Статистика 1

Семерикова Елена Вячеславовна
к.э.н., Старший преподаватель Департамента прикладной экономики ФЭН НИУ ВШЭ

Статистика помогает принимать уверенные решения...

...когда мы располагаем неполной информацией.


Представьте себе, что мы хотим узнать средний вес...
...всей рыбы в озере.
Ловись, ловись, рыбка...
Если мы узнаем, сколько в среднем весит одна рыбка...
...мы сможем понять, сколько рыбешек нам нужно ловить каждый день, чтобы спасти наших питомцев от голодной смерти!
Если бы мы осушили озеро и взвесили каждую рыбку...

...то получили бы всю необходимую информацию
и высчитали средний вес.

Но по очевидным причинам мы не можем этого сделать.


С другой стороны, если мы поймаем 100 рыбешек и взвесим их...
Эти 100 рыбок весят 112 кг.
Следовательно, в среднем одна рыбешка весит 1,12 кг!
...мы получим неполную информацию о всей рыбе в озере.


Статистика предполагает использование той рыбы, которую мы поймали...
...чтобы сделать доверительное суждение относительно всей рыбы в этом озере.
...чтобы судить о той, которая осталась в озере.
Правда? Как же это работает?

Если мы не поймаем всю

рыбу,

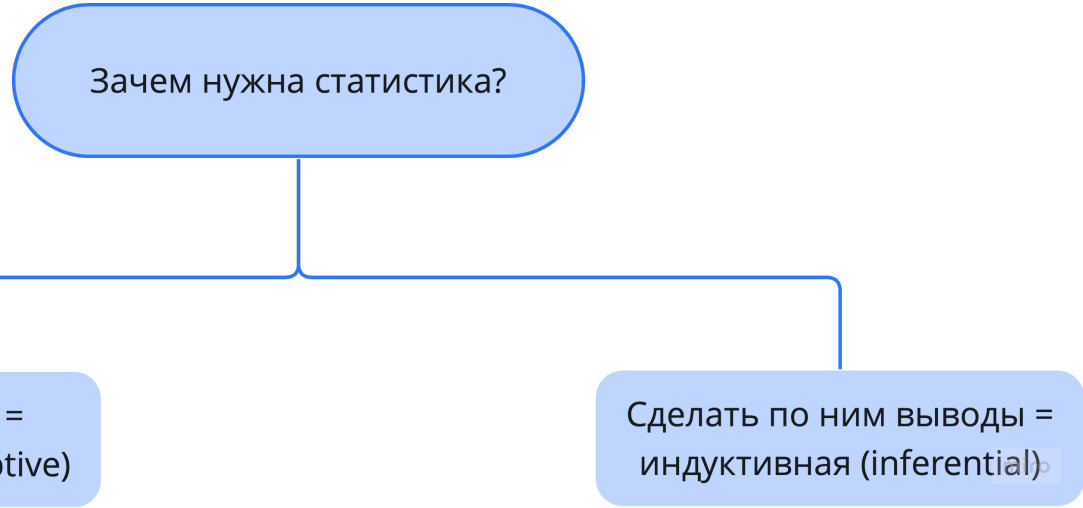
Мы никогда не сможем узнать

со 100% определенностью,

что там происходит внизу

Мы можем использовать статистику, чтобы делать доверительные предположения

Но их нельзя использовать как неоспоримый факт



Описательная статистика:
описываем выборку

Индуктивная статистика:
на основе выборки делаем заключение о свойствах генеральной совокупности

Говорим сегодня об этом, а именно….

описание статистических данных, представления их в форме таблиц, распределений, характеристик распределения и пр.

формулировка выводов с помощью методов, основанных на теории вероятностей

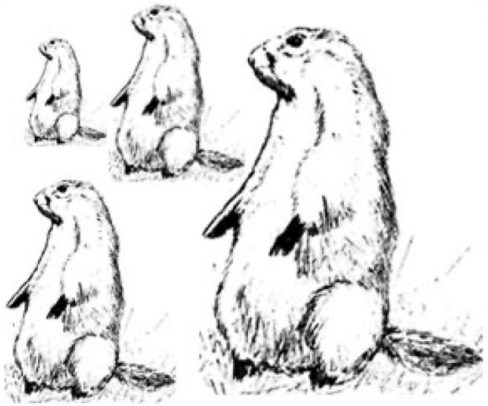

Какие данные мы получаем?
О ком или о чем данные?
Какие характеристики они описывают?
Генеральная совокупность (Population)
Выборка (Sample) (подмножество ГС)
Количественные / Качественные
Дискретные / Непрерывные

Изучаем популяцию сурков

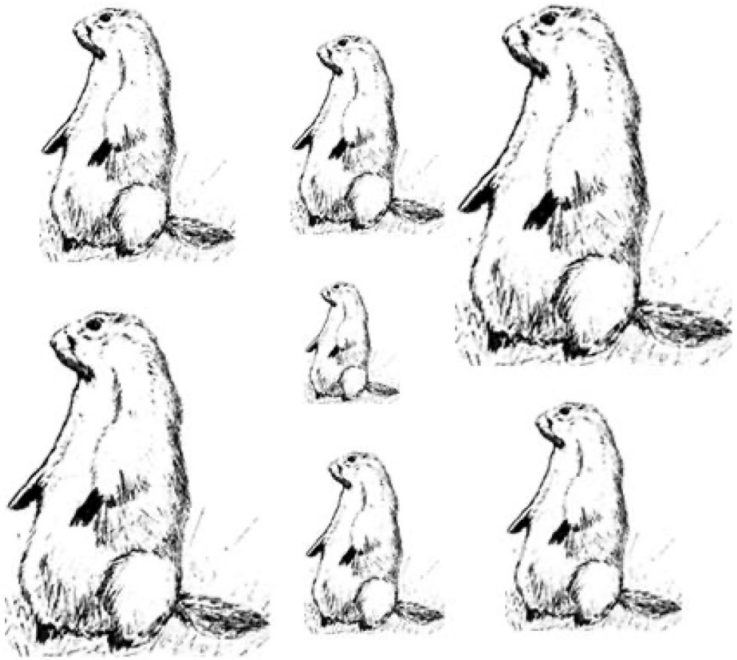
Популяция = Генеральная совокупность

наблюдение

ВЫБОРКА



популяция – совокупность всех интересующих нас объектов



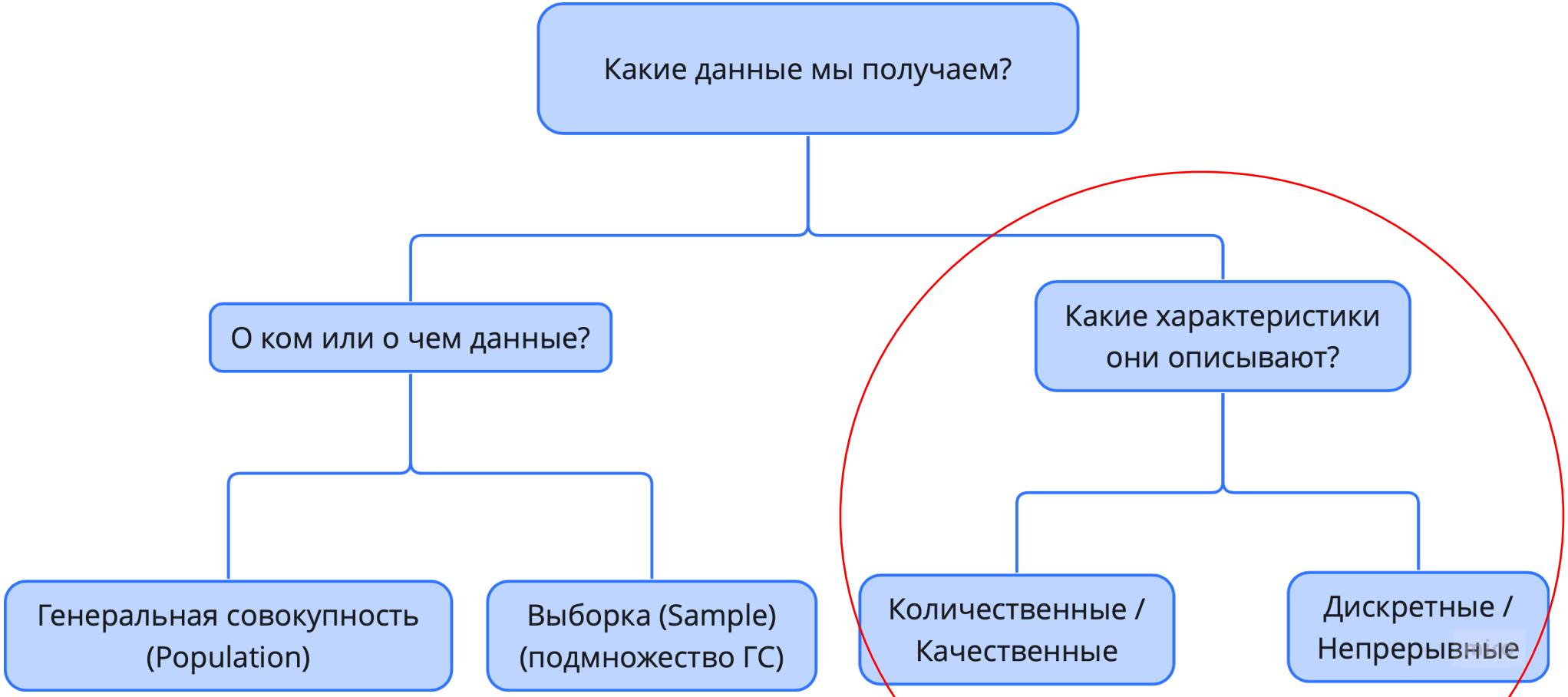
Выборка и генеральная совокупность

Выборка – случайным образом отобранная часть генеральной совокупность

Выборка должна быть репрезентативна – отражать свойства популяции т.е. случайной

(все особи имеют одинаковые шансы попасть в выборку)

Рис. Случайный сурок в выборке


Какие данные мы получаем?
О ком или о чем данные?
Какие характеристики они описывают?
Генеральная совокупность (Population)
Выборка (Sample) (подмножество ГС)
Количественные / Качественные
Дискретные / Непрерывные

Что можно измерить у сурка?

Вес
Возраст
Длину хвоста

Пол
Цвет шерсти

Целесообразно ли считать среднее в этом случае? Как можно интерпретировать результат?

