Московский политехнический университет

Математические методы анализа данных

Лежнина Юлия Аркадьевна

Москва 2022

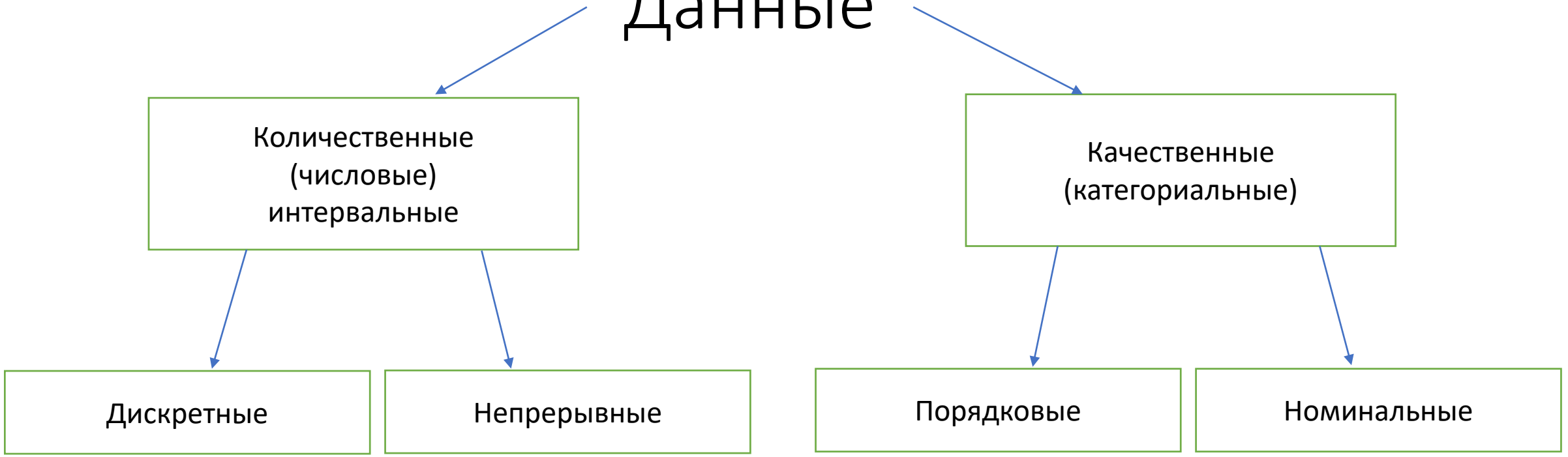
План лекции 2

- Типы данных
- Выборки
- Меры среднего
- Меры вариативности
- Меры и типы переменных
- Формирование выборки
- Ошибки выборки
- Пропущенные данные
- Неопределенные данные

