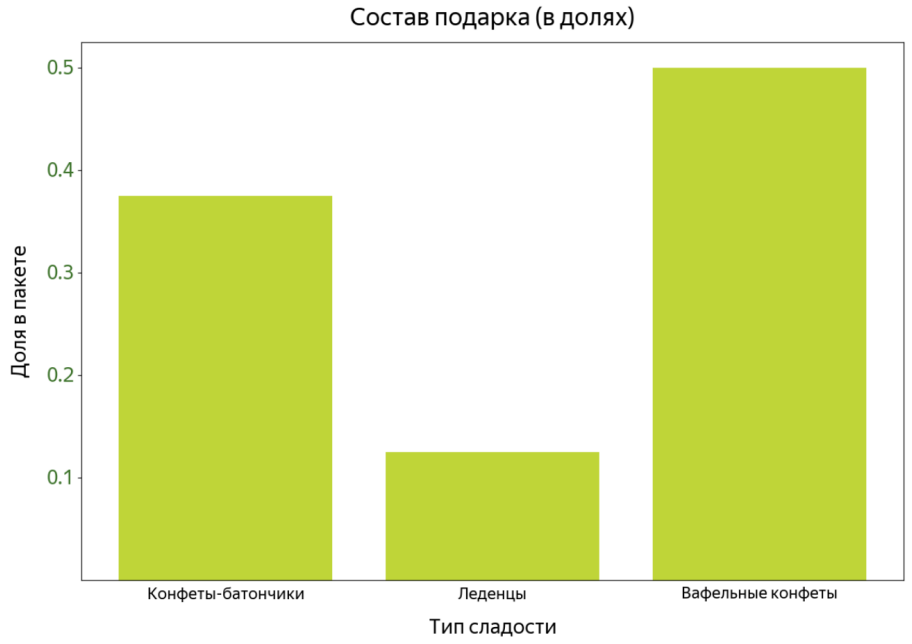


Визуализация данных

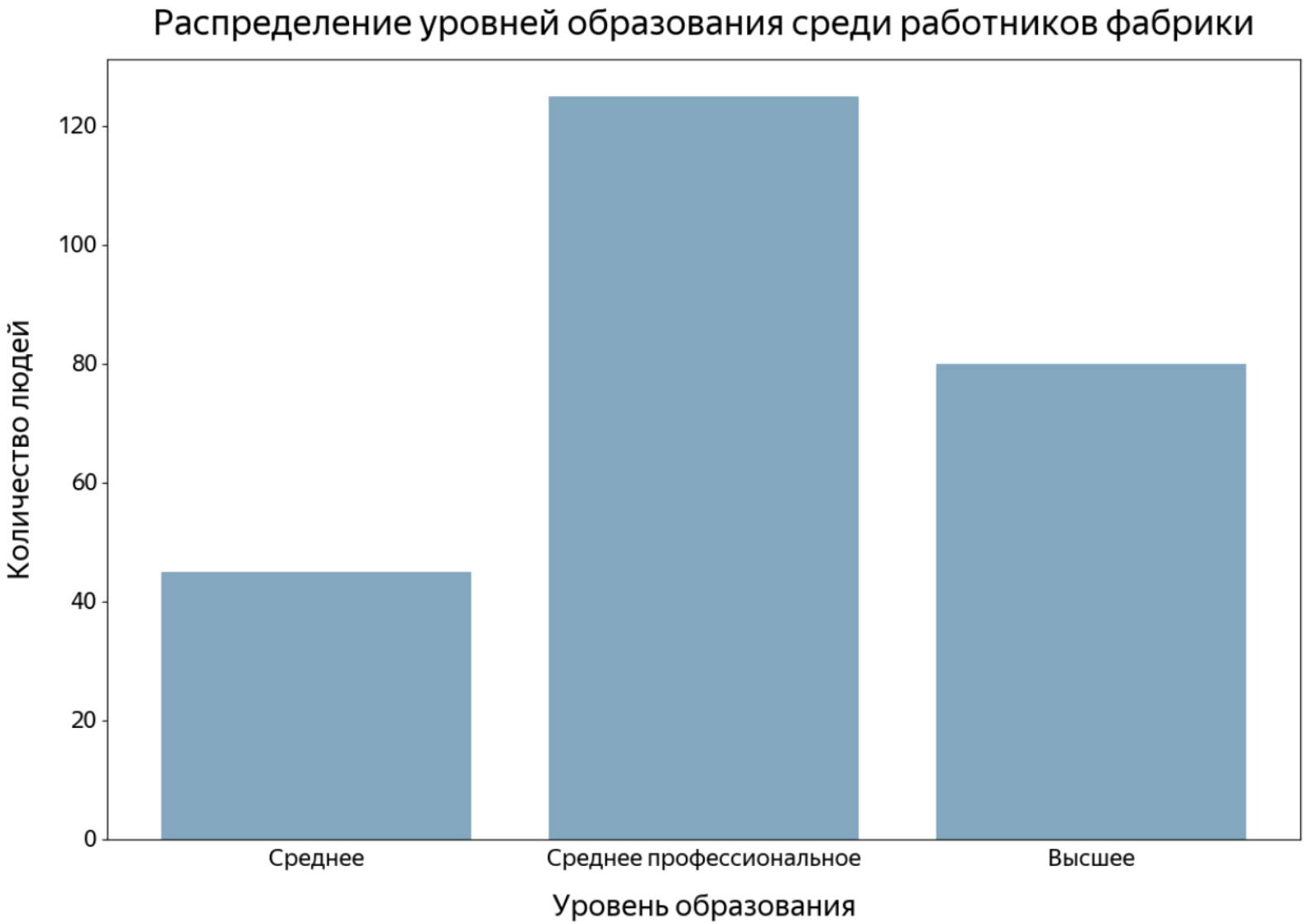
Для категориальных и порядковых данных обычно считают количество объектов в каждой категории или определяют долю каждой категории от общего числа наблюдений. Такие данные хорошо описываются с помощью **столбчатых диаграмм (барплотов)**.

По горизонтали показываем категории, а по вертикали — числовые значения. Примеры — в галерее, листайте.





Визуализация данных

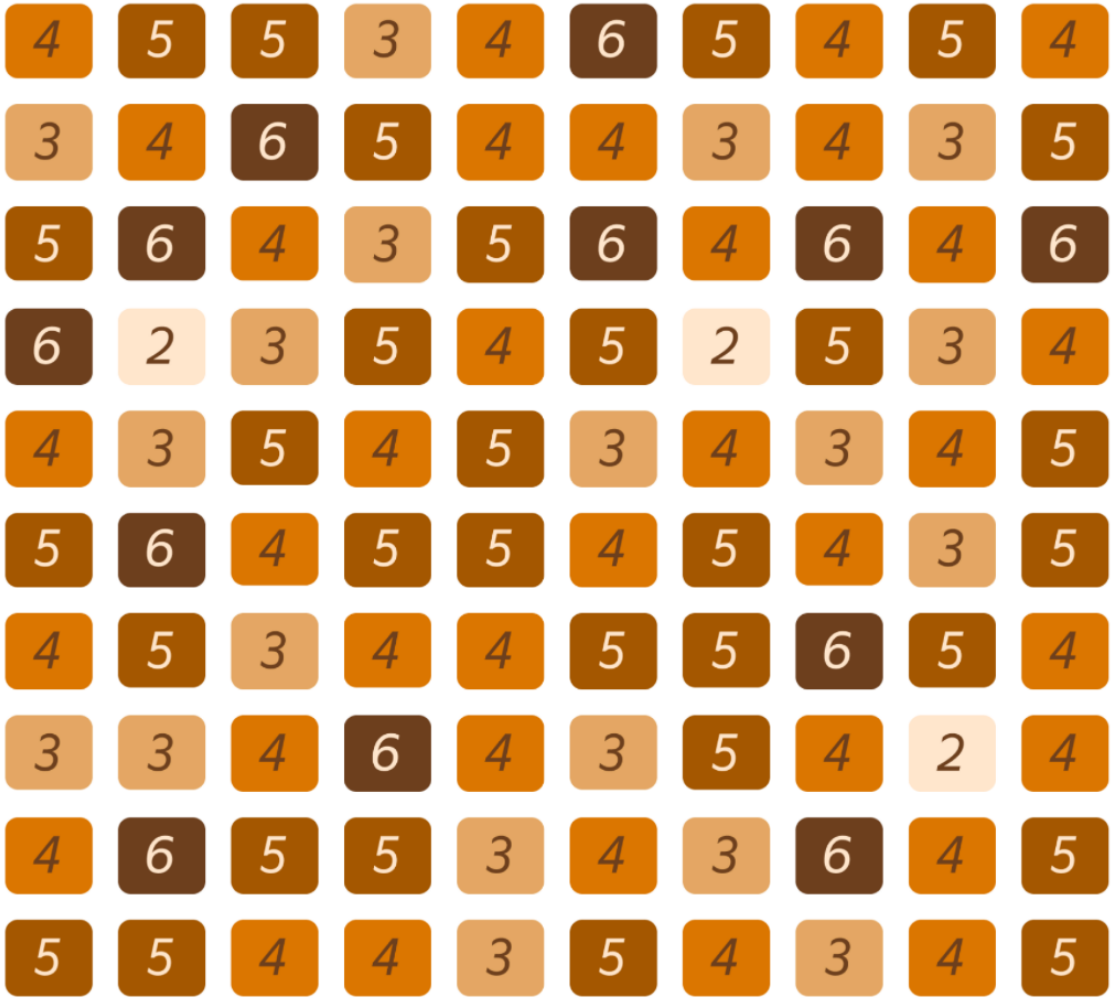


Визуализация числовых данных

Для примера возьмём 100 ирисок «Первичный ключик», исследуем содержание в них сахара. Измерим количество сахара с точностью до грамма в каждой ириске из набора:

