L7 Визуализация данных и введение в статистический анализ

Ксения Балабаева
20/10/2025

Меры центральной тенденции: где находится "центр" данных?

Меры центральной тенденции описывают типичное или центральное значение в наборе данных. Две основные меры — среднее и медиана — по-разному определяют "центр" и имеют различные свойства.

Среднее (Mean, μ)

Сумма всех значений, деленная на их количество.

Формула: $\mu = (\Sigma x_i) / n$

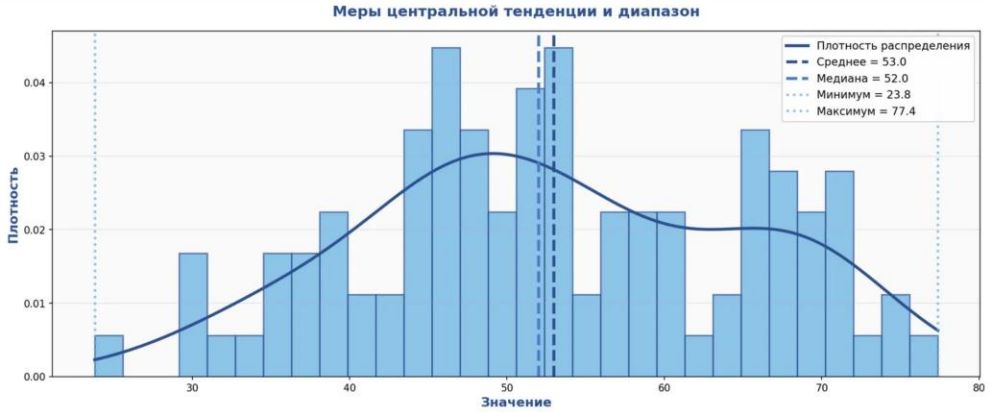
Свойство: Чувствительно к выбросам — экстремальные значения сильно влияют на среднее.

Медиана (Median)

Значение, делящее упорядоченный набор данных пополам.

50% наблюдений меньше медианы, 50% больше

Свойство: Устойчива к выбросам — экстремальные значения не влияют на медиану.



Минимум и максимум: границы диапазона данных

Минимум и максимум определяют границы, в которых находятся все значения данных. Разность между ними называется диапазоном и показывает общий разброс данных.

min

Минимум (min)

Наименьшее значение в наборе данных. Определяет нижнюю границу распределения.

max

Максимум (max)

Наибольшее значение в наборе данных. Определяет верхнюю границу распределения.

R

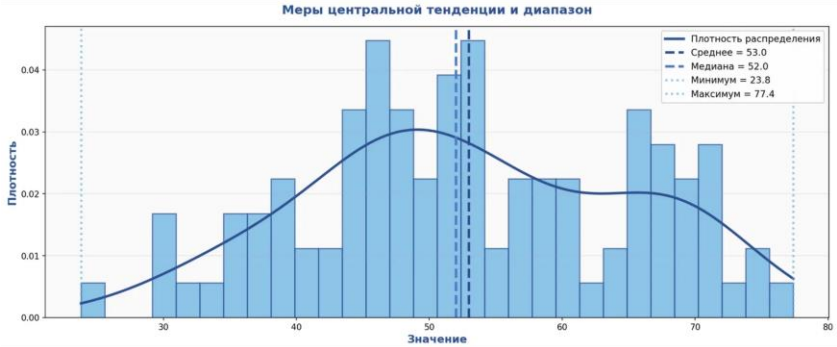
Диапазон (Range)

Разность между максимумом и минимумом.

Формула: Range = max – min

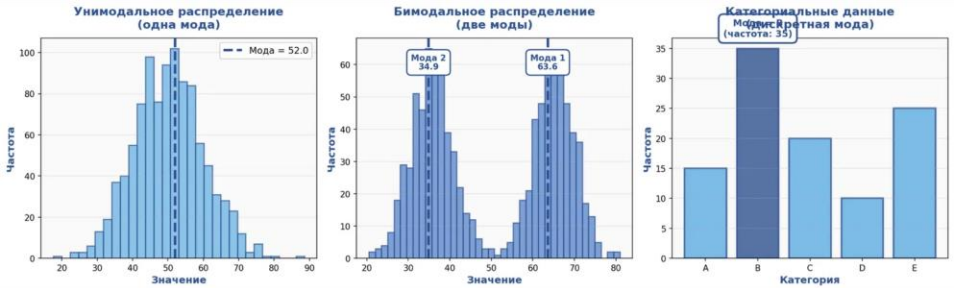
Применение

Быстрая оценка разброса данных, выявление экстремальных значений, проверка диапазона допустимых значений.