Матрица данных

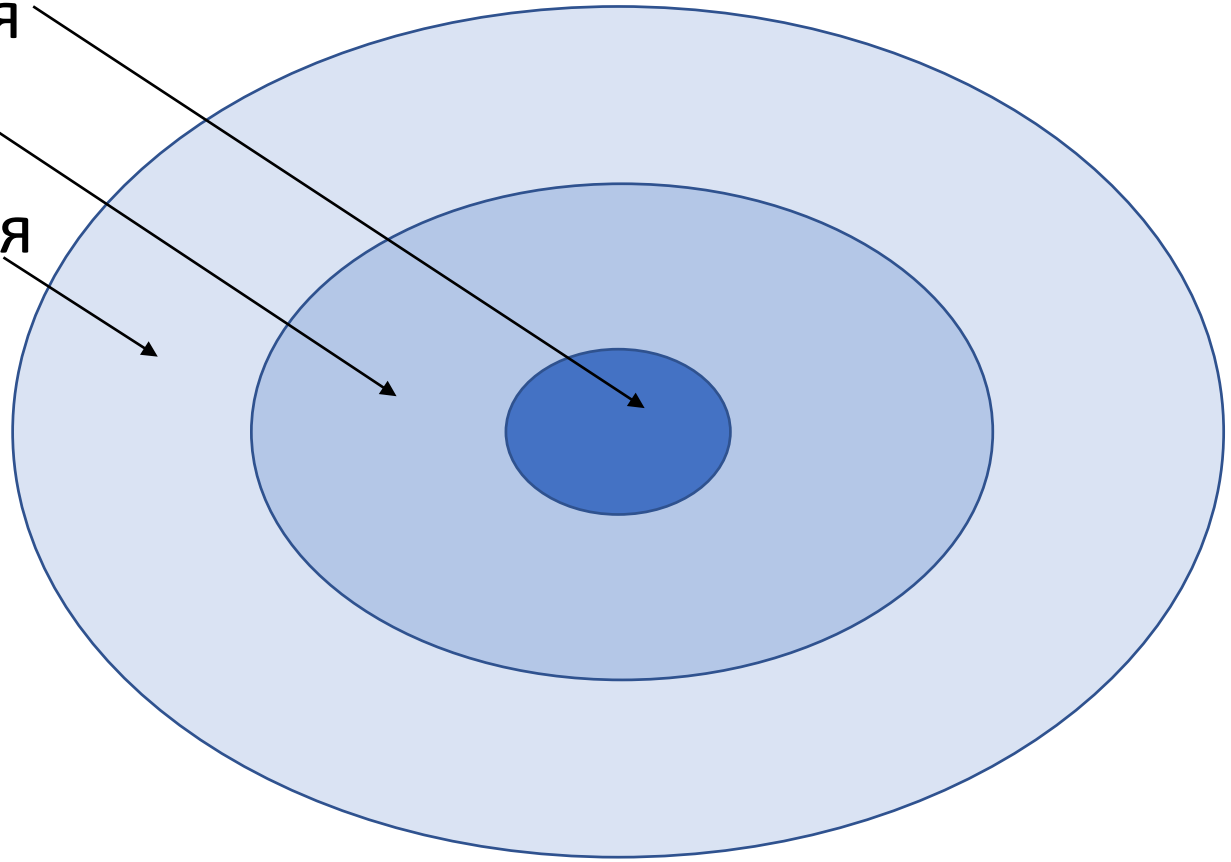
	Признак 1	Признак 2	.....	Признак n
Объект 1	x_{11}	x_{12}	.....	x_{1n}
Объект2	x_{21}	x_{22}	.....	x_{2n}
.....	.....	.....	.....	.....
Объект m	x_{m1}	x_{m2}	.....	x_{mn}

Данные


Количественные
(числовые)
интервальные
Качественные
(категориальные)
Дискретные
Непрерывные
Порядковые
Номинальные

Шкалы типов данных

- Номинальная
- Порядковая
- Интервальная



Пример матрицы данных

№	Населённый пункт	Семейное положение	Пол	Количество подчиненных	Доход	Удовлетворенность жизнью
01	Областной центр	Вдовец (вдова)	Ж	-	13 000	Полностью удовлетворена
02	Областной центр	Живёте вместе, но не зарегистрированы	Ж	-	20 000	И да, и нет
03	Областной центр	Состоите в зарегистрированном браке	Ж	-	17 000	И да, и нет
04	Областной центр	Разведены и в браке не состоите	Ж	-	45 000	Скорее удовлетворена
05	Областной центр	Никогда в браке не состояли	М	-	25 000	Не очень удовлетворён
06	Областной центр	Никогда в браке не состояли	М	-	30 000	Скорее удовлетворён
07	Областной центр	Разведены и в браке не состоите	Ж	30	35 000	Скорее удовлетворена
08	Областной центр	Никогда в браке не состояли	М	-	30 000	Скорее удовлетворён
09	Областной центр	Состоите в зарегистрированном браке	М	3	40 000	Скорее удовлетворён
10	Областной центр	Состоите в зарегистрированном браке	Ж	15	25 000	Скорее удовлетворена