Визуализация числовых данных

Для примера возьмём 100 ирисок «Первичный ключик», исследуем содержание в них сахара. Измерим количество сахара с точностью до грамма в каждой ириске из набора:

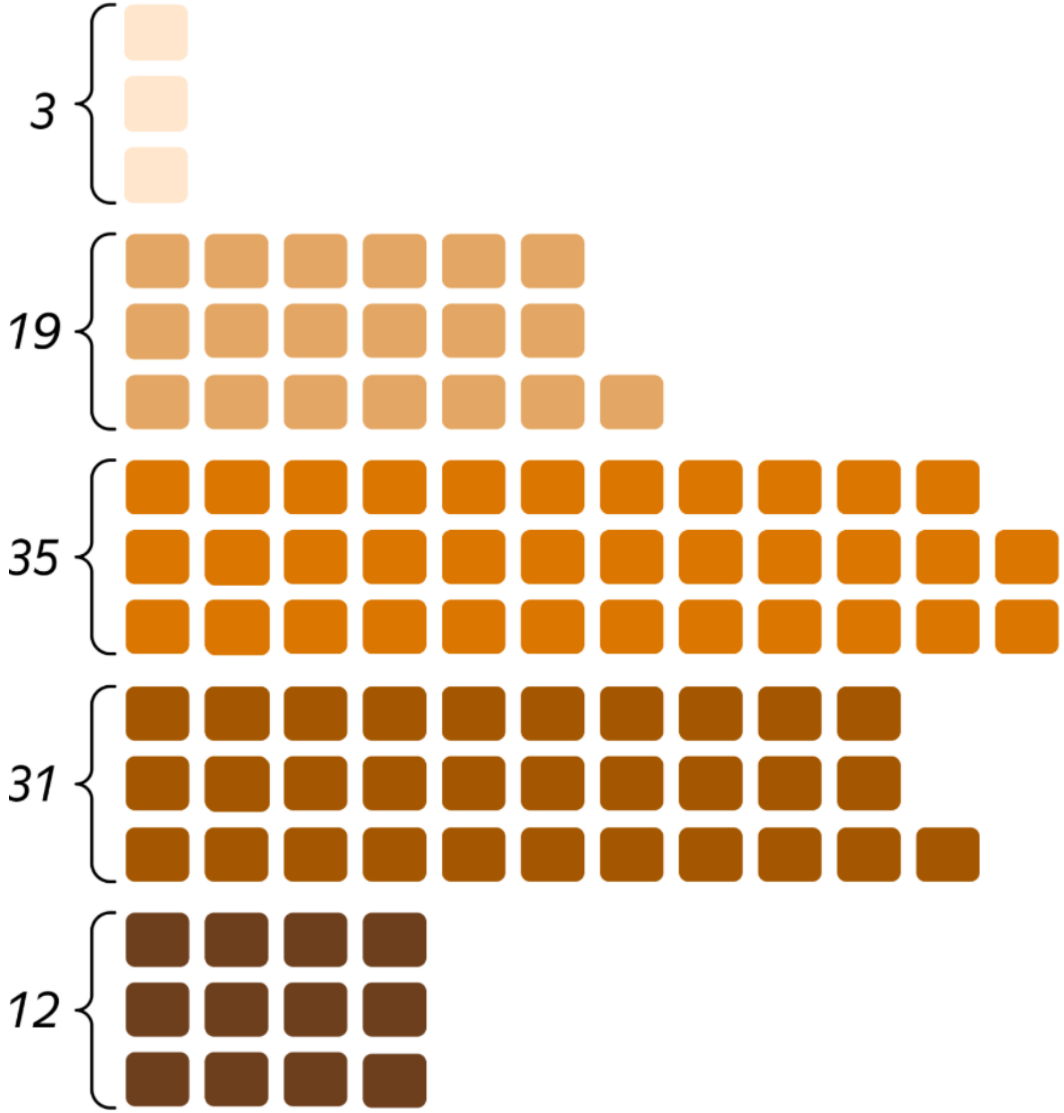


Ириски лежат хаотично. По данным трудно сделать какие-то выводы.

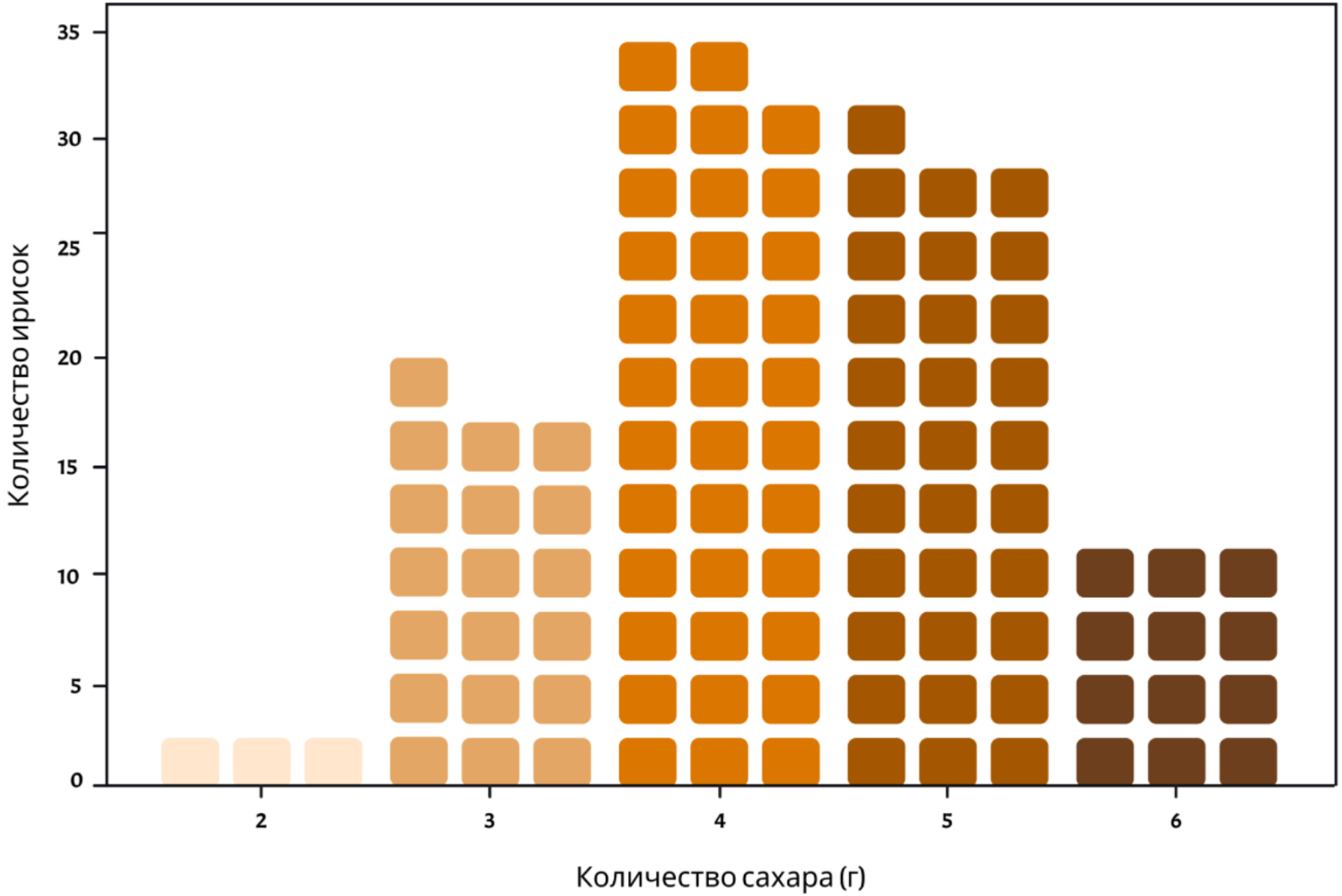
Визуализация числовых данных

Для примера возьмём 100 ирисок «Первичный ключик», исследуем содержание в них сахара. Измерим количество сахара с точностью до грамма в каждой ириске из набора:

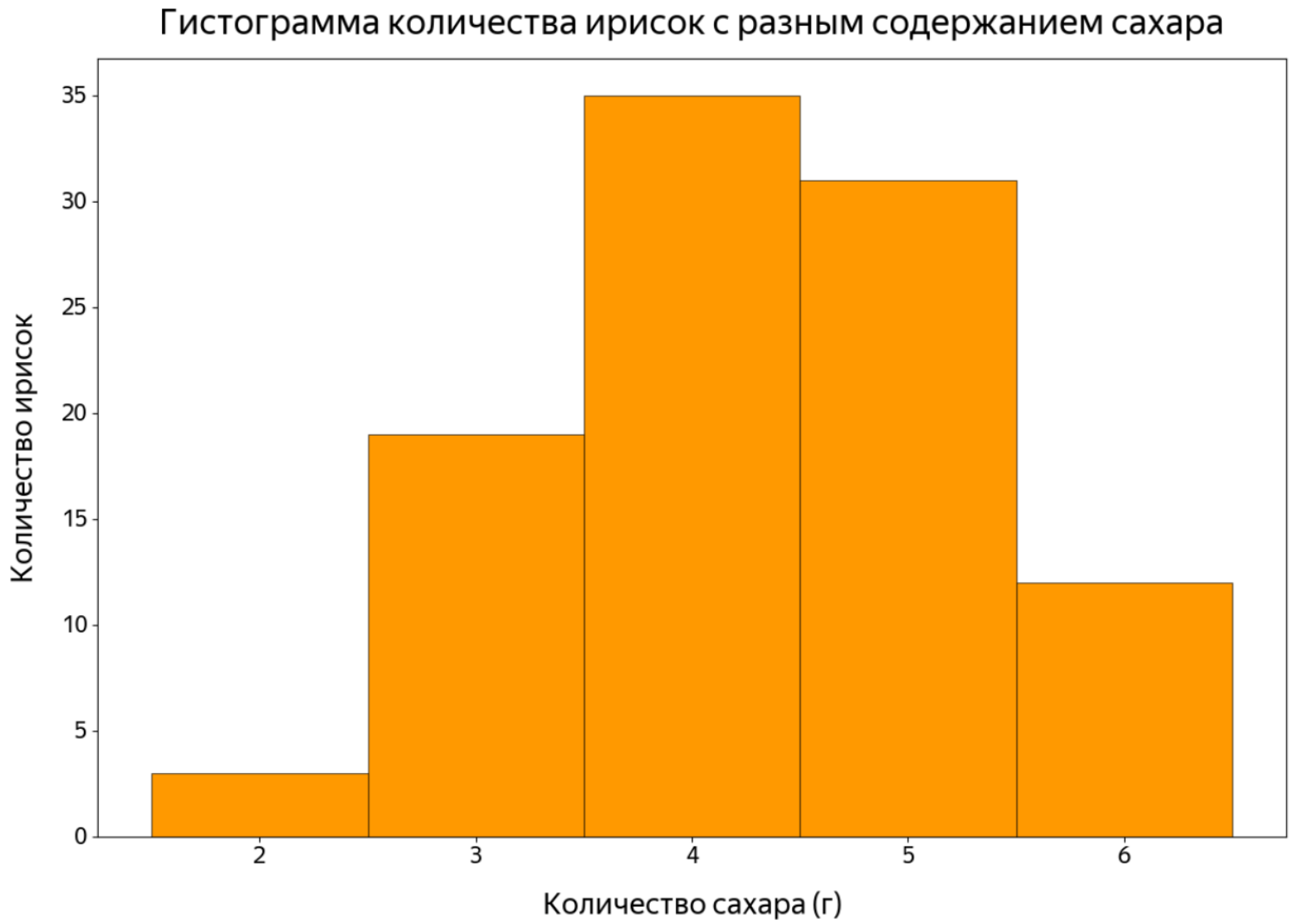
Ириски можно сгруппировать по количеству сахара в них. Получили barplot.



Визуализация числовых данных



Визуализация числовых данных



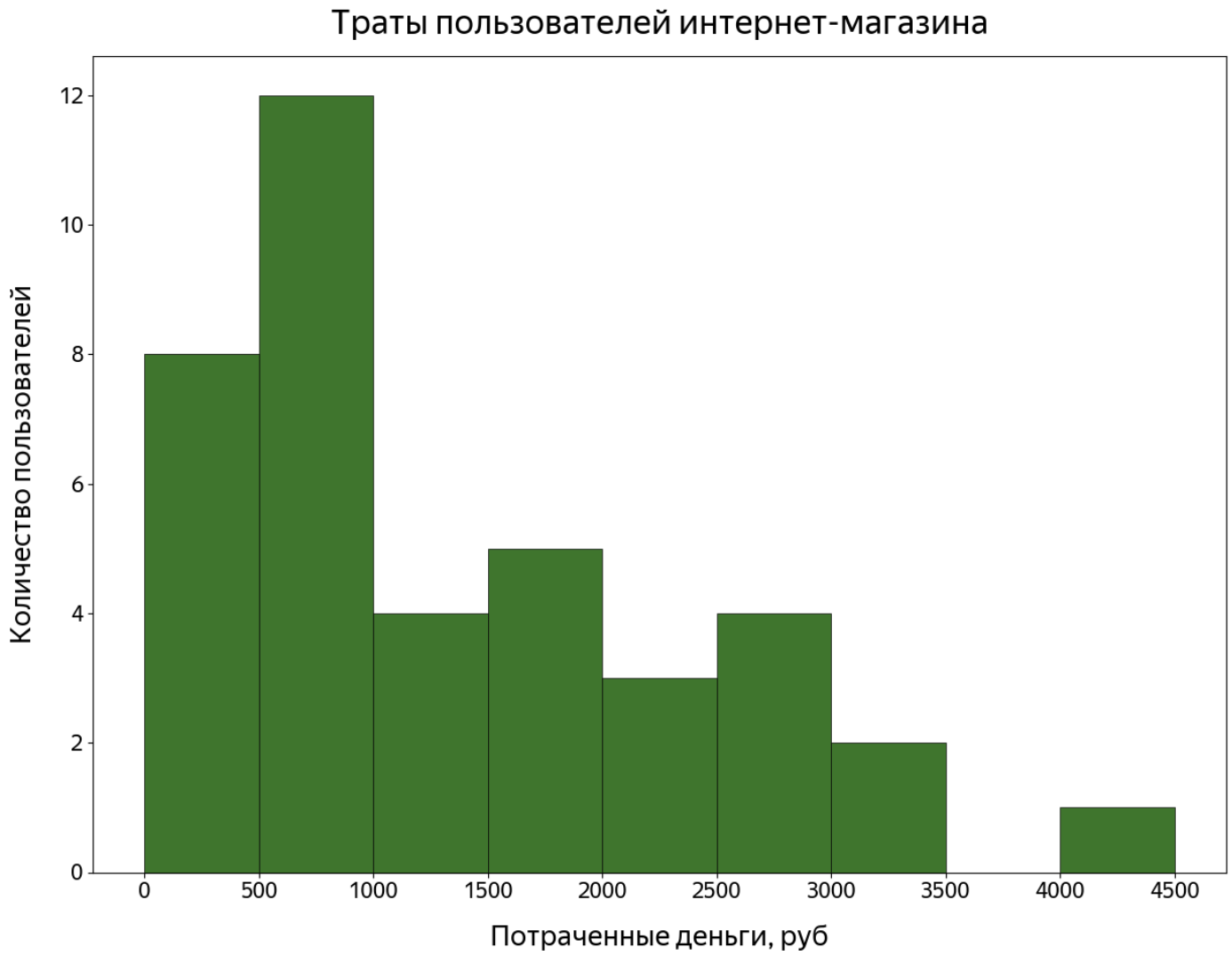
Визуализация числовых данных

Упражнение 1

Аналитик интернет-магазина подарков «Ненужная прелесть» собрал данные по покупателям за март и построил по ним гистограмму.

Проанализируйте график и выберите все верные утверждения.

- ❑ Большее количество пользователей делают небольшие покупки — до 1000 рублей.
- ❑ 7 покупателей потратили за март от 1500 до 2000 рублей.
- ❑ Покупателей, потративших 2500–3000, больше, чем покупателей, потративших 3000–3500.
- ❑ Есть небольшая часть пользователей, которые совершают довольно крупные покупки: более 4000 рублей.

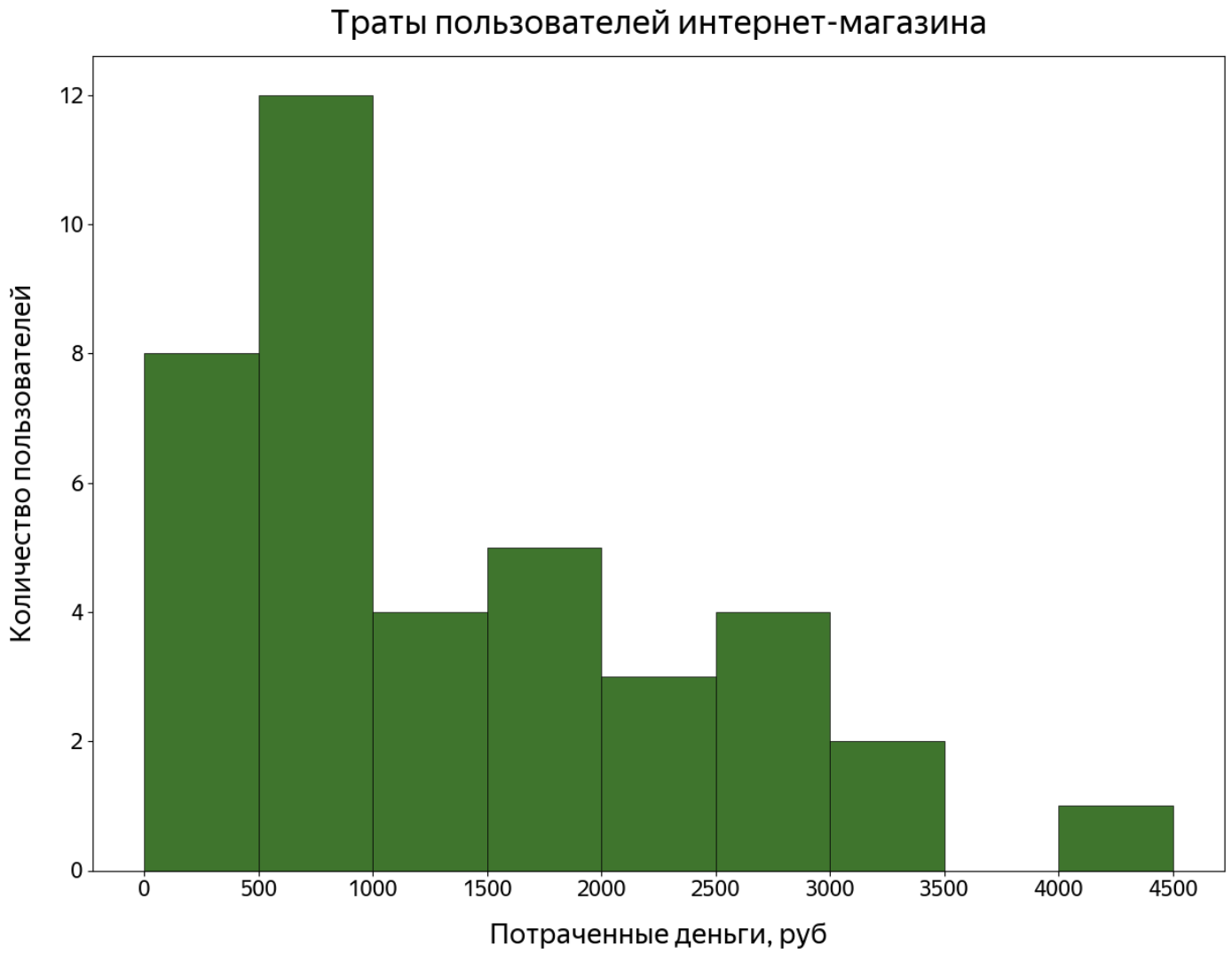


Упражнение 1

Аналитик интернет-магазина подарков «Ненужная прелесть» собрал данные по покупателям за март и построил по ним гистограмму.