Мода: наиболее частое значение в данных

Мода (mode) — значение, которое встречается в наборе данных наиболее часто. В отличие от среднего и медианы, мода применима как к числовым, так и к категориальным данным.



Типы распределений по количеству мод:

Унимодальное распределение

Имеет одну моду — один явный пик частоты. Характерно для нормального распределения.

Бимодальное распределение

Имеет две моды — два пика частоты. Может указывать на наличие двух различных групп в данных.

Мультимодальное распределение

Имеет более двух мод — несколько пиков частоты. Указывает на сложную структуру данных.

Особенности и применение

Для категориальных данных: Мода — единственная мера центральной тенденции, применимая к номинальным данным (например, цвет, пол, категория товара).

Устойчивость: Мода не зависит от выбросов, так как учитывает только частоту появления значений.

Квантили: деление распределения на равные части

Квантиль — значение, ниже которого находится определенная доля наблюдений в упорядоченном наборе данных.

Квартили делят данные на 4 части:

Q1 (25%) — Первый квартиль
25% данных меньше этого значения

Q2 (50%) — Второй квартиль = Медиана
50% данных меньше этого значения

Q3 (75%) — Третий квартиль
75% данных меньше этого значения

Межквартильный размах (IQR)

Содержит центральные 50% данных.

Формула: IQR = Q3 – Q1

Процентили делят данные на 100 частей (например, 95-й процентиль означает, что 95% данных меньше этого значения).

Стандартное отклонение: мера разброса данных

Стандартное отклонение показывает, насколько в среднем значения отклоняются от среднего. Это одна из важнейших мер разброса данных, широко используемая в статистике и машинном обучении.

Дисперсия (Variance, σ²)

Среднее квадратов отклонений от среднего значения. Измеряется в квадратах единиц исходных данных.

$$\sigma^2 = \Sigma(x_i - \mu)^2 / n$$

где x_i — значения данных, μ — среднее, n — количество наблюдений

Стандартное отклонение: мера разброса данных

Стандартное отклонение показывает, насколько в среднем значения отклоняются от среднего. Это одна из важнейших мер разброса данных, широко используемая в статистике и машинном обучении.

Дисперсия (Variance, σ^2)

Среднее квадратов отклонений от среднего значения. Измеряется в квадратах единиц исходных данных.

$$\sigma^2 = \Sigma(x_i - \mu)^2 / n$$

где x_i — значения данных, μ — среднее, n — количество наблюдений

Стандартное отклонение (σ)

Квадратный корень из дисперсии. Измеряется в тех же единицах, что и исходные данные, что делает его более интерпретируемым.

$$\sigma = \sqrt{(\sigma^2)}$$

или эквивалентно:

$$\sigma = \sqrt{[\Sigma(x_i - \mu)^2 / n]}$$

Интерпретация

Чем больше σ, тем больше разброс данных. Малое стандартное отклонение означает, что значения сконцентрированы близко к среднему. Большое стандартное отклонение указывает на широкий разброс значений. Преимущество σ перед σ^2 в том, что оно измеряется в тех же единицах, что и исходные данные (например, если данные в сантиметрах, то σ также в сантиметрах).

Правило трех сигм: распределение данных вокруг среднего

Для нормального распределения правило трех сигм описывает, какая доля данных находится в пределах одного, двух и трех стандартных отклонений от среднего значения.

Интервалы стандартных отклонений

±1σ (одна сигма)

Содержит **~68.27%** всех данных

±2σ (две сигмы)

Содержит **~95.45%** всех данных

±3σ (три сигмы)

Содержит **~99.73%** всех данных

Применение и Z-score

Выявление аномалий: Значения за пределами ±3σ считаются редкими и потенциально аномальными.