Данные: Российский мониторинг экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE)», проводимый Национальным исследовательским университетом — Высшей школой экономики и ЗАО «Демоскоп»

Основные понятия математической статистики

ВЫБОРКА

- Последовательность независимых случайных величин x_1,x_2,...,x_n, соответствующих всем возможным результатам n статистических экспериментов и имеющих одинаковый закон распределения вероятностей со случайной величиной xi, называется выборкой объёма n, порождённой случайной величиной xi. Если xi — дискретная случайная величина, то выборкой объёма n называется любое подмножество n объектов генеральной совокупности объёма N, выбранное равновероятно среди всех таких подмножеств.

Основные понятия математической статистики

Пусть X_{1}, X_{2},…,X_{n} - конечная выборка из некоторого распределения, определённая на некотором вероятностном пространстве. Перенумеруем последовательность в порядке неубывания, так что

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}.$$

Эта последовательность называется **вариационным рядом**. Вариационный ряд и его члены являются порядковыми статистиками. Случайная величина x(k) называется **k-той порядковой статистикой** исходной выборки.

Выборочная квантиль

$$X_{(1)} < X_{(2)} < X_{(3)} < \ldots < X_{(n)}$$

$$t_\alpha = X_{([\alpha \cdot n])}$$

α — заданная вероятность
n — объём выборки

Случайная величина:

802, 851, 851, 863, 870, 870, 870, 894, 897, 899, 901,
905, 906, 906, 910, 914, 925, 936, 945, 952, 953, 978

α=0,9 => 1-α=0,1 n=22

[(1-α)n] = [0,1 22] = 2

x(2)=851
P(X>851)=90%

Меры центральной тенденции

- Арифметическое среднее $\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$

- Рассмотрим два распределения объема N=20:

X1	**10 000**	**20 000**	**30 000**	**40 000**	**50 000**
n	7	5	3	3	2

$\bar{x}$ =24 000

Мода – наиболее часто встречающееся значение

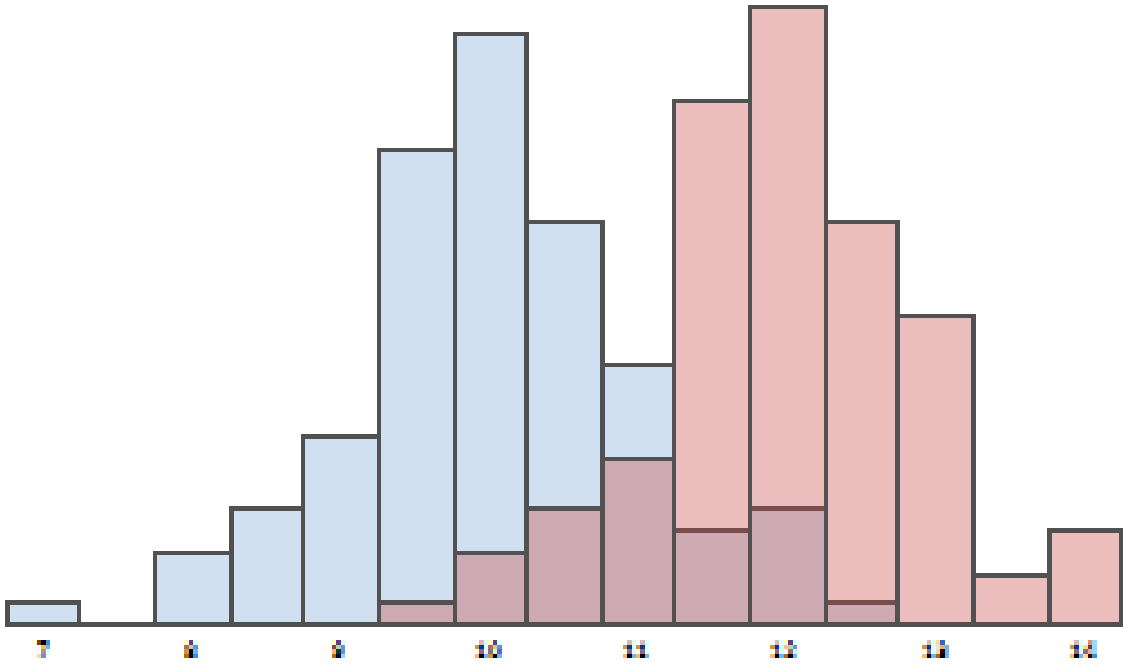
X2	**10 000**	**20 000**	**30 000**	**150 000**
n	10	4	5	1