Проанализируйте график и выберите все верные утверждения.

Большее количество пользователей делают небольшие покупки — до 1000 рублей.
7 покупателей потратили за март от 1500 до 2000 рублей.
Покупателей, потративших 2500–3000, больше, чем покупателей, потративших 3000–3500.
Есть небольшая часть пользователей, которые совершают довольно крупные покупки: более 4000 рублей.



Обозначения

Обозначения

Обычно в задачах есть величина, которую измеряют в разные моменты времени или для разных объектов. Например, температуру могут измерять каждый день, прибыль — каждый месяц, содержание сахара — для каждой ириски. Саму величину обозначают заглавной латинской буквой, её конкретные значения — той же латинской буквой, но строчной, а наблюдения нумеруют, и эти номера пишут как индекс у каждого значения. Например:

- T — температура (сама величина). Значение температуры в первый день — t_1, во второй — t_2, в третий — t_3 и так далее.
- P — прибыль (величина). Размер прибыли в первый месяц — p_1, во второй — p_2 и так далее.

Когда измеряют несколько величин, их обозначают разными буквами. Если в задаче величина одна, то часто её обозначают X. Например, если в задаче идёт речь только о содержании сахара в ирисках, то можно обозначить: X — содержание сахара в ирисках. Тогда содержание сахара в первой ириске — x_1, во второй — x_2 и так далее.

Набор данных, который получается после измерений, называют **выборкой**.
Выборка содержания сахара в трёх ирисках: $x_1,\ x_2,\ x_3$.

Среднее

Чтобы вычислить выборочное среднее, нужно сложить все значения и разделить на их количество.

Например, для пяти ирисок:

$$\overline{x} = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5).$$

Вообще, в математике записывают в виде общей суммы. Для этого используют символ $\sum$. Это заглавная греческая буква «сигма», в математике она означает суммирование.

Для выборки из n элементов среднее вычисляют по формуле:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

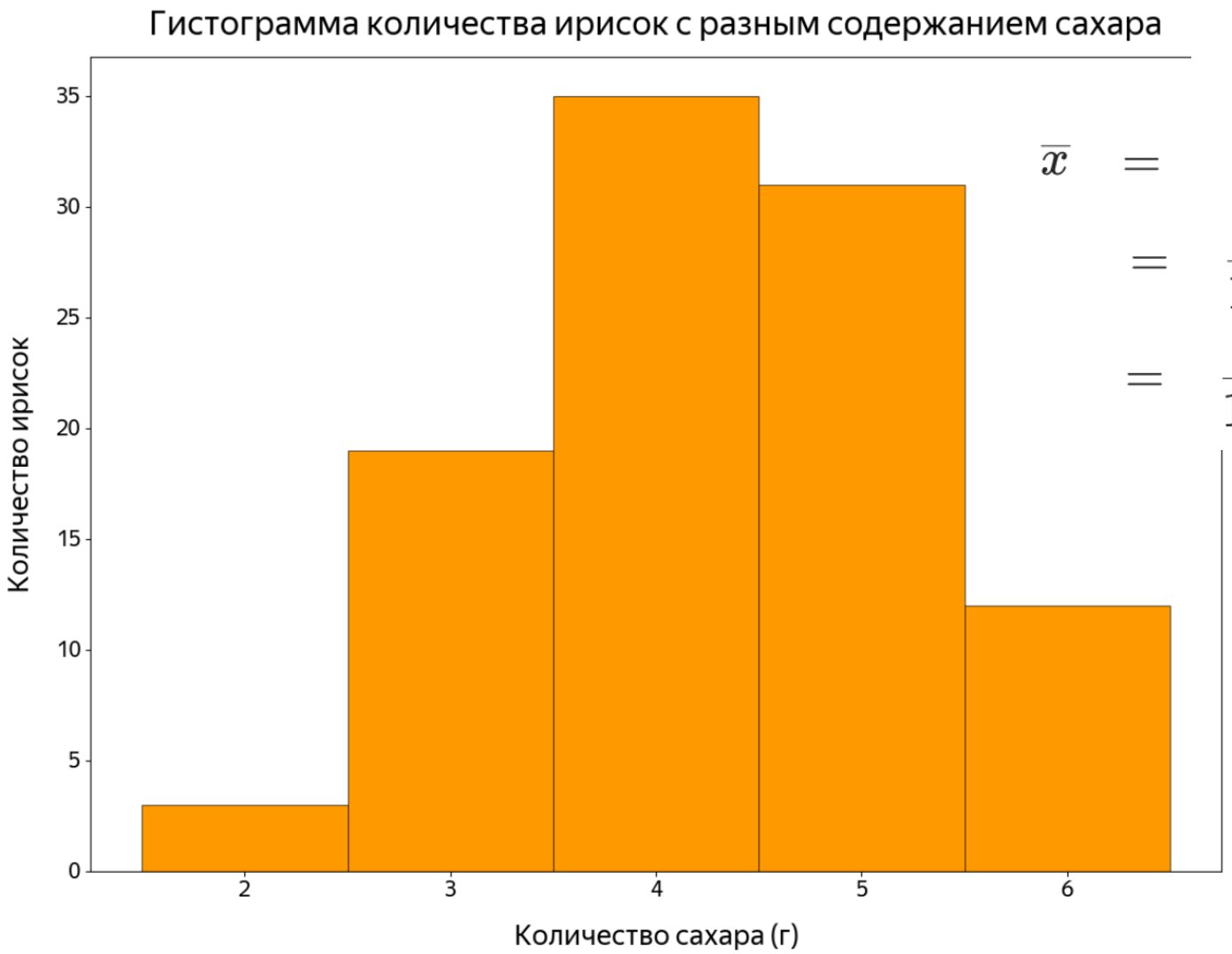
Среднее

Сигму всегда можно «развернуть» в длинную сумму:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5.$$

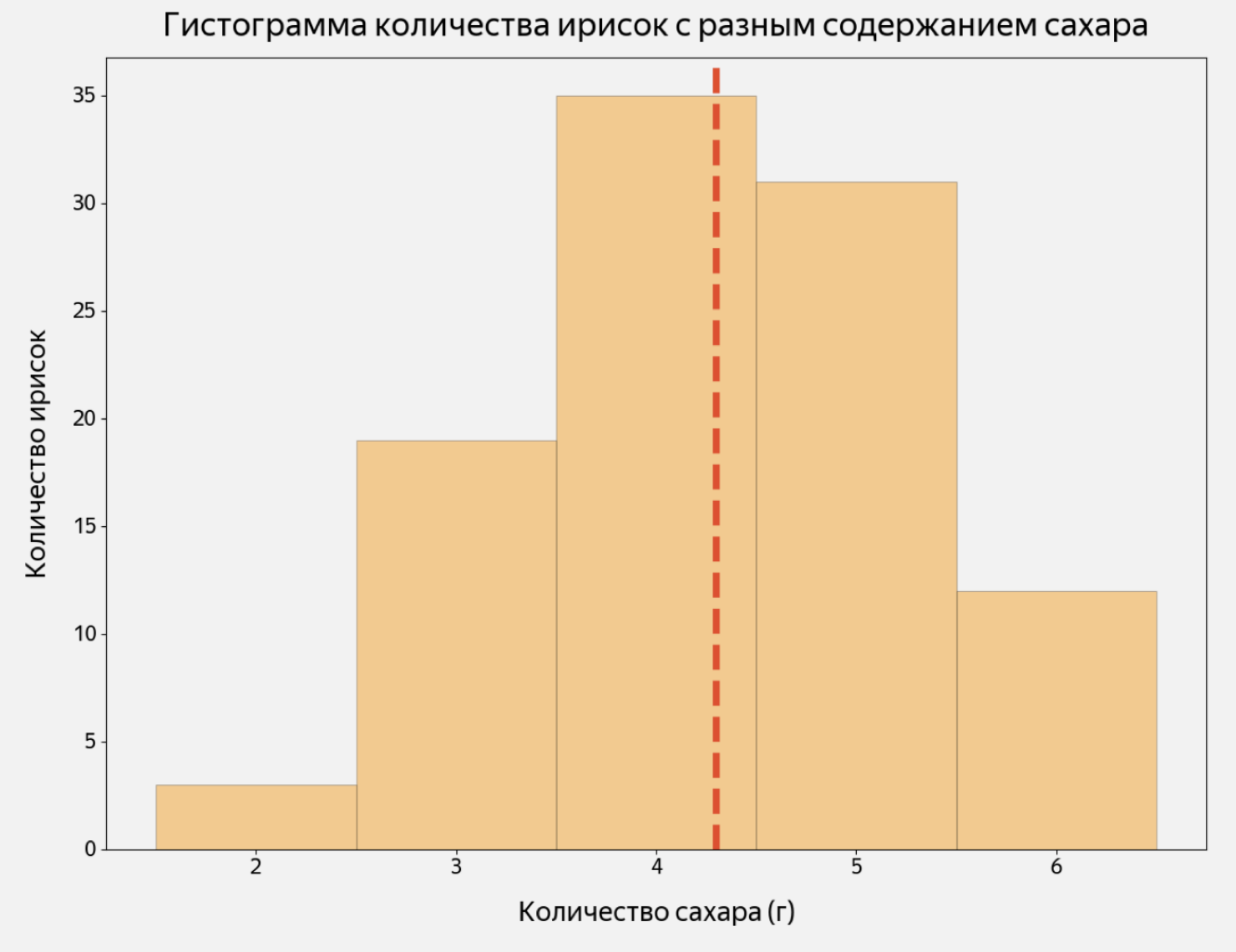
Среднее

Найдём среднее этой выборки:



$$\begin{aligned}\overline{x} &= \\ &= \frac{1}{100}(2\cdot 3 + 3\cdot 19 + 4\cdot 35 + 5\cdot 31 + 6\cdot 12) = \\ &= \frac{1}{100}(6 + 57 + 140 + 155 + 72) = 4.3.\end{aligned}$$

Среднее


Гистограмма количества ирисок с разным содержанием сахара
Количество ирисок
0
5
10
15
20
25
30
35
2
3
4
5
6
Количество сахара (г)

Упражнение 2

Василий ведёт трекер полезных привычек в приложении. Новая привычка — пить достаточное количество чистой воды. Василий ежедневно записывает количество, а в конце недели приложение анализирует данные. Если среднее количество выпитой воды лежит в диапазоне $2.0 - 2.2$ литра в день, программа засчитывает неделю как удачную и даёт пользователю звёздочку.