Стандартизация (z-score):

$z = (x - \mu) / \sigma$

Показывает, на сколько стандартных отклонений значение отличается от среднего.

Box Plot: визуализация всех ключевых статистик

Box Plot (ящик с усами) объединяет медиану, квартили, межквартильный размах, минимум, максимум и выбросы в одной компактной визуализации, позволяя быстро оценить распределение данных.

Элементы Box Plot

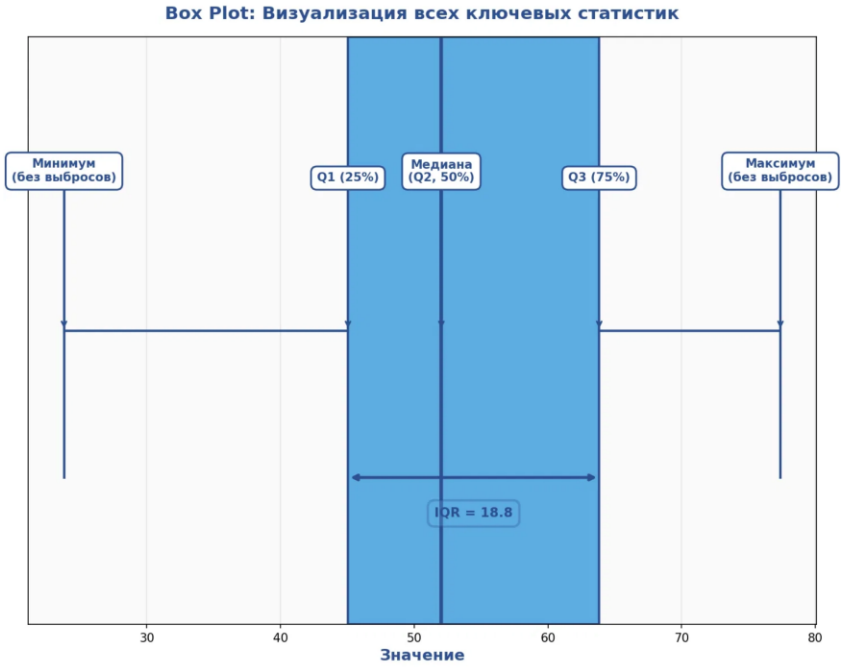
- **Центральная линия:** медиана (Q2, 50-й процентиль)
- **Границы коробки:** Q1 (25%) и Q3 (75%)
- **Высота коробки:** IQR (межквартильный размах)
- **Усы (whiskers):** минимум и максимум без выбросов
- **Точки за усами:** выбросы (outliers)

Правило выбросов

Значения за пределами интервала **[Q1 − 1.5×IQR, Q3 + 1.5×IQR]** считаются выбросами и отображаются отдельными точками.

Преимущество

Компактное представление распределения, легко сравнивать несколько групп данных, устойчивость к выбросам.



Гистограмма и Kernel Density Estimation (KDE)

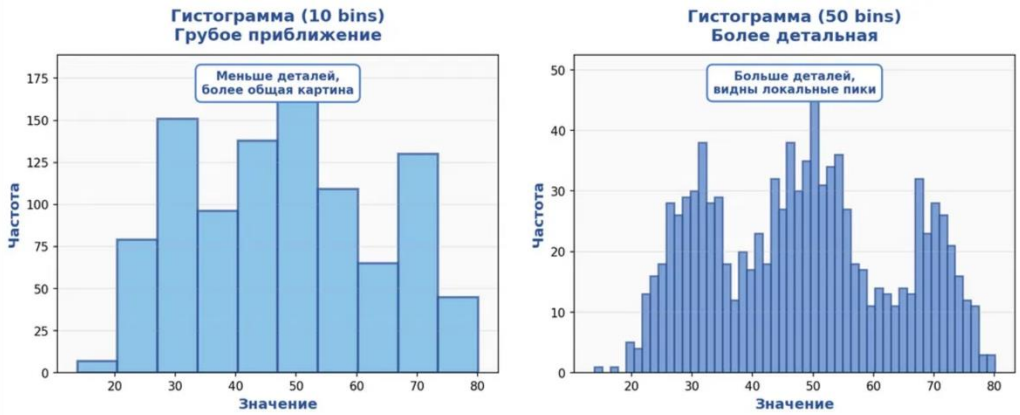
Гистограмма (Histogram)

Визуализация распределения данных путем разбиения диапазона на интервалы (bins) и подсчета частоты попадания значений в каждый интервал.