$\bar{x}$ =24 000

Мода

Бимодальное распределение

Мода – наиболее часто встречающееся значение



Медиана

- Распределения, для которых $\bar{x}$ =24 000
- мода равна 10 000

X1	**10 000**	**20 000**	**30 000**	**40 000**	**50 000**
n	7	5	3	3	2

Медиана =15 000

X2	**10 000**	**20 000**	**30 000**	**150 000**
n	10	4	5	1

Медиана = 10 000

Меры вариативности

- Размах - расстояние между минимальным и максимальным значениями признака

X1	**10 000**	**20 000**	**30 000**	**40 000**	**50 000**
n	7	5	3	3	2

- равен 40 000

X2	**10 000**	**20 000**	**30 000**	**150 000**
n	10	4	5	1

- равен 140 000

Дисперсия и среднеквадратическое отклонение

- **Выборочная дисперсия** в математической статистике — это оценка теоретической дисперсии распределения, рассчитанная на основе данных выборки. Виды выборочных дисперсий:

- Смещённая $$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2.$$

- Несмещённая или исправленная $$S^2 = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \bar{X})^2.$$

- Среднеквадратическое отклонение $$S_0 = \sqrt{\frac{n}{n-1}S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

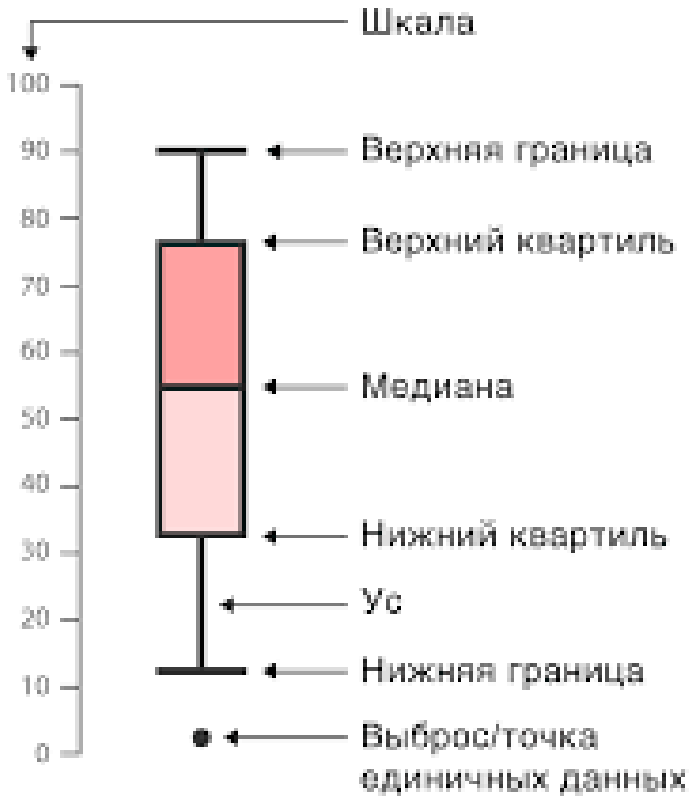
Меры менее чувствительные к выбросам

- **Среднее абсолютное отклонение**, или просто среднее отклонение (англ. MAD, mean absolute deviation) — величина, используемая для оценки прогнозных функций