Получит ли Василий звёздочку за эту неделю?

Данные, которые записал Василий:

День	Количество выпитой воды, л
1	1.7
2	1.9
3	2.4
4	1.5
5	2.3
6	2.5
7	2.1

Упражнение 2

Василий ведёт трекер полезных привычек в приложении. Новая привычка — пить достаточное количество чистой воды. Василий ежедневно записывает количество, а в конце недели приложение анализирует данные. Если среднее количество выпитой воды лежит в диапазоне $2.0 - 2.2$ литра в день, программа засчитывает неделю как удачную и даёт пользователю звёздочку.

Данные, которые записал Василий:

День	Количество выпитой воды, л
1	1.7
2	1.9
3	2.4
4	1.5
5	2.3
6	2.5
7	2.1

Получит ли Василий звёздочку за эту неделю?

Да, получит.

Вычислим среднее количество выпитой в день воды:

$$\overline{x} = \frac{1}{7}\sum_{i=1}^{7} x_i = \frac{1}{7} \cdot (1.7 + 1.9 + 2.4 + 1.5 + 2.3 + 2.5 + 2.1) = \frac{14.4}{7} \approx 2.06.$$

Полученное число лежит в нужном диапазоне: от 2.0 до 2.2 литра в день. Значит, приложение засчитает неделю как удачную и выдаст Василию звёздочку.

Упражнение 3

На фабрике производят ещё ириски «Лисп-лисп». Возьмём выборку из 100 таких ирисок и измерим в них содержание сахара:

Вычислите среднее содержание сахара в наборе ирисок «Лисп-лисп».

Количество сахара (г)	Количество ирисок «Лисп-лисп»
2	1
3	5
4	14
5	27
6	33
7	20

Упражнение 3

На фабрике производят ещё ириски «Лисп-лисп». Возьмём выборку из 100 таких ирисок и измерим в них содержание сахара:

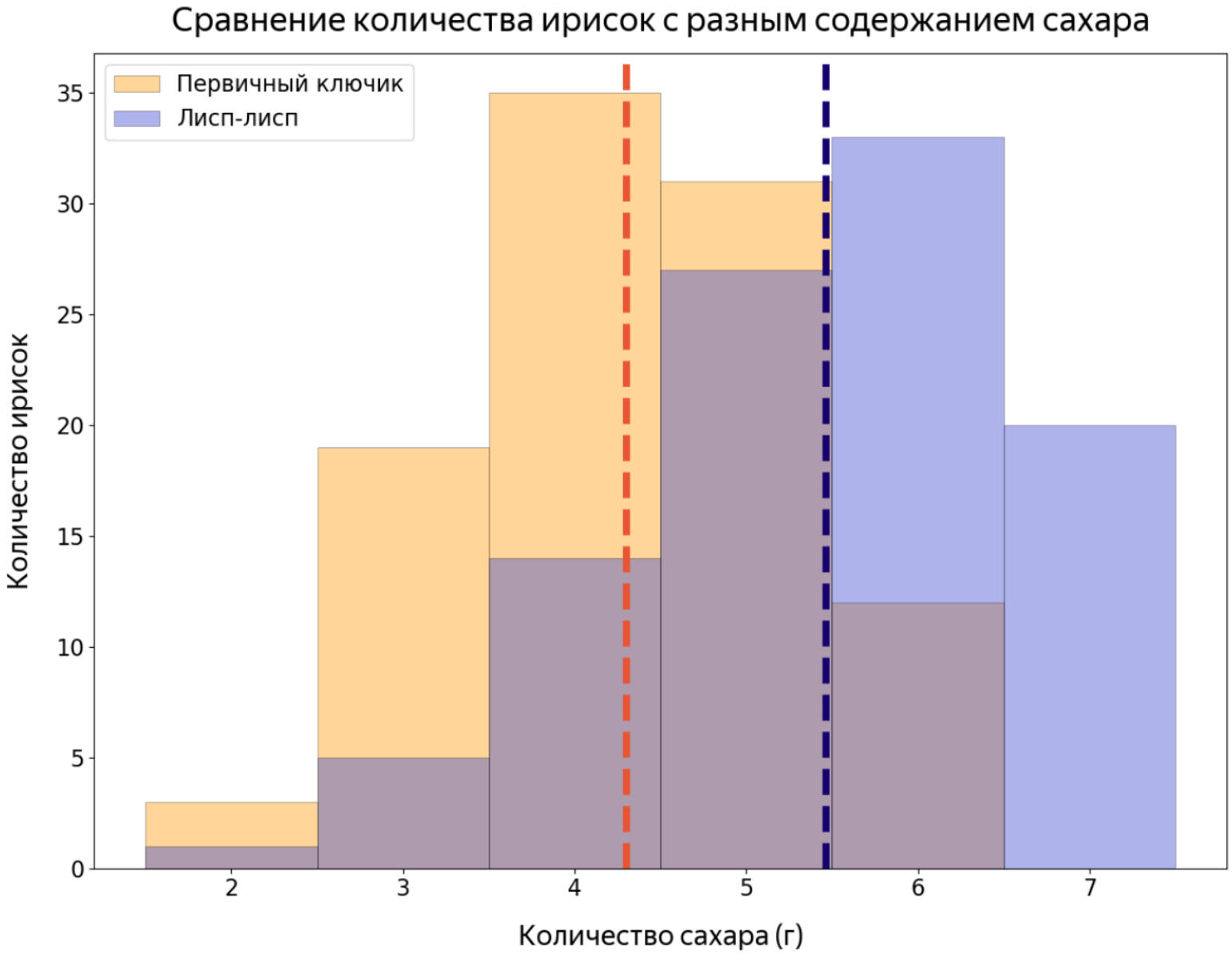
Количество сахара (г)	Количество ирисок «Лисп-лисп»
2	1
3	5
4	14
5	27
6	33
7	20

Вычислите среднее содержание сахара в наборе ирисок «Лисп-лисп».

$$\begin{aligned} \overline{x} &= \frac{1}{100}(2 \cdot 1 + 3 \cdot 5 + 4 \cdot 14 + 5 \cdot 27 + 6 \cdot 33 + 7 \cdot 20) = \\ &= \frac{1}{100}(2 + 15 + 56 + 135 + 198 + 140) = 5.46. \end{aligned}$$

Ответ: 5.46

Сравнение средних



Практика

Задача 1

Определите тип данных и разнесите их по соответствующим корзинкам.

Ценовой сегмент: премиум, средний, эконом	Производительность станка	Тип продажи: опт, офлайн-розница, онлайн-розница
Процент какао в шоколаде	Ароматы продуктов	Объём чана для варки