В данном случае лучше использовать процентное описание: 60% сурков-девочек, 40% сурков-мальчиков

Типы признаков

Количественные

Качественные

Пример:

- Данные о поездках такси в Москве

Год	Сезон	Время суток	В пути, мин	Расстояние, км	Стоимость, руб	Руб/км	Температура	Час пик	Временной интервал, ч
2021	лето	день	89	58	1483	25,6	32	1	17:00-20:00
2021	лето	день	82	61	1859	30,5	32	0	14:00-16:00
2021	лето	ночь	17	15	680	45,3	15	0	23:00-24:00
2021	весна	день	27	8	410	51,3	16	0	13:00-14:00
2020	осень	вечер	26	23	410	17,8	5	0	21:00-23:00
2020	осень	день	42	23	654	28,4	9	1	17:00-20:00
2020	осень	утро	27	17	339	19,9	10	0	11:00-13:00
2020	осень	утро	32	19	472	24,8	11	0	11:00-13:00
2020	весна	утро	37	32	639	20,0	7	0	11:00-13:00
2020	зима	день	46	8,5	378	44,5	1	0	14:00-16:00
2020	зима	день	76	58	1362	23,5	2	1	17:00-20:00
2020	зима	утро	56	27	789	29,2	1	0	11:00-13:00
2020	зима	утро	69	30	755	25,2	-2	0	11:00-13:00
2019	лето	день	54	27	1016	37,6	22	0	14:00-16:00
2019	лето	день	6	4	199	49,8	13	0	14:00-16:00
2019	лето	утро	68	58	1623	28,0	29	1	8:00-10:00
2019	весна	день	20	8	260	32,5	7	0	14:00-16:00
2019	весна	вечер	19	8	300	37,5	4	0	21:00-23:00
2019	весна	вечер	37	39	801	20,5	0	0	21:00-23:00
2019	весна	день	25	16	477	29,8	0	1	17:00-20:00

Источник: данные агрегатора такси

Качественные признаки

- Примеры качественных признаков:
- Пол
- Порода животного
- Место жительства (село/город)

Качественная

Год	Сезон	Время суток	В пути, мин	Расстояние, км	Стоимость, руб	Стоимость/км, руб	Температура, С	Час пик	Балл пробки
2020	осень	вечер	26	23	410	17,8	+5	нет	3
2020	осень	день	42	23	654	28,4	+9	да	7
2020	зима	день	76	58	1362	23,5	+2	да	5
2019	весна	день	20	8	260	32,5	+7	нет	6
2019	весна	вечер	37	39	801	20,5	0	нет	4

Источник: данные агрегатора такси

Качественные признаки

Сезон	Время суток	Час пик	Балл пробки
осень	вечер	нет	3
осень	день	да	7
зима	день	да	5
весна	день	нет	6
весна	вечер	нет	4

Номинальные

Порядковые

Источник: данные агрегатора такси

Количественные признаки

Примеры количественных признаков:

- Доход домохозяйства
- Рост
- Вес

Год	Сезон	Время суток	В пути, мин	Расстояние, км	Стоимость, руб	Стоимость/км, руб	Температура, С	Час пик	Балл пробки
2020	осень	вечер	26	23	410	17,8	+5	нет	3
2020	осень	день	42	23	654	28,4	+9	да	7
2020	зима	день	76	58	1362	23,5	+2	да	5
2019	весна	день	20	8	260	32,5	+7	нет	6
2019	весна	вечер	37	39	801	20,5	0	нет	4

Источник: данные агрегатора такси

Количественные переменные

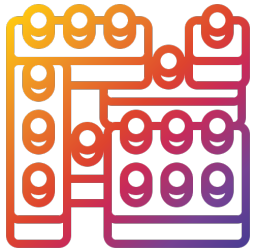
В пути, мин	Расстояние, км	Стоимость, руб	Стоимость/км, руб	Год	Температура, C	Температура, F
26	23	410	17,8	2020	+5	41
42	23	654	28,4	2020	+9	48,2
76	58	1362	23,5	2020	+2	35,6
20	8	260	32,5	2019	+7	44,6
37	39	801	20,5	2019	0	32

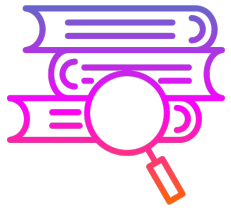
Относительные

Интервальные

Дискретные и непрерывные переменные

Дискретные ← Можно сделать из непрерывных переменных дискретные — **Непрерывные**





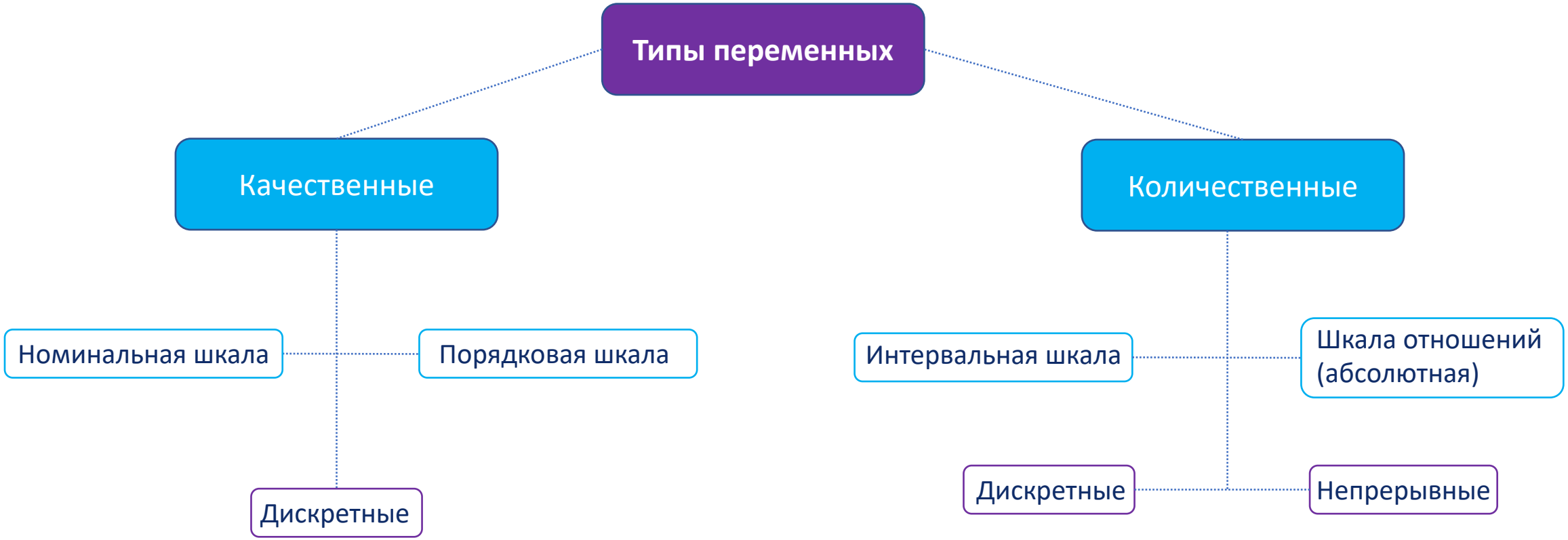
Качественные переменные:
- мальчики и девочки

Количественные переменные:
- количество деталей
- количество книг на полке

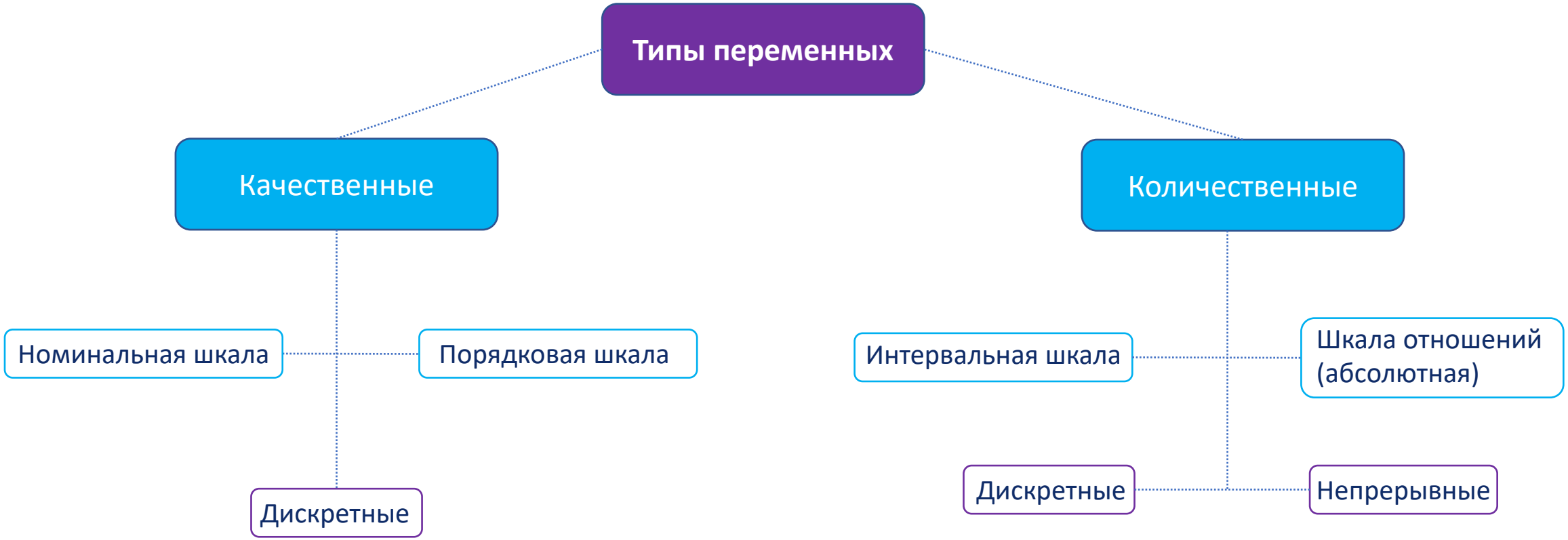
А из количественных переменных качественные

Количественные переменные:
- скорость
- частота пульса
- высота прыжка
- размер дистанции

Типы шкал


Типы переменных
Качественные
Количественные
Номинальная шкала
Порядковая шкала
Дискретные
Интервальная шкала
Шкала отношений (абсолютная)
Дискретные
Непрерывные

Типы шкал


Типы переменных
Качественные
Количественные
Номинальная шкала
Порядковая шкала
Дискретные
Интервальная шкала
Шкала отношений (абсолютная)
Дискретные
Непрерывные

