Конспект по теме «Описательная статистика»

Среднее

Типы данных:

- **Категориальные** данные описывают качественные характеристики и могут быть разделены на различные группы или категории. Например, пол, цвет глаз или марка автомобиля.
- **Порядковые** данные, как и категориальные, описывают качественные характеристики, но их значения можно проранжировать. Например, размер одежды, уровень образования.
- **Числовые** данные — измеримые или счётные значения. Например, возраст, доход, рост.

Методы визуализации: столбчатая диаграмма и гистограмма.

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

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

где n — количество элементов в выборке.

Медиана

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

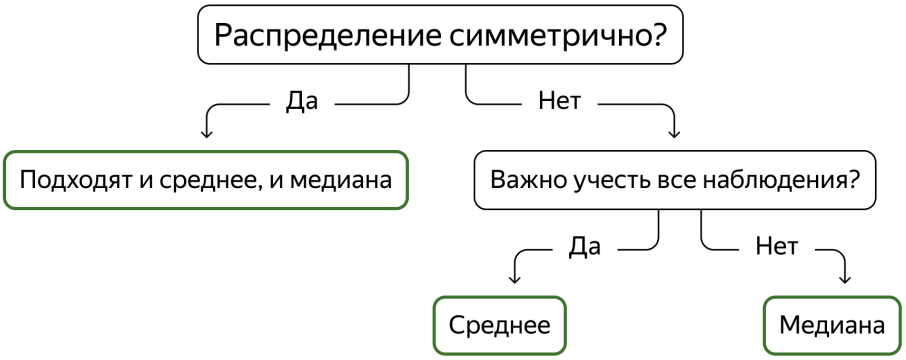
Алгоритм вычисления медианы:

1. Упорядочить элементы в списке по возрастанию.
2. Посчитать количество элементов в списке.
3. а) Если число элементов в списке нечётное, найти число, стоящее посередине.

б) Если число элементов чётное, найти два числа, которые находятся посередине, сложить их и результат разделить пополам.

Распределение называется **симметричным**, если оно выглядит одинаково с обеих сторон от своей середины.

Как выбрать меру центральной тенденции



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.

Алгоритм определения α-квантиля

1. Отсортируйте набор данных по возрастанию.
2. Найдите позицию квантиля по формуле: $n \cdot \alpha$, где n — количество элементов в наборе, α — доля, которая нас интересует.
3. Определите значение квантиля:
   а) Если позиция квантиля — целое число,
   α-квантиль равен значению, которое соответствует этой позиции в упорядоченном наборе данных.
   б) Если позиция квантиля — дробное число, возьмите среднее значение между двумя ближайшими соседями.

Перцентиль — это то же самое, что и квантиль, но в процентах.

Квартили делят выборку на 4 равные части:

- $\mathrm{Q1}$ (первый квартиль) — это 0.25-квантиль,
- $\mathrm{Q2}$ (второй квартиль) — это 0.5-квантиль (медиана),
- $\mathrm{Q3}$ (третий квартиль) — это 0.75-квантиль.

Межквартильный размах $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$.

Выброс — это значение или набор значений в наборе данных, который сильно отличается от остальных.

Алгоритм отсеивания выбросов:

1. Отсортируем данные в возрастающем порядке.
2. Вычислим первый квартиль $\mathrm{Q1}$ (0.25-квантиль) и третий квартиль $\mathrm{Q3}$ (0.75-квантиль).
3. Рассчитайте межквартильный размах: $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$.
4. Определите границы выбросов:
   - Нижняя граница: $\mathrm{Q1} - 1.5 \cdot \mathrm{IQR}$,
   - Верхняя граница: $\mathrm{Q3} + 1.5 \cdot \mathrm{IQR}$.
5. Отсейте все значения, которые лежат за пределами этих границ, — они считаются выбросами.

Диаграмма «ящик с усами»:

Ящик:

- Левая граница ящика — первый квартиль ($\mathrm{Q1}$), то есть 25% данных лежат ниже этой точки.
- Медиана (второй квартиль, $\mathrm{Q2}$) представлена вертикальной линией внутри ящика. Медиана делит данные на две равные части: половина данных лежит ниже этой линии, а половина — выше.
- Правая граница ящика — третий квартиль ($\mathrm{Q3}$), то есть 75% данных лежат ниже этой точки.

Усы:

- Левый ус ограничивается значением $\mathrm{Q1} - 1.5 \cdot \mathrm{IQR}$.
- Правый ус ограничивается значением $\mathrm{Q3} + 1.5 \cdot \mathrm{IQR}$.

Выбросы обозначают кружочками.

Дисперсия

Дисперсия — это статистический показатель, который описывает разброс значений в наборе данных относительно среднего значения.

$$\mathrm{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Стандартное отклонение — это квадратный корень дисперсии, оно измеряется в тех же единицах, что и исходные данные.

$s_X = \sqrt{Var(X)}$.