Категориальные

Порядковые

Числовые

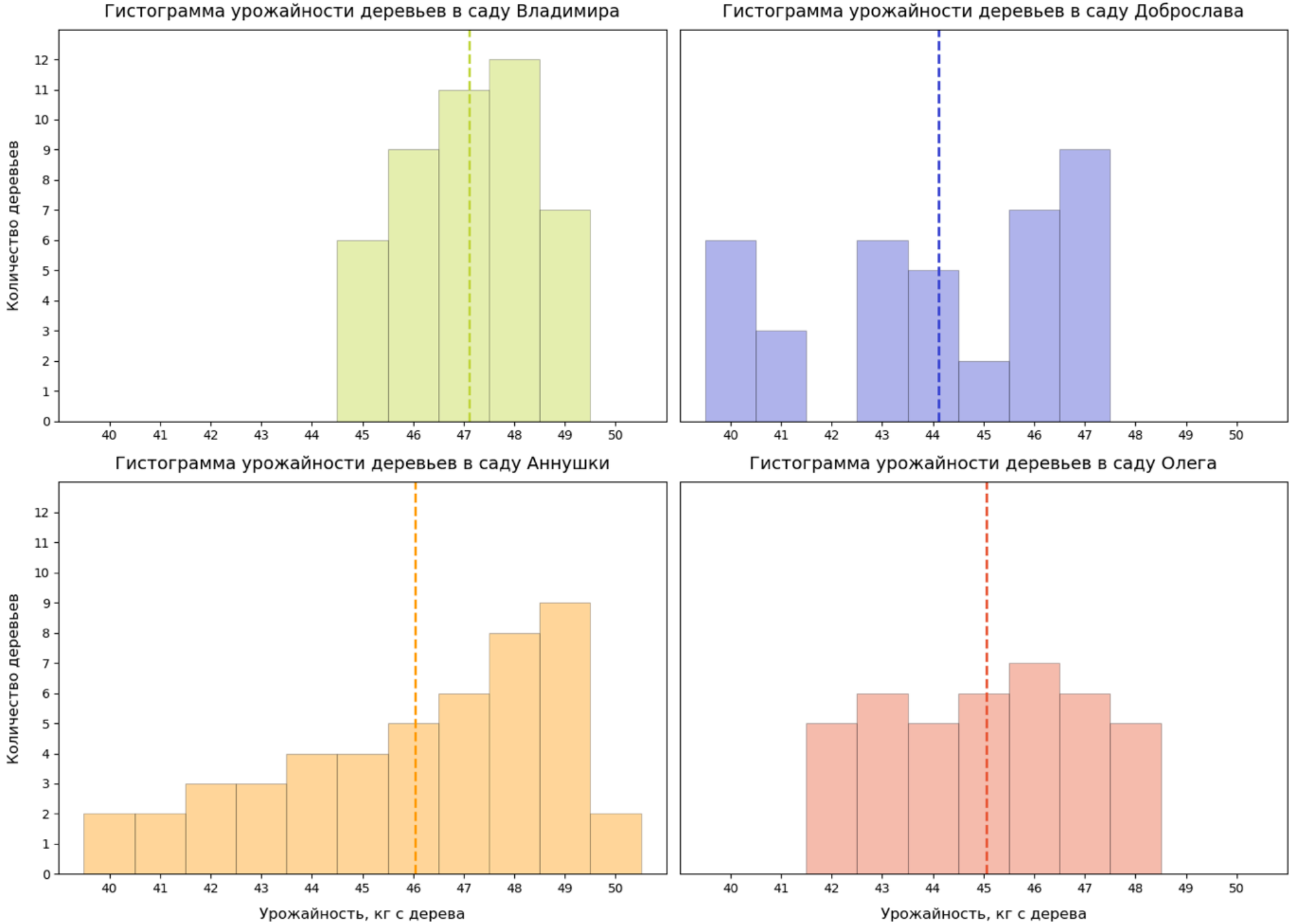
Задача 1

Определите тип данных и разнесите их по соответствующим корзинкам.

Категориальные	Порядковые	Числовые
Тип продажи: опт, офлайн-розница, онлайн-розница	Ценовой сегмент: премиум, средний, эконом	Объём чана для варки
Ароматы продуктов		Процент какао в шоколаде

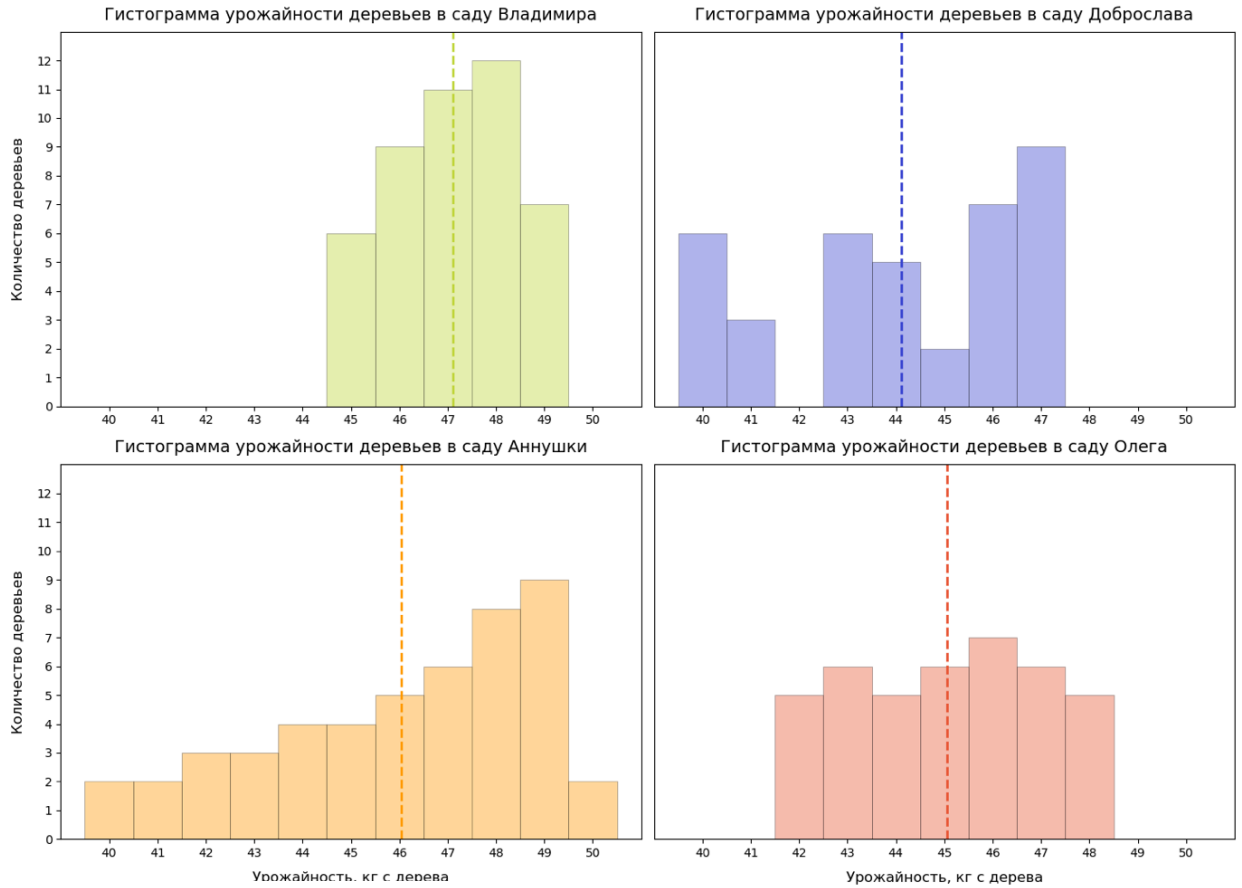
Задача 2

Расставьте фермерские хозяйства в порядке возрастания средней урожайности деревьев.



Задача 2

Расставьте фермерские хозяйства в порядке возрастания средней урожайности деревьев.



Доброслав < Олег < Аннушка < Владимир

Упорядочим средние урожайности: $44 < 45 < 46 < 47$. Расставим фермеров в соответствующем порядке: Доброслав, Олег, Аннушка, Владимир.

Задача 3

Спортсмен Яблочко готовится бежать кросс и ежедневно тренируется.
Тренер свёл информацию о его тренировках в таблицу.

Рассчитайте среднюю дистанцию,
которую пробегает во время тренировки Яблочко.

Выберите один верный вариант ответа.

- ◯ 8.7
- ◯ 7
- ◯ 10
- ◯ 8.5

День	Километраж
1	7
2	8
3	8
4	9
5	9
6	8
7	9
8	10
9	10
10	9

Задача 3

Спортсмен Яблочко готовится бежать кросс и ежедневно тренируется.
Тренер свёл информацию о его тренировках в таблицу.

Рассчитайте среднюю дистанцию,
которую пробегает во время тренировки Яблочко.

Выберите один верный вариант ответа.

День	Километраж
1	7
2	8
3	8
4	9
5	9
6	8
7	9
8	10
9	10
10	9

✓ 8.7

В таблице есть повторяющиеся данные: три дня Яблочко пробегал по 8 км, четыре дня по 9 км, два дня по 10 км. Учтём это при расчёте среднего:

$$\frac{1 \cdot 7 + 3 \cdot 8 + 4 \cdot 9 + 2 \cdot 10}{10} = \frac{7 + 24 + 36 + 20}{10} = \frac{87}{10} = 8.7.$$

Медиана

Зарплата

Каждый год совет акционеров фабрики собирает совещание, на котором обсуждают разные важные вопросы, в том числе условия труда и оплаты сотрудников. На совещании директор предоставляет разные данные: о количестве инцидентов с нарушением техники безопасности, о количестве дней нетрудоспособности сотрудников и, конечно, о заработной плате.

В прошлом году старый директор уволился и вскрылись интересные истории. Руководитель фабрики из года в год заявлял о стабильном уровне роста заработной платы, совет акционеров был доволен. Когда пришёл новый директор, обнаружилось, что зарплата росла только у старого руководителя, а у обычных сотрудников долгие годы уровень дохода был постоянным.