Ключевой параметр: количество bins. Малое количество дает грубое приближение, большое — более детальную картину, но может показывать шум.

Преимущества гистограммы:

- Простота интерпретации
- Показывает фактические частоты
- Быстрое построение



Гистограмма и Kernel Density Estimation (KDE)

Гистограмма (Histogram)

Визуализация распределения данных путем разбиения диапазона на интервалы (bins) и подсчета частоты попадания значений в каждый интервал.

Ключевой параметр: количество bins. Малое количество дает грубое приближение, большое — более детальную картину, но может показывать шум.

Kernel Density Estimation (KDE)

Непрерывная сглаженная оценка плотности вероятности. Вместо дискретных столбцов создает плавную кривую, используя ядерную функцию (kernel) вокруг каждой точки данных.

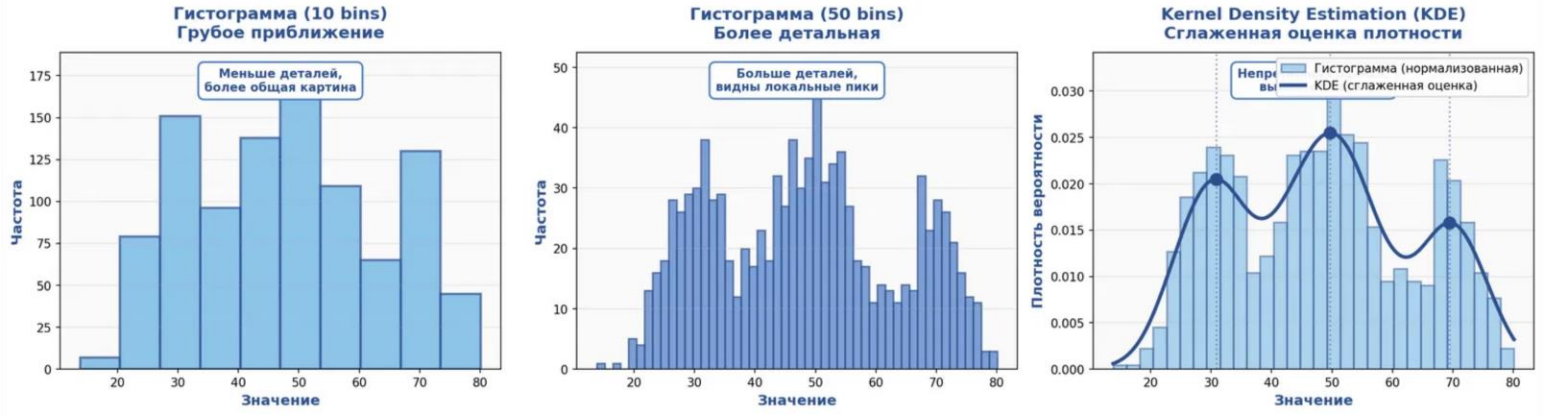
Ключевой параметр: bandwidth (ширина полосы). Контролирует степень сглаживания: малый bandwidth → детально, большой → сглажено.