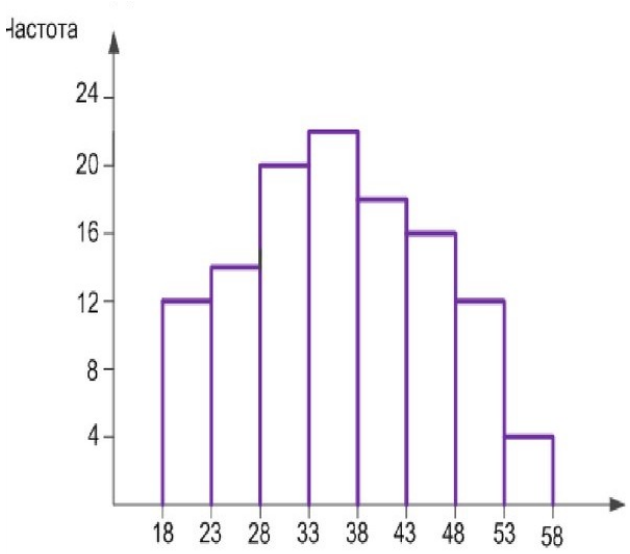
- Данные о поездках такси в Москве

Год	Сезон	Время суток	В пути, мин	Расстояние, км	Стоимость, руб	Руб/км	Температура	Час пик	Временной интервал, ч
2021	лето	день	89	58	1483	25,6	32	1	17:00-20:00
2021	лето	день	82	61	1859	30,5	32	0	14:00-16:00
2021	лето	ночь	17	15	680	45,3	15	0	23:00-24:00
2021	весна	день	27	8	410	51,3	16	0	13:00-14:00
2020	осень	вечер	26	23	410	17,8	5	0	21:00-23:00
2020	осень	день	42	23	654	28,4	9	1	17:00-20:00
2020	осень	утро	27	17	339	19,9	10	0	11:00-13:00
2020	осень	утро	32	19	472	24,8	11	0	11:00-13:00
2020	весна	утро	37	32	639	20,0	7	0	11:00-13:00
2020	зима	день	46	8,5	378	44,5	1	0	14:00-16:00
2020	зима	день	76	58	1362	23,5	2	1	17:00-20:00
2020	зима	утро	56	27	789	29,2	1	0	11:00-13:00
2020	зима	утро	69	30	755	25,2	-2	0	11:00-13:00
2019	лето	день	54	27	1016	37,6	22	0	14:00-16:00
2019	лето	день	6	4	199	49,8	13	0	14:00-16:00
2019	лето	утро	68	58	1623	28,0	29	1	8:00-10:00
2019	весна	день	20	8	260	32,5	7	0	14:00-16:00
2019	весна	вечер	19	8	300	37,5	4	0	21:00-23:00
2019	весна	вечер	37	39	801	20,5	0	0	21:00-23:00
2019	весна	день	25	16	477	29,8	0	1	17:00-20:00

Источник: данные агрегатора такси

Как анализировать данные (распределение)?

- Визуально
- Численно
  - Центр
  - Разброс
  - Другие характеристики



Распределение

Частотное распределение переменной – соответствие между значениями переменной и количеством таких значений в выборке

Распределение графически

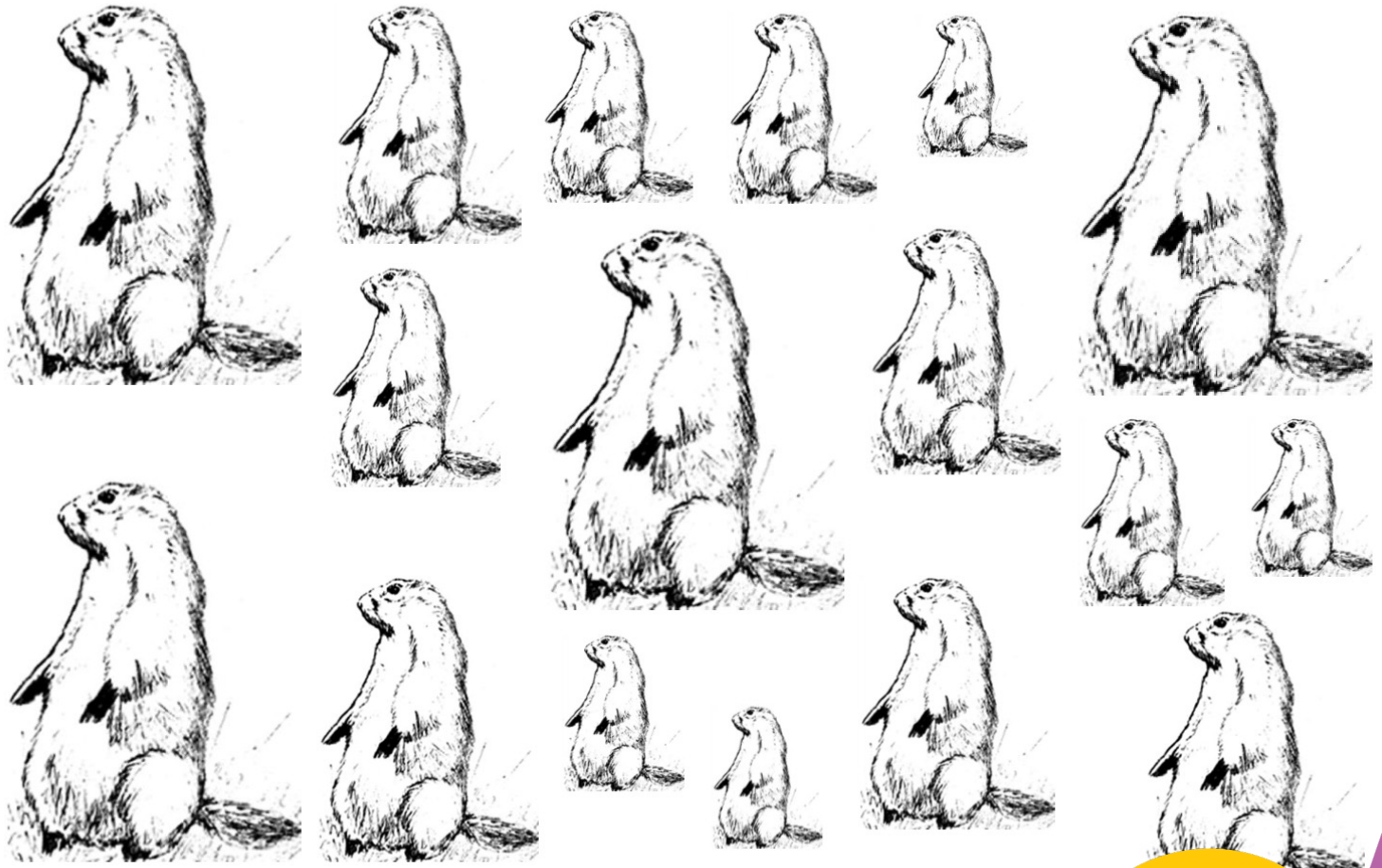
Гистограмма - графическое представление частотного распределения, разбитого по интервалам, где высота столбика отражает ЧАСТОТУ

Частота – то, сколько раз встретилось данное значение переменной

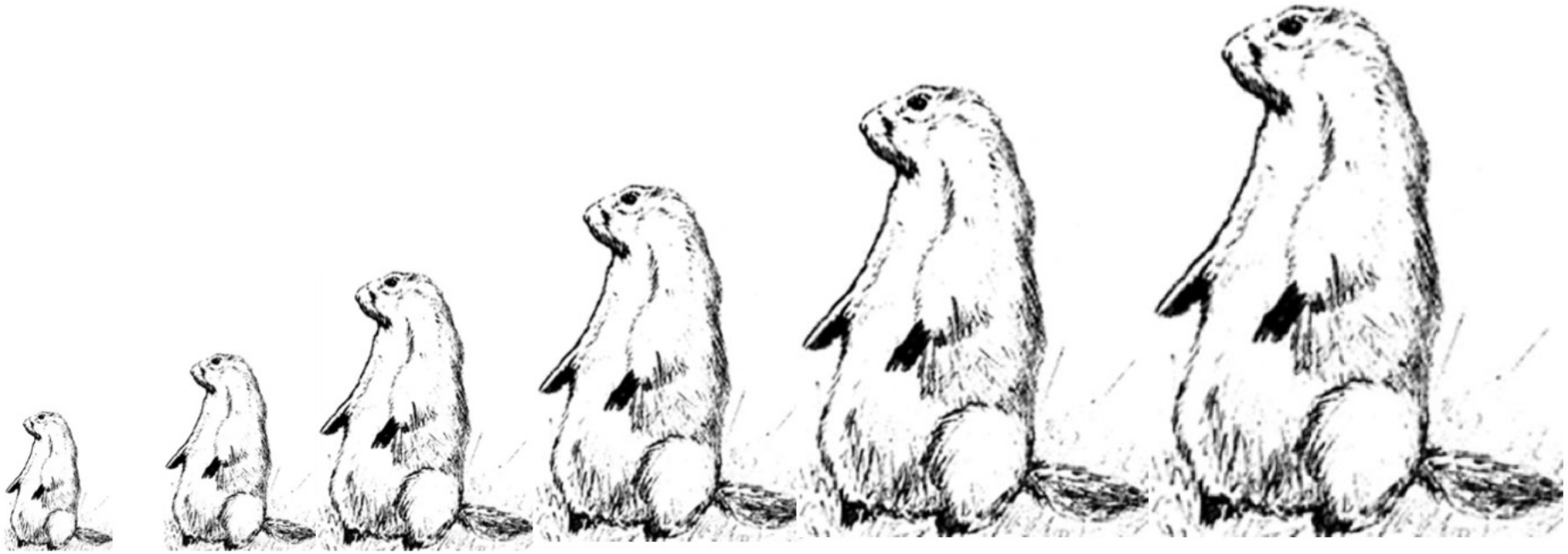
Карл Пирсон (1857 – 1936)