Зарплата

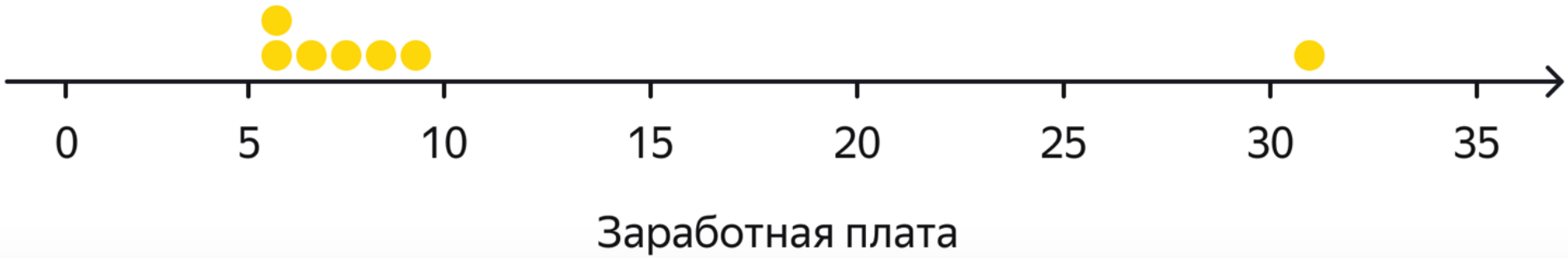
Всё зависит от того, как считать. Посмотрим на данные о заработной плате:

Сотрудник	Заработная плата, тыс. фантиков
Иннокентий Павлович, кондитер	10
Афанасий Борисович, электрик	6
Игорь Нахопетович, директор	31
Алгебрина Рубиновна, повар	8
Семён Владленович, водитель	6
Тамара Георгиевна, технолог	9
Ирма Робертовна, тестировщик	7

Зарплата

Сотрудник	Заработная плата, тыс. фантиков
Иннокентий Павлович, кондитер	10
Афанасий Борисович, электрик	6
Игорь Нахопетович, директор	31
Алгебрина Рубиновна, повар	8
Семён Владленович, водитель	6
Тамара Георгиевна, технолог	9
Ирма Робертовна, тестировщик	7

Визуализируем:

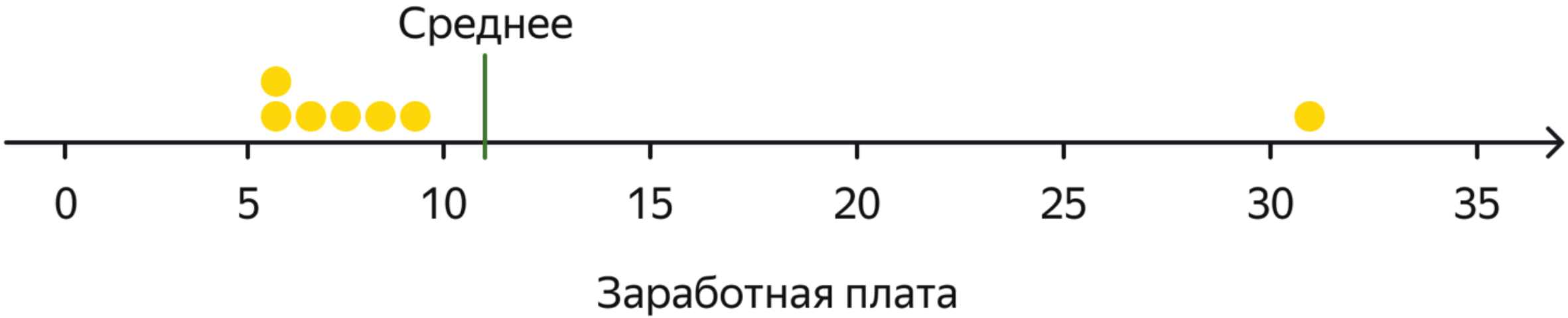


Зарплата

Обозначим зарплату как s (от англ. salary — зарплата) и посчитаем среднее значение:

$$\overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i = \frac{1}{7}\cdot(10+6+31+8+6+9+7) = 11.$$

Из-за выбивающейся зарплаты среднее работает плохо.



Медиана

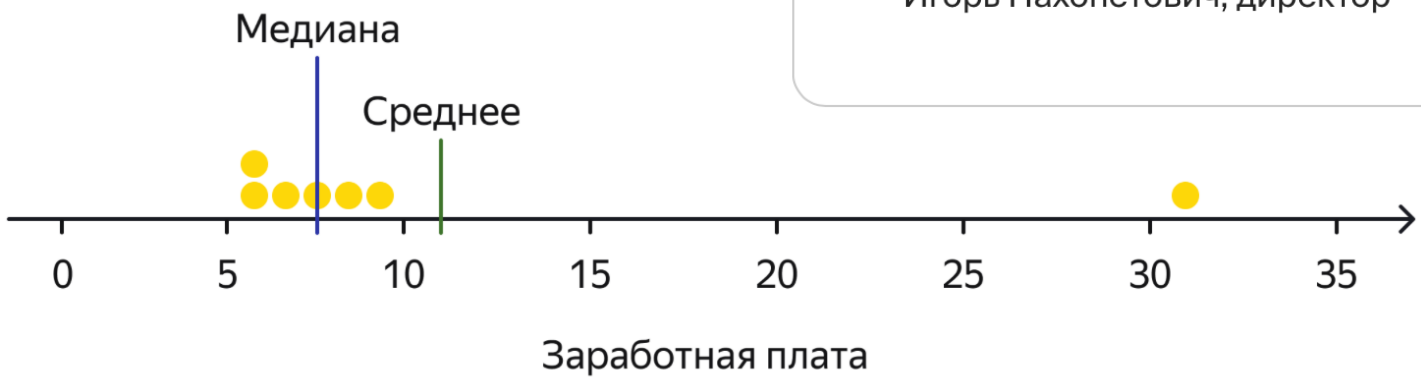
Сначала расставим все объекты по возрастанию.

Сначала расставим все объекты по возрастанию.

А теперь возьмём число, которое стоит посередине списка. Это значение 8.

Мы нашли медиану!

Сотрудник	Заработная плата, тыс. фантиков	Порядковый номер
Афанасий Борисович, электрик	6	1
Семён Владленович, водитель	6	2
Ирма Робертовна, тестировщик	7	3
Алгебрина Рубиновна, повар	8	4
Тамара Георгиевна, технолог	9	5
Иннокентий Павлович, кондитер	10	6
Игорь Нахопетович, директор	31	7



Видно, что для этого набора данных медиана лучше отражает центр распределения.

Медиана

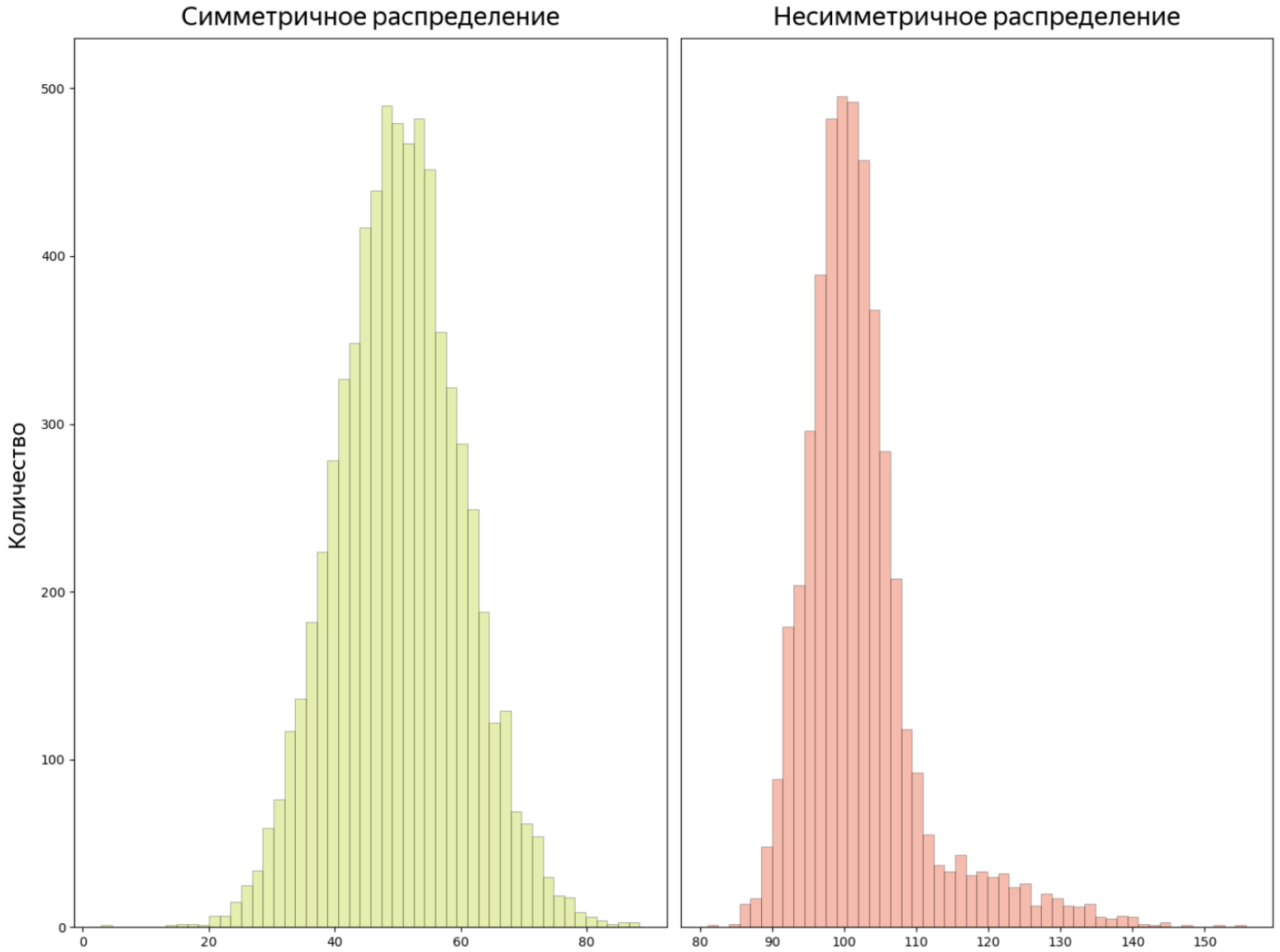
Медиана — это наблюдение, которое делит весь набор данных на две равные части: меньше него 50% наблюдений и больше него тоже 50% наблюдений.