Преимущества гистограммы:

- Простота интерпретации
- Показывает фактические частоты
- Быстрое построение

Преимущества KDE:

- Непрерывная оценка плотности
- Четко выявляет моды
- Не зависит от выбора границ bins



Корреляция

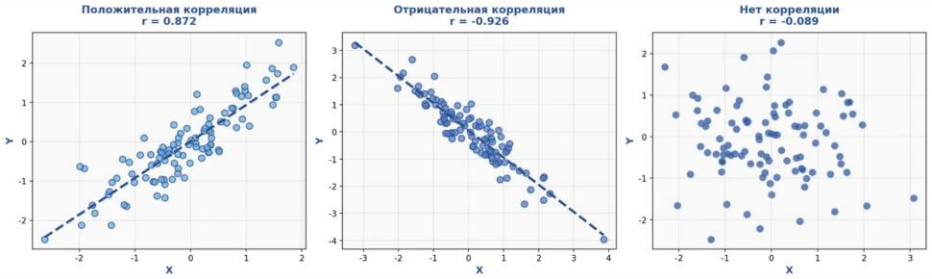
Коэффициент корреляции Пирсона (r) — мера линейной зависимости между двумя переменными. Показывает, насколько сильно и в каком направлении связаны переменные.

Формула корреляции:

$$r = \Sigma[(x_i - \mu_x)(y_i - \mu_y)] \;/\; \sqrt{[\Sigma(x_i - \mu_x)^2 \times \Sigma(y_i - \mu_y)^2]}$$

Диапазон значений: $-1 \le r \le 1$

r = 1	Идеальная положительная корреляция
r = 0	Нет линейной корреляции
r = −1	Идеальная отрицательная корреляция



Интерпретация силы связи:

|r| > 0.7 — сильная связь
0.3 < |r| < 0.7 — умеренная связь
|r| < 0.3 — слабая связь

Корреляция как косинус угла между векторами

Векторное представление: Переменные X и Y можно представить как векторы в n-мерном пространстве, где n — количество наблюдений. Корреляция имеет элегантную геометрическую интерпретацию через угол между этими векторами.

Формула через скалярное произведение:

$$r = \cos(\theta) = (X \cdot Y) / (||X|| \times ||Y||)$$

где θ — угол между центрированными векторами X и Y

Геометрический смысл:

θ ≈ 0° (малый угол)

r ≈ 1 — положительная корреляция

Векторы направлены в одну сторону

θ ≈ 90°

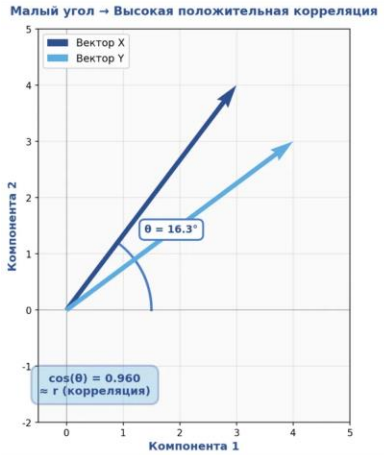
r ≈ 0 — нет корреляции

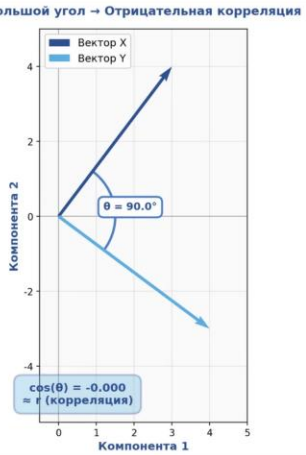
Векторы перпендикулярны

θ ≈ 180° (большой угол)

r ≈ −1 — отрицательная корреляция

Векторы направлены в противоположные стороны





Интуиция

Корреляция показывает, насколько векторы "смотрят в одном направлении". Чем меньше угол между ними, тем сильнее положительная корреляция.

Ключевые выводы

1. **Меры центральной тенденции** (среднее, мода, медиана) показывают типичное значение в данных
2. **Меры разброса** (min, max, IQR, σ) характеризуют вариативность и разброс данных
3. **Квантили** делят распределение на части и помогают понять структуру данных
4. **Правило трех сигм** позволяет выявлять аномалии в нормально распределенных данных
5. **Корреляция** измеряет линейную взаимосвязь между переменными ($-1 \leq r \leq 1$)
6. **Геометрическая интерпретация:** корреляция = косинус угла между векторами

Практическое применение

Эти статистики составляют фундамент анализа данных, машинного обучения и научных исследований. Они позволяют описывать данные, выявлять закономерности, обнаруживать аномалии и принимать обоснованные решения на основе количественных показателей.