Предложил гистограмму в 1891 году

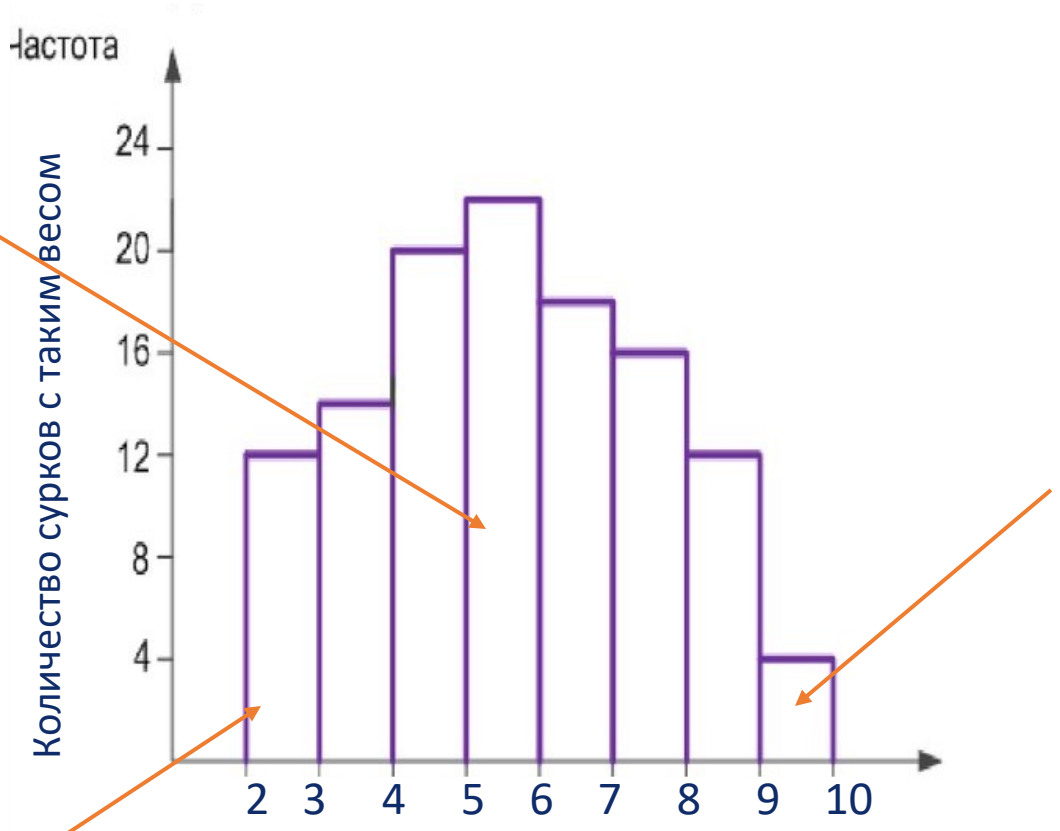
Возьмем N сурков



Упорядочим сурков и разобьем на интервалы



Нетяжелый сурок



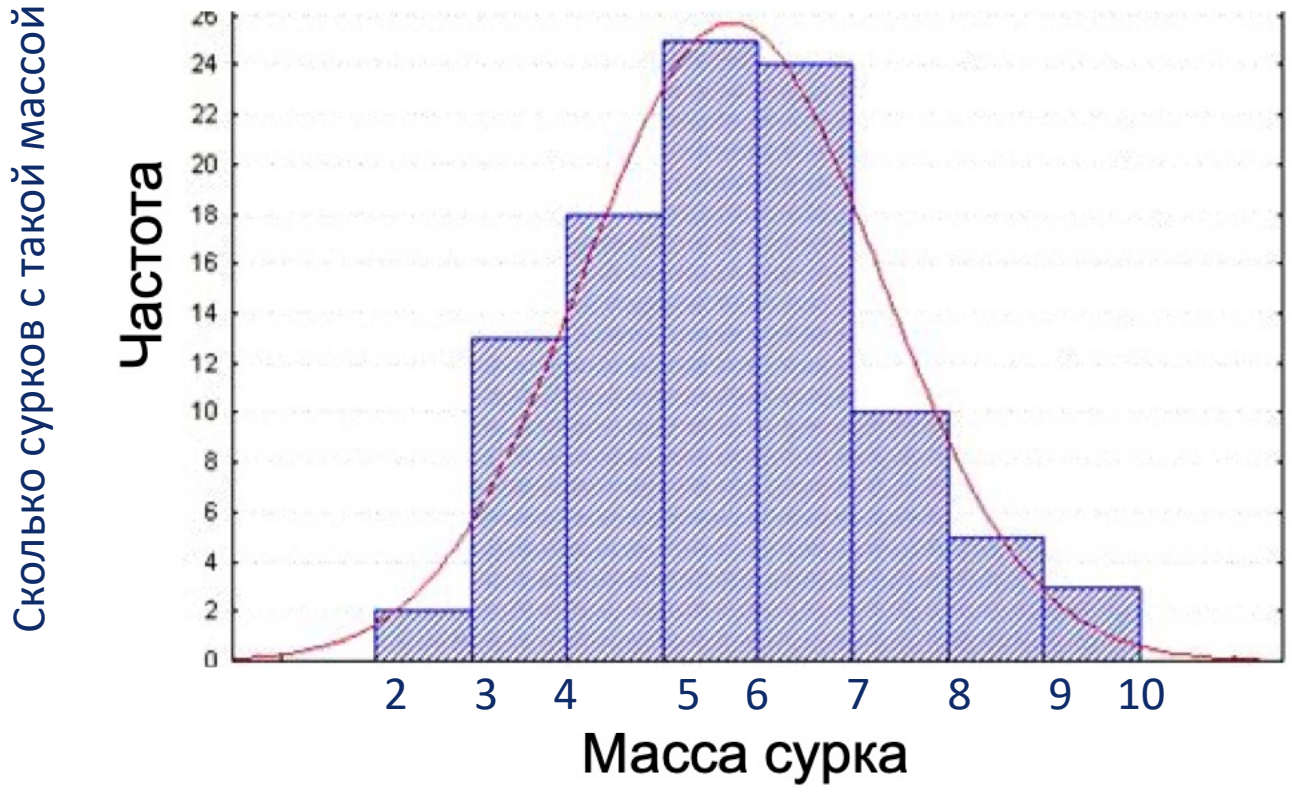
Очень тяжелый сурок

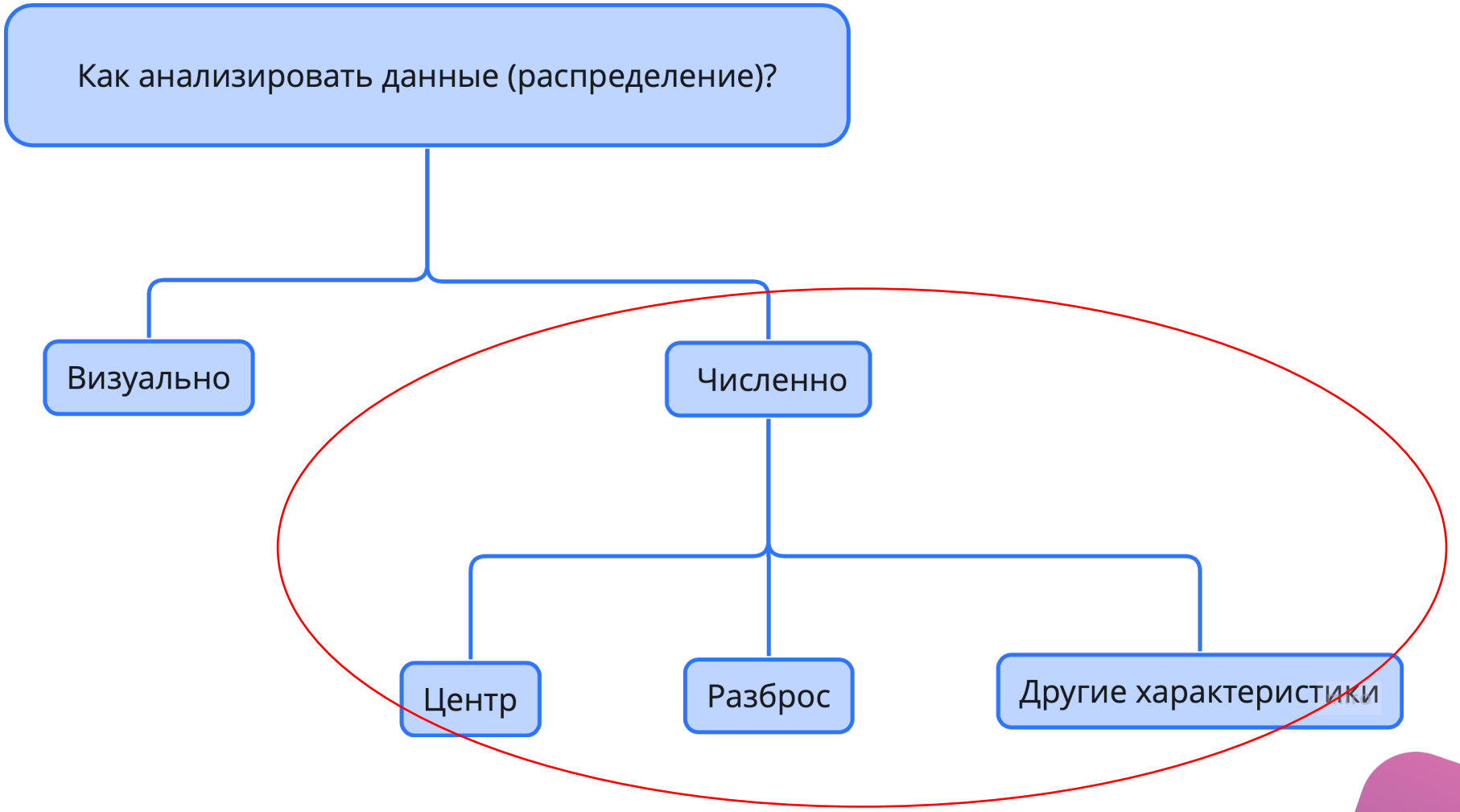
Степной сурок, или байбак принимает солнечные ванны. Зверек нагулял жир перед спячкой и выглядит упитанным.

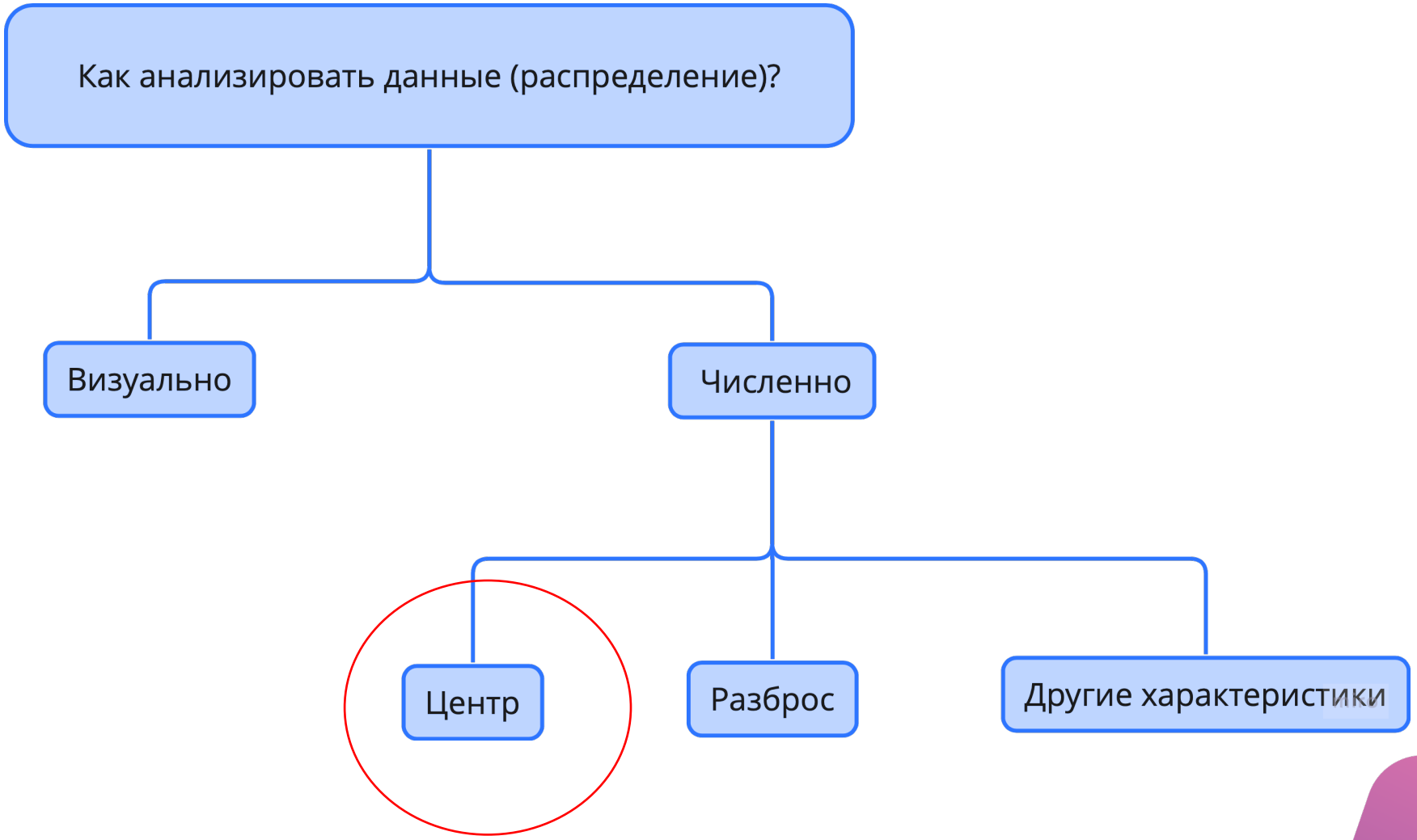
Малый сурок

Источник: https://animalsglobe.ru/surki

Гистограмма – зеркало распределения




Как анализировать данные (распределение)?
Визуально
Численно
Центр
Разброс
Другие характеристики


Как анализировать данные (распределение)?
Визуально
Численно
Центр
Разброс
Другие характеристики

Меры центра: мода, медиана, среднее

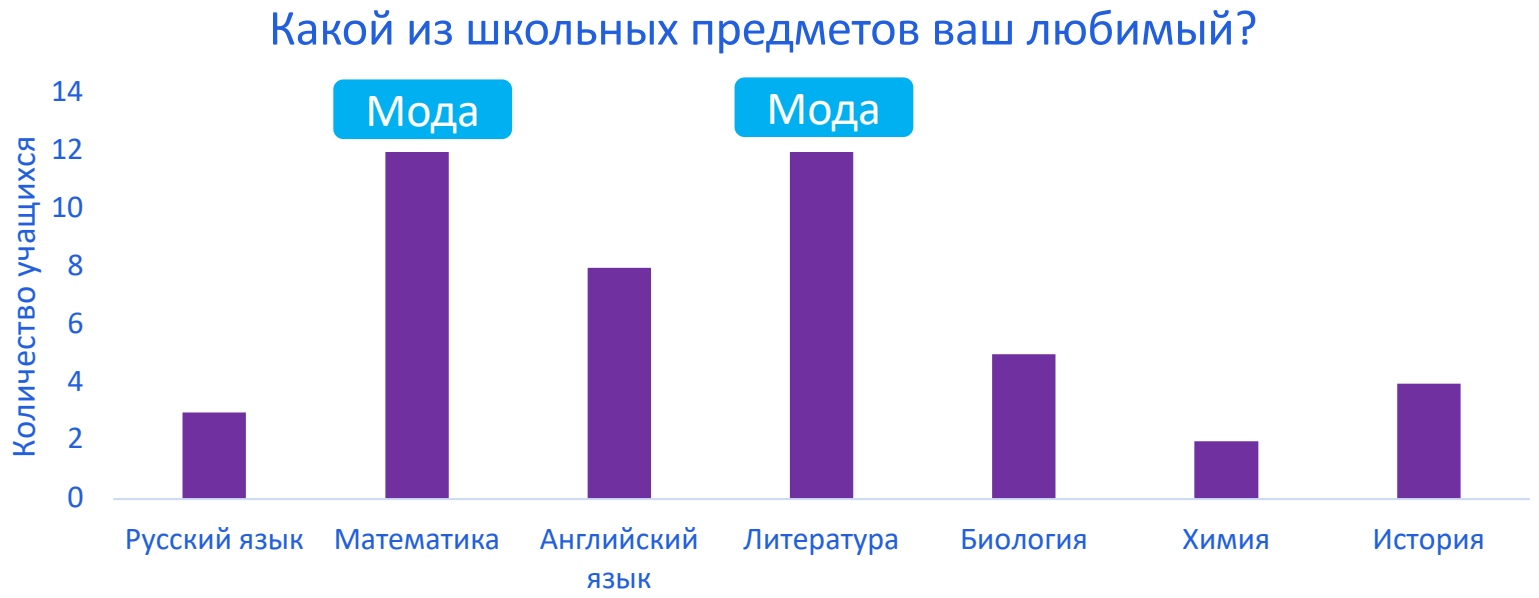
- **Меры центра**, или **центральной тенденции** – это числа, которые описывают множество значений одним числом (к примеру, среднюю температуру месяца вместо значений по каждому дню)

Меры центра: мода, медиана, среднее

	Определение	**Измерение**
Мода	Самое распространенное значение	Номинальный, порядковый, количественный
Медиана	Среднее значение упорядоченного ряда, как минимум половина значений больше или меньше медианы	Порядковый, количественный
Среднее (арифметическое, геометрическое, квадратическое)	1. $\overline{x_{\text{арифм}}} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum_1^n x_i}{n}$ 2. $x_{\text{геом}} = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$ 3. $x_{\text{квадр}} = \sqrt{\frac{\sum_1^n x_i^2}{n}}$	Количественный

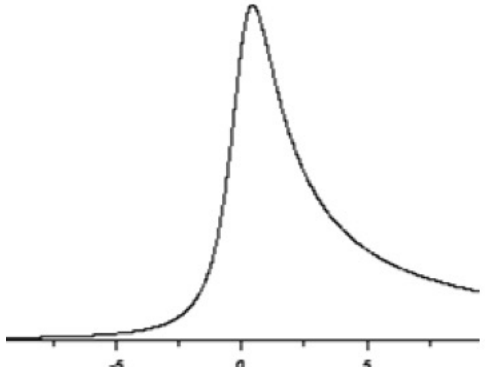
Мода

- **Мода** — значение во множестве наблюдений, которое встречается наиболее часто. (Мода = типичность)
- Иногда в совокупности встречается более чем одна мода: бимодальное распределение

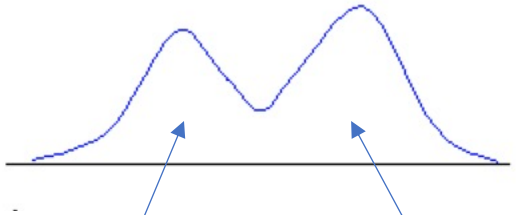


Распределения бывают разные по количеству мод

унимодальное



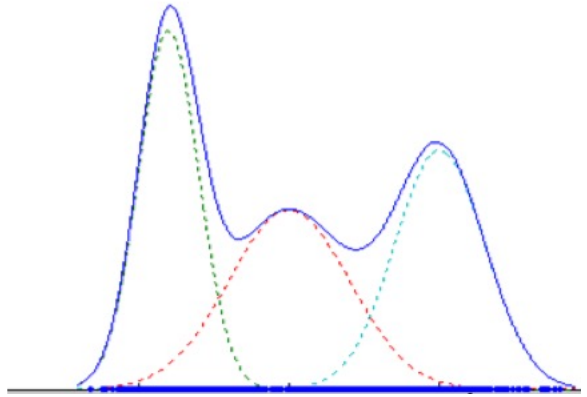
бимодальное



Группа легких сурков

Группа тяжелых сурков

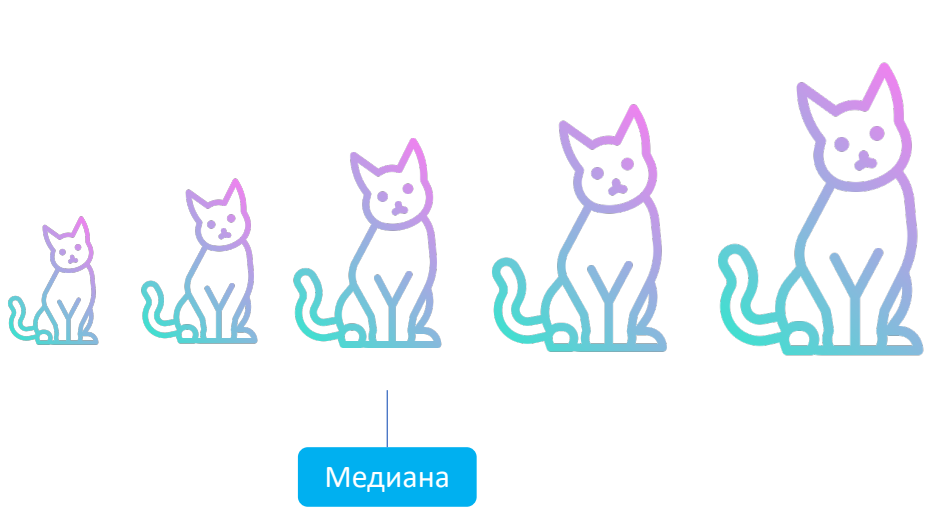
мультимодальное

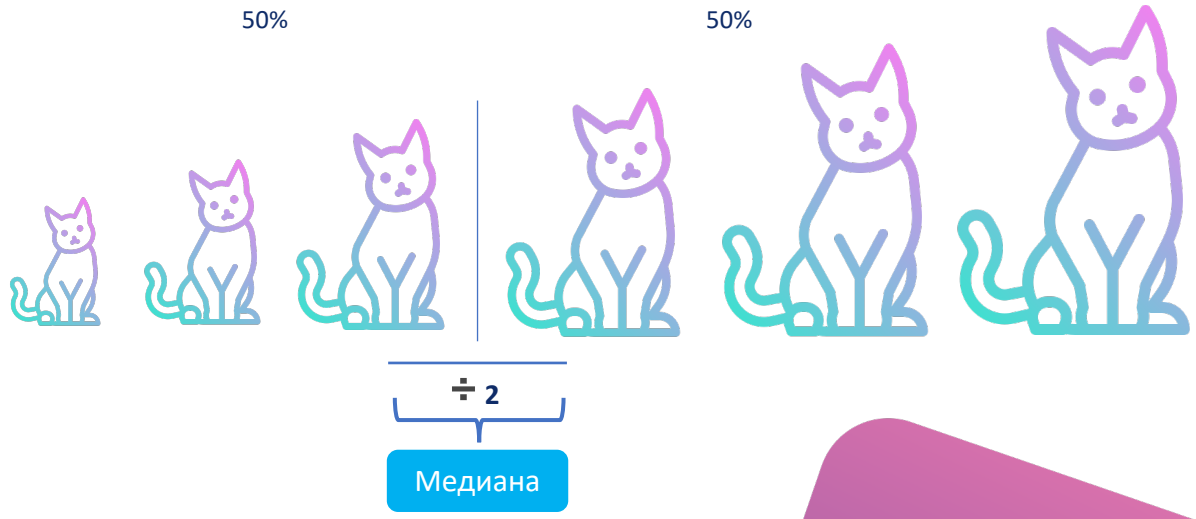


Внутри группы сурки больше похожи друг на друга по весу, чем по всей выборке

Медиана

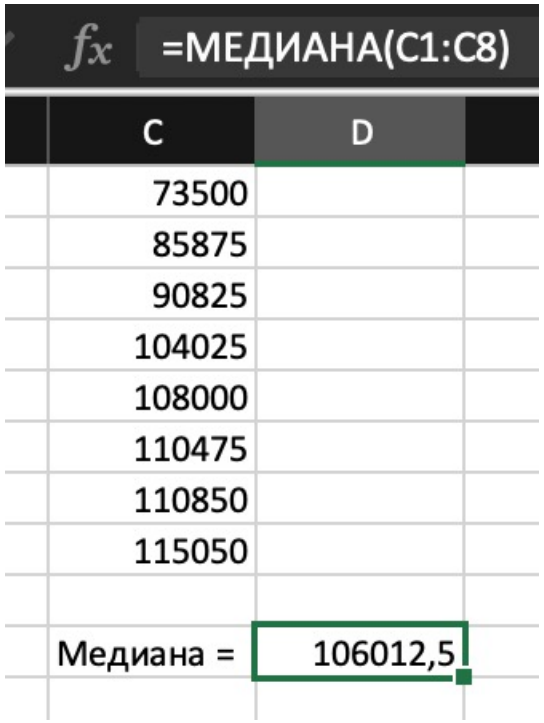
- **Медиана** или **серединное значение** набора чисел — число, которое находится в середине этого набора, если его упорядочить по возрастанию, то есть такое число, что половина из элементов набора не меньше него, а другая половина не больше
- В случае нечетного количества наблюдений медиана – серединное значение упорядоченного ряда
- В случае четного количества наблюдений медиана – среднее арифметическое двух серединных наблюдений



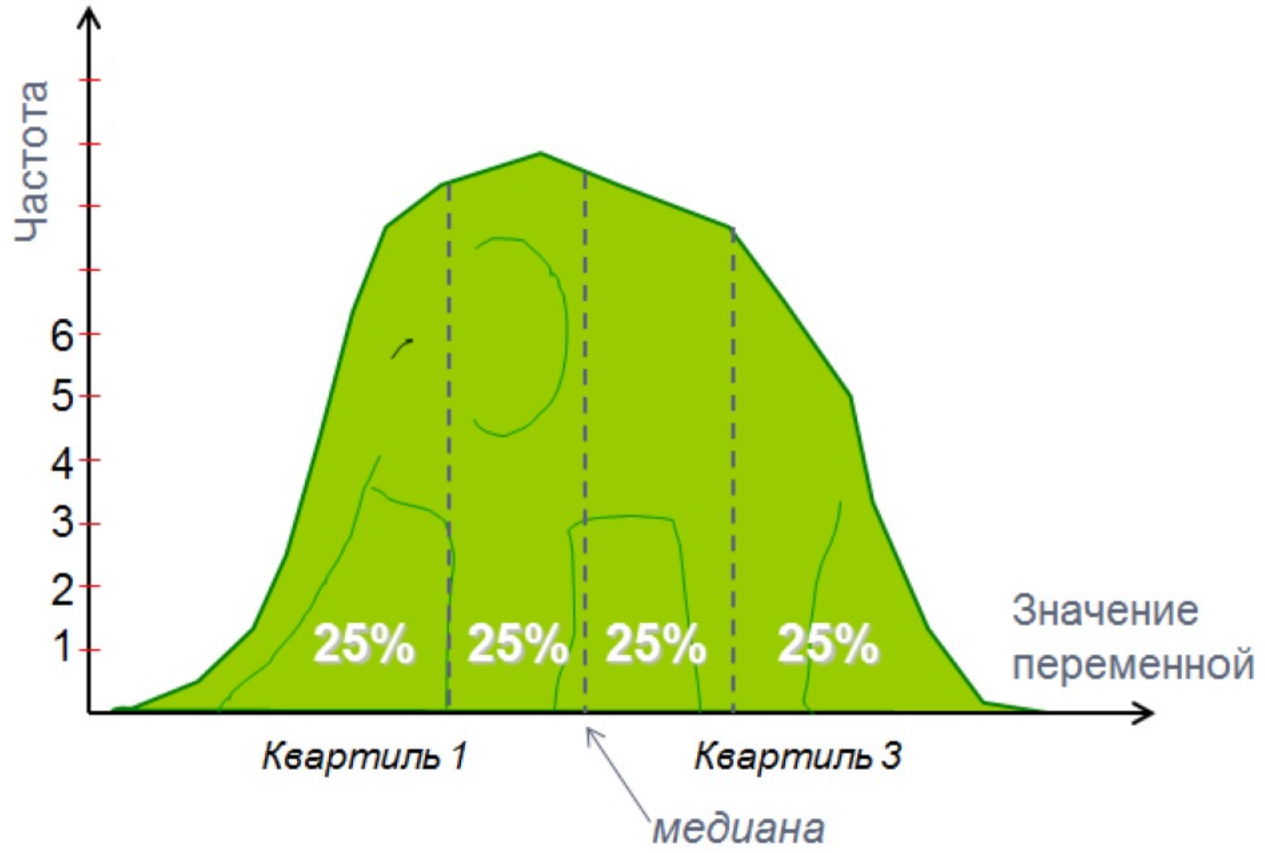


Медиана: четное и нечетное количество наблюдений

- **Нечетное количество наблюдений:**
- Есть данные о зарплате работников компании:
- 73500 ₽, 85875 ₽, 104025 ₽, 108000 ₽, 110475 ₽, 110850 ₽, 115050 ₽
- В данном случае медиана: 108000 ₽
- **Четное количество наблюдений:**
- Есть данные о зарплате другой группы работников компании :
- 73500 ₽, 85875 ₽, 90825 ₽, 104025 ₽, 108000 ₽, 110475 ₽, 110850 ₽, 115050 ₽
- В данном случае медиана: $\tilde{x} = \frac{104025+108000}{2} = 106012{,}5$ ₽



Медиана


Частота
6
5
4
3
2
1
25%
25%
25%
25%
Значение переменной
Квартиль 1
Квартиль 3
медиана

Среднее арифметическое

- **Среднее арифметическое** – то сумма отдельных значений, деленная на их количество
- **Самая известная** и часто используемая метрика статистики
  - $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Средняя зарплата работников компании:
  - $\bar{x} = \frac{73500+85875+90825+104025+108000+110475+110850+115050}{7} = \frac{798600}{7} \approx 114085{,}7$ ₽
- Медиана зарплат $106012{,}5$ ₽, а среднее – $114085{,}7$ ₽

Качественные переменные: мода, медиана, среднее

Сезон	Время суток	Час пик	Балл пробки
осень	вечер	нет	3
осень	день	да	7
зима	день	да	5
весна	день	нет	6
весна	вечер	нет	4

Признак	Среднее	Медиана	Мода
Сезон	?	?	?
Время суток	?	?	?
Час пик	?	?	?
Балл пробки	?	?	?

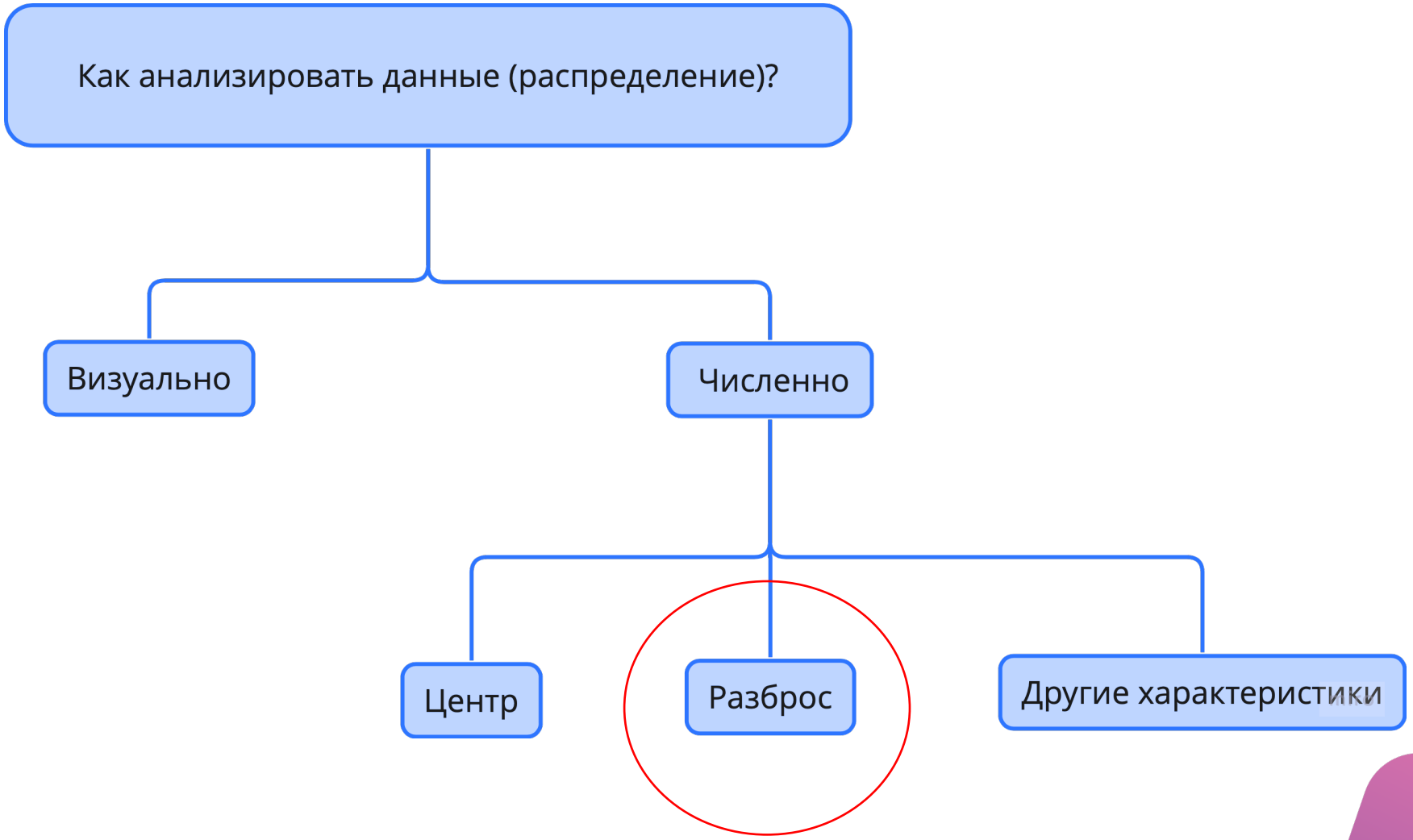
Качественные переменные: мода, медиана, среднее

Сезон	Время суток	Час пик	Балл пробки
осень	вечер	нет	3
осень	день	да	7
зима	день	да	5
весна	день	нет	6
весна	вечер	нет	4

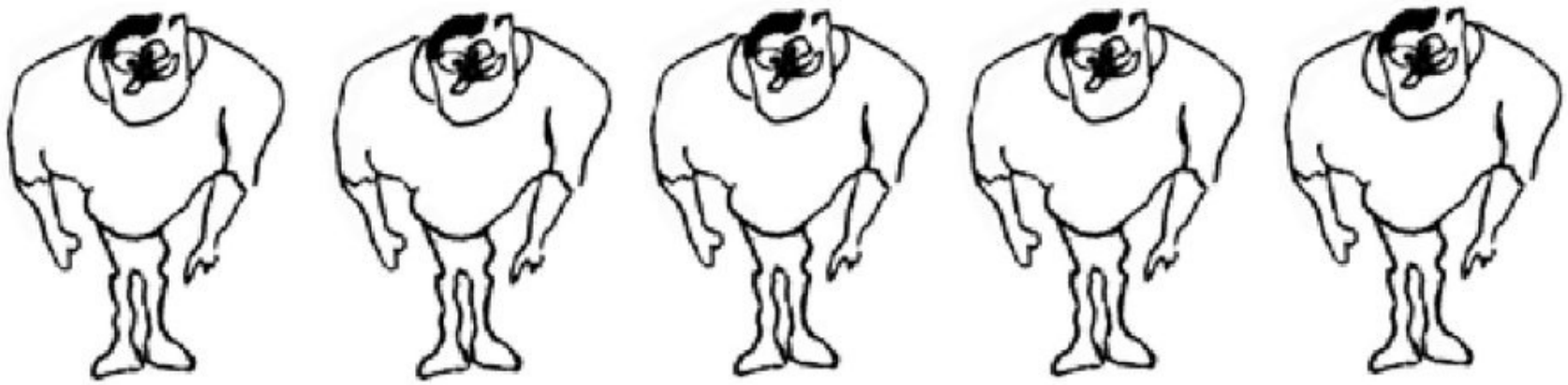
Признак	Среднее	Медиана	Мода
Сезон	-	-	+
Время суток	-	-	+
Час пик	-	-	+
Балл пробки	-	+	+

- **Какие типы переменной и меры центра подходят? Выберите наиболее подходящий.**
- Количество заработанных за ЛЭШ хсевро
  - ✓ Тип переменной: количественная
  - ✓ Мера центра: среднее арифметическое
- Доверие своему вожатому (от очень доверяю до очень не доверяю)
  - ✓ Тип переменной: порядковый
  - ✓ Мера центра: медиана
- Принадлежность к программе (БКиАД или Олимпиадная программа)
  - ✓ Тип переменной: номинальная
  - ✓ Мера центра: мода


В среднем
у каждого
по два пирожка
DEMOTIVATORS.RU
Статистика - вещь упрямая

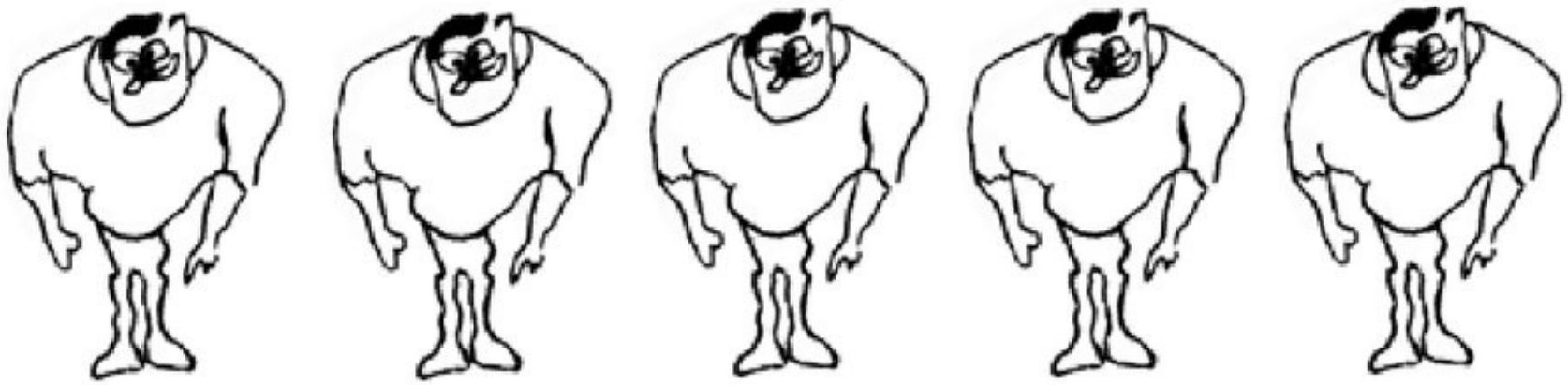

Как анализировать данные (распределение)?
Визуально
Численно
Центр
Разброс
Другие характеристики

Меры изменчивости (разброса)



Средний вес команды = 95 кг

Меры изменчивости (разброса)



Средний вес команды = 95 кг

Меры изменчивости (разброса)

Меры разброса

	Определение	Измерение
Размах (range)	Разница между самым большим и самым маленьким значением	Количественный
Дисперсия (variance)	Среднеквадратичное отклонение всех наблюдений от среднего значения	Количественный
Стандартное отклонение (standard deviation)	Квадратный корень из дисперсии	Количественный
Межквартильный размах (interquartile range)	Разница между 1-м и 3-м квартилем	Порядковый, количественный
Коэффициент вариации (coefficient of variance)	Стандартное отклонение, деленное на среднее	Количественный

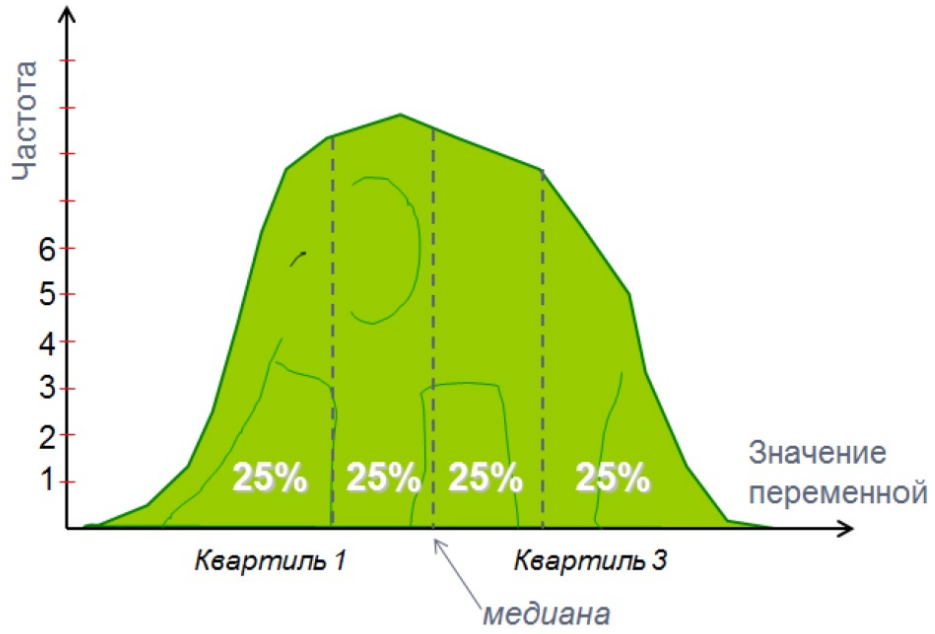
Абсолютные

Размах

- **Размах** – разница между самым большим и самым малым наблюдением
- Дает относительно **мало информации**
- Подвержен **выбросам**
- **Максимум** (max) – наибольшее значение, **минимум** (min) – наименьшее значение
  - $R = x_{max} - x_{min}$
- **Пример:**
- 73500 ₽, 85875 ₽, 90825 ₽, 104025 ₽, 108000 ₽, 110475 ₽, 110850 ₽, 115050 ₽
- Разброс зарплат работников компании:
  - $R = 115050 - 73500 = 41550$ ₽

Межквартильный размах

- Квартили делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана)
- Межквартильный размах – разница между 3м и 1м квартилем



Дисперсия: что это такое?

- **Дисперсия** – это разброс значений какой то величины от его среднего (математического ожидания)
- Для чего нужна дисперсия?
- Предположим, что у нас есть данные о средней зарплате на предприятии – 70000 ₽. Прожиточный минимум в регионе – 12000 ₽. Можем ли мы говорить, что на предприятии все в порядке с доходом работников?
- **Конечно, нет.**
- Может быть такая ситуация, что доходы распределились следующим образом:
- 8000 ₽, 10000 ₽, 7000 ₽, 155000 ₽, 170000 ₽
- А может быть следующая ситуация:
- 60000 ₽, 80000 ₽, 70000 ₽, 50000 ₽, 90000 ₽

Среднее = 70 000 ₽, Дисперсия = 7 159 500 000 ₽,
Ст. отклонение = 84 614 ₽

Среднее = 70 000 ₽, Дисперсия = 250 000 000₽,
Ст. отклонение = 15 811,4 ₽

Дисперсия и стандартное отклонение

- Дисперсия
- $s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Стандартное отклонение
  - $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Качественные переменные: дисперсия

Сезон	Время суток	Час пик	Балл пробки
осень	вечер	нет	3
осень	день	да	7
зима	день	да	5
весна	день	нет	6
весна	вечер	нет	4

	Номинальные	Порядковые
Разброс	?	?
Дисперсия	?	?
Стандартное отклонение	?	?
Межквартильный разброс	?	?

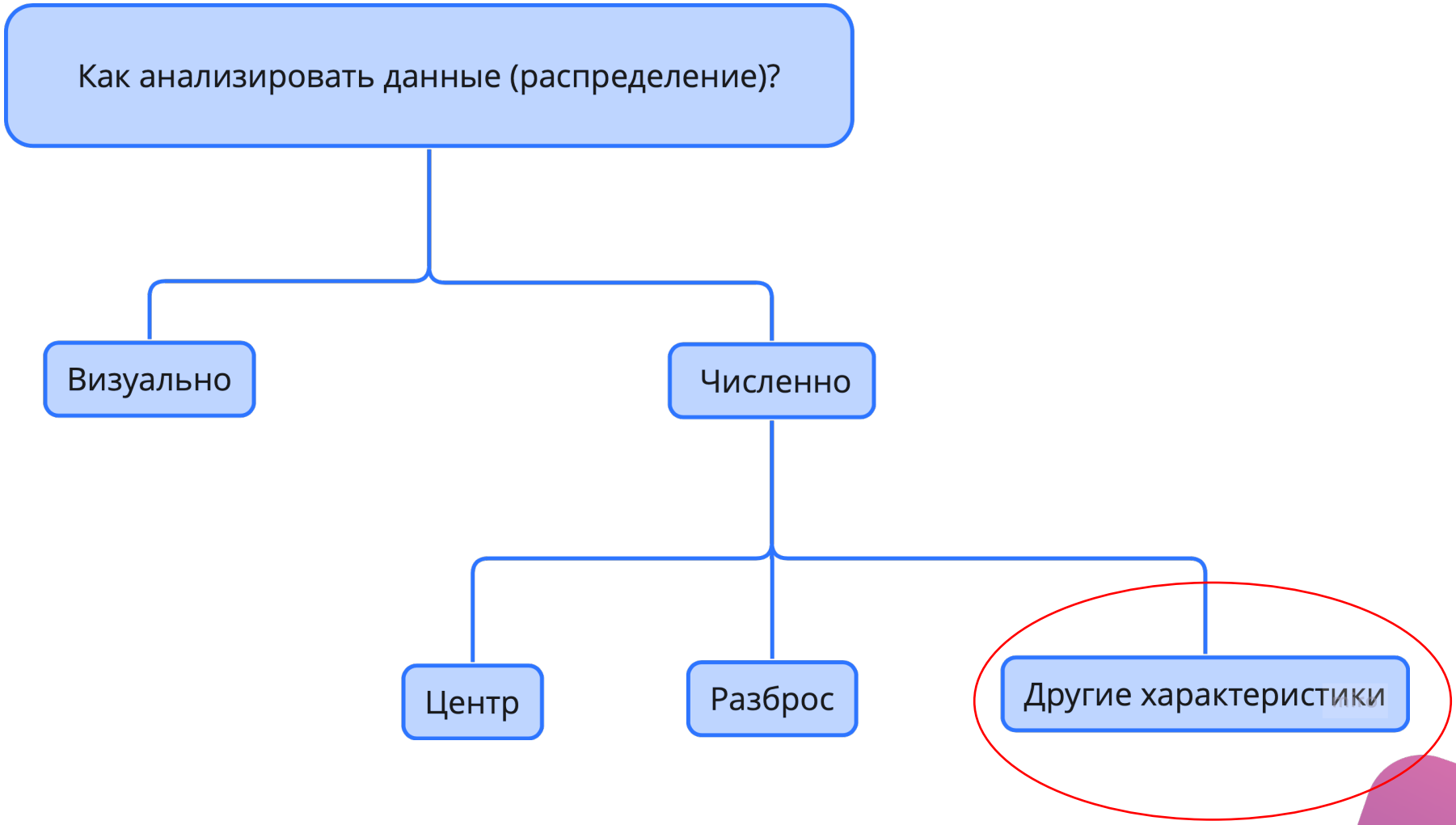
Источник: данные агрегатора такси

Качественные переменные: дисперсия

Сезон	Время суток	Час пик	Балл пробки
осень	вечер	нет	3
осень	день	да	7
зима	день	да	5
весна	день	нет	6
весна	вечер	нет	4

	Номинальные	Порядковые
Разброс	-	-
Дисперсия	-	-
Стандартное отклонение	-	-
Межквартильный разброс	-	+

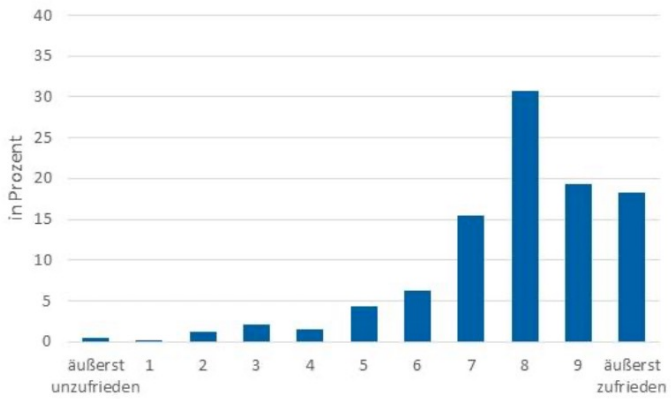
Источник: данные агрегатора такси

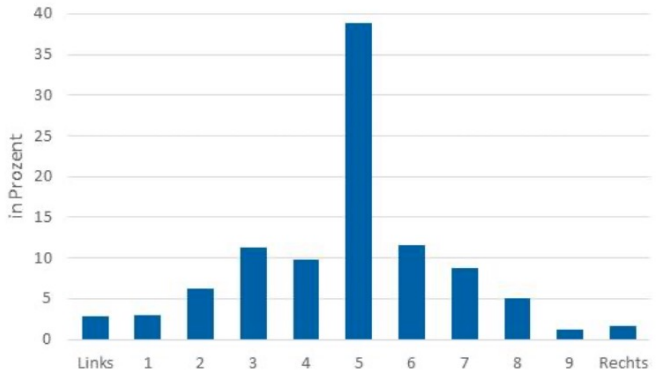

Как анализировать данные (распределение)?
Визуально
Численно
Центр
Разброс
Другие характеристики

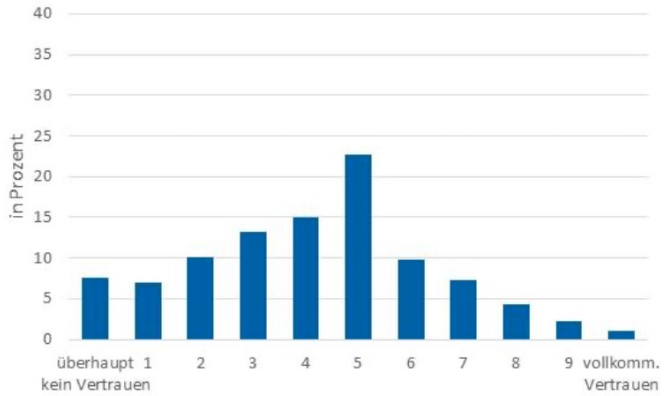
Асимметрия (skewness)

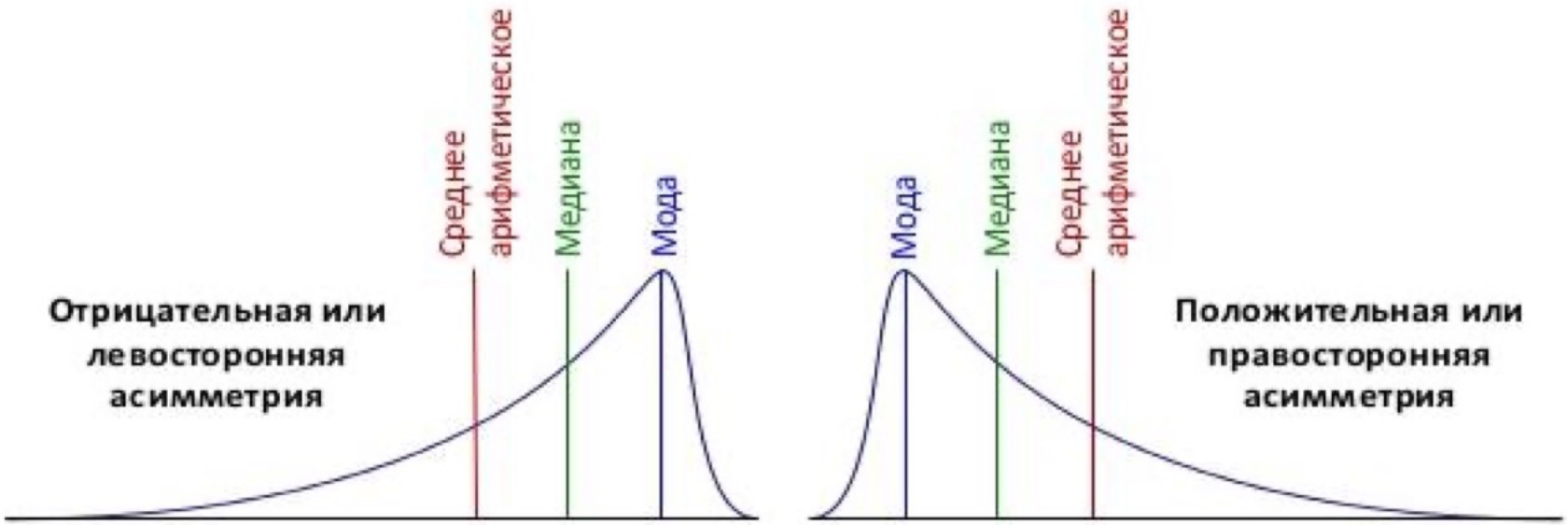
= СКОС(выделяете нужные ячейки)

Асимметрия (skewness)

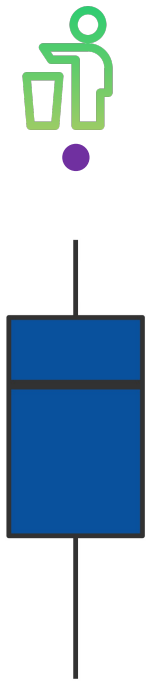




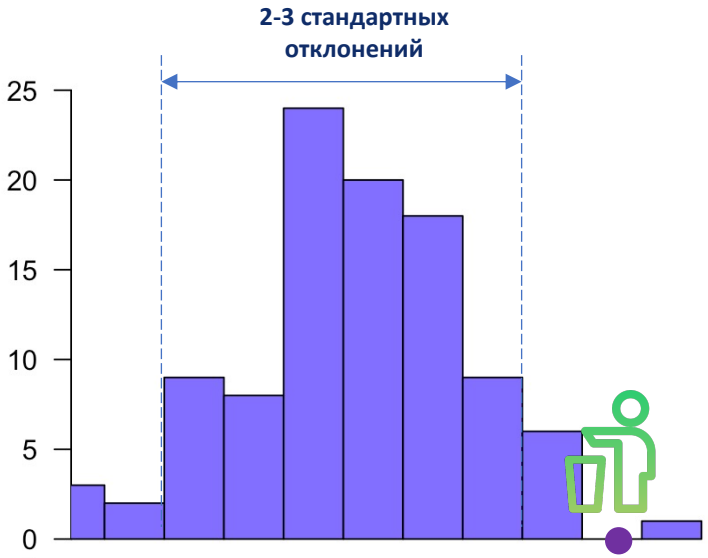



Среднее арифметическое
Медиана
Мода
Отрицательная или левосторонняя асимметрия
Мода
Медиана
Среднее арифметическое
Положительная или правосторонняя асимметрия

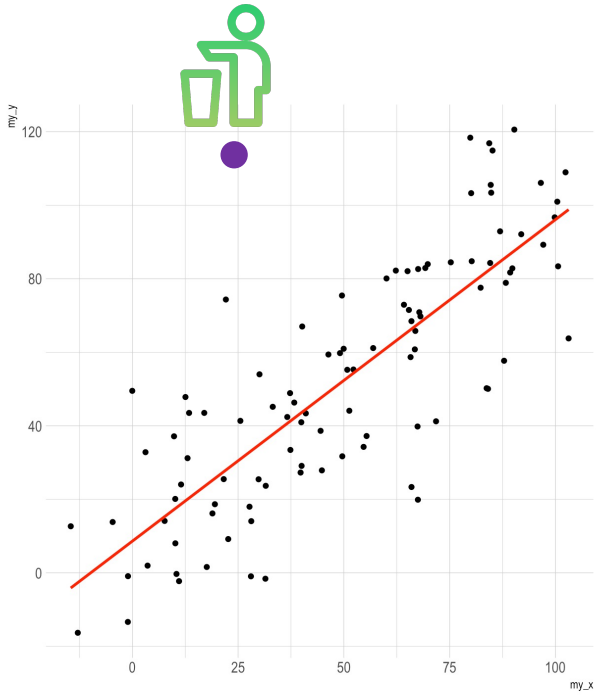
Что может исказить результаты анализа?



Box Plot



Распределение



Регрессия

Источник: данные агрегатора такси

Задание после лекции командам:

1. Выделите в базе интересующие качественные и количественные признаки.

2. Какие из пройденных инструментов анализа могут быть им полезны для анализа качественных признаков? Какие будут полезны для анализа количественных признаков?

3. Можете ли вы как-то разделить выборку так, чтобы наблюдения были больше похожи друг на друга? По какому признаку лучше это сделать?