Пример вычисления медианы

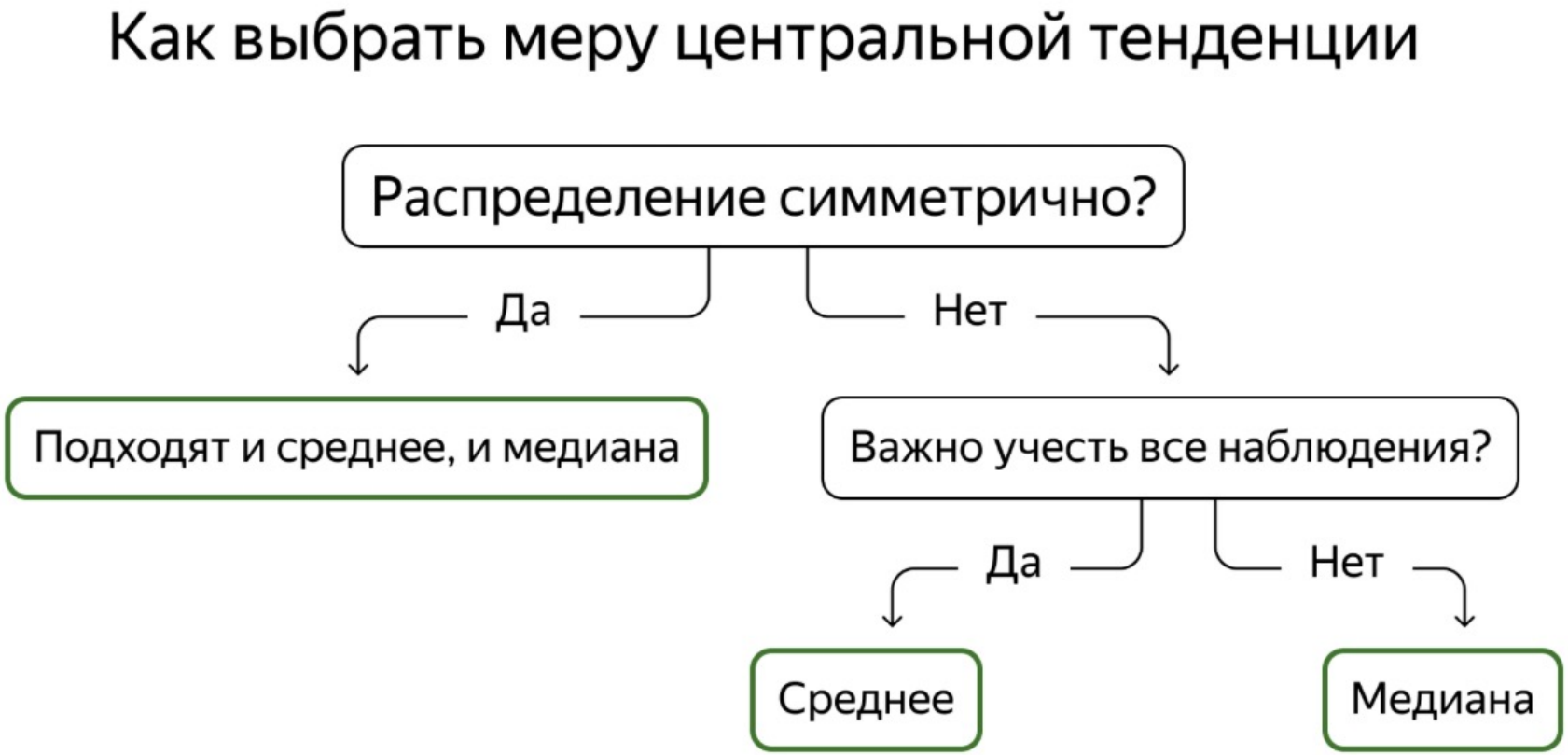
Сотрудник	Заработная плата, тыс. фантиков	Порядковый номер
Афанасий Борисович, электрик	6	1
Семён Владленович, водитель	6	2
Алгебрина Рубиновна, повар	7	3
Ирма Робертовна, тестировщик	8	4
Тамара Георгиевна, технолог	9	5
Иннокентий Павлович, кондитер	10	6
Хомяк Игоревич, зам. директора	27	7
Игорь Нахопетович, директор	31	8

Свойства распределения

Симметричность распределения



Медиана или среднее?



Квантиль

От медианы к квантилю

В прошлом уроке вы познакомились с понятием медианы. Она представляет собой значение, которое делит упорядоченный набор данных на две равные части. Оказывается, медиана — это частный случай квантиля. Медиана помогает отсечь ровно половину данных, а квантиль — любую заданную часть. Разберёмся на примерах.

Рассмотрим простой набор данных:

10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

Найдём 0.2-квантиль.

Это означает, что нам надо найти такое число в заданном наборе, что 20% от общего количества значений будут меньше или равны ему.

Квантиль

Рассмотрим простой набор данных:

10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

Найдём 0.2-квантиль.

Это означает, что нам надо найти такое число в заданном наборе, что 20% от общего количества значений будут меньше или равны ему.

Для этого обычно выполняют следующие шаги:

1. Сортируют набор данных по возрастанию.
2. Находят номер элемента по такой формуле: $n \cdot \alpha$, где n — количество элементов в наборе, α — доля, которая нас интересует.

В нашем примере 10 значений. Они уже отсортированы. Порядковый номер искомого значения равен $10 \cdot 0.2 = 2$. Соответственно, значение 20 и будет 0.2 0.2-квантилем. Заметим, что справа от него будет 100%−20%=80% значений.

Квантиль

Число X является α-квантилем набора данных, если оно делит этот набор данных таким образом, что $\alpha\%$ наблюдений меньше или равны X и $(100 - \alpha)\%$ наблюдений больше или равны X.

Упражнение 4

На кондитерской фабрике есть специальный цех по производству зефира. В этом цехе очень важно поддерживать определённый уровень влажности, иначе зефир может получиться очень сухим или не высохнет вообще. Поэтому в цехе устанавливают специальное оборудование и собирают данные о влажности воздуха. Соберём данные за 20 дней и найдём значение 0.05-квантиля:

$$35, 37, 36, 38, 26, 41, 33, 32, 39, 47, 38, 30, 34, 36, 37, 42, 29, 45, 44, 27.$$

Сначала отсортируем эти значения по возрастанию:

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

У нас в данных 20 значений, а значит, позиция значения, левее которого будет 5% значений, равна: $20 \cdot 0.05 = 1$. Тогда 0.05-квантиль на этих данных будет первое значение в отсортированном наборе — оно равно 26.

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

Найдите значение 0.1-квантиля для тех же данных.

Упражнение 4

На кондитерской фабрике есть специальный цех по производству зефира. В этом цехе очень важно поддерживать определённый уровень влажности, иначе зефир может получиться очень сухим или не высохнет вообще. Поэтому в цехе устанавливают специальное оборудование и собирают данные о влажности воздуха. Соберём данные за 20 дней и найдём значение 0.05-квантиля:

$$35, 37, 36, 38, 26, 41, 33, 32, 39, 47, 38, 30, 34, 36, 37, 42, 29, 45, 44, 27.$$

Сначала отсортируем эти значения по возрастанию:

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

У нас в данных 20 значений, а значит, позиция значения, левее которого будет 5% значений, равна: $20 \cdot 0.05 = 1$. Тогда 0.05-квантиль на этих данных будет первое значение в отсортированном наборе — оно равно 26.

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

Найдите значение 0.1-квантиля для тех же данных.

Порядковый номер: $20 \cdot 0.1 = 2$.

Соответствующее значение: 27.

Ответ: 27

Упражнение 4

На кондитерской фабрике есть специальный цех по производству зефира. В этом цехе очень важно поддерживать определённый уровень влажности, иначе зефир может получиться очень сухим или не высохнет вообще. Поэтому в цехе устанавливают специальное оборудование и собирают данные о влажности воздуха. Соберём данные за 20 дней и найдём значение 0.05-квантиля:

$$35, 37, 36, 38, 26, 41, 33, 32, 39, 47, 38, 30, 34, 36, 37, 42, 29, 45, 44, 27.$$

Сначала отсортируем эти значения по возрастанию:

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

У нас в данных 20 значений, а значит, позиция значения, левее которого будет 5% значений, равна: $20 \cdot 0.05 = 1$. Тогда 0.05-квантиль на этих данных будет первое значение в отсортированном наборе — оно равно 26.

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

Найдите значение 0.2-квантиля для тех же данных.

Упражнение 4

На кондитерской фабрике есть специальный цех по производству зефира. В этом цехе очень важно поддерживать определённый уровень влажности, иначе зефир может получиться очень сухим или не высохнет вообще. Поэтому в цехе устанавливают специальное оборудование и собирают данные о влажности воздуха. Соберём данные за 20 дней и найдём значение 0.05-квантиля:

$$35, 37, 36, 38, 26, 41, 33, 32, 39, 47, 38, 30, 34, 36, 37, 42, 29, 45, 44, 27.$$

Сначала отсортируем эти значения по возрастанию:

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

У нас в данных 20 значений, а значит, позиция значения, левее которого будет 5% значений, равна: $20 \cdot 0.05 = 1$. Тогда 0.05-квантиль на этих данных будет первое значение в отсортированном наборе — оно равно 26.

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

Найдите значение 0.9-квантиля для тех же данных.

Округление квантилей

Иногда количество элементов в выборке не кратно проценту квантиля, из-за чего произведение $n \cdot \alpha$ даёт нецелое число. Например, для 0.33-квантиля в примере с зефиром: $n \cdot \alpha = 6.6$. В таком случае обычно используют различные методы интерполяции для вычисления значения квантиля. Их много, мы можем выбрать один из них.

Договоримся, что если при умножении $n \cdot \alpha$ получается дробное число, то мы будем брать среднее значение двух ближайших соседей.

Найдём 0.33-квантиль для наших данных:

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

Порядковый номер искомого числа: $20 \cdot 0.33 = 6.6$. Два соседа — это значения под номерами 6 и 7:

$$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$$

Тогда искомый квантиль равен $\dfrac{33 + 34}{2} = 33.5$. И это число строго удовлетворяет определению: в наборе ровно 33% данных меньше или равны 33.5.

Упражнение 5

Найдите 0.48-квантиль для зефирного набора:

$26, 27, 29, 30, 32, 33, 34, 35, 36, 36, 37, 37, 38, 38, 39, 41, 42, 44, 45, 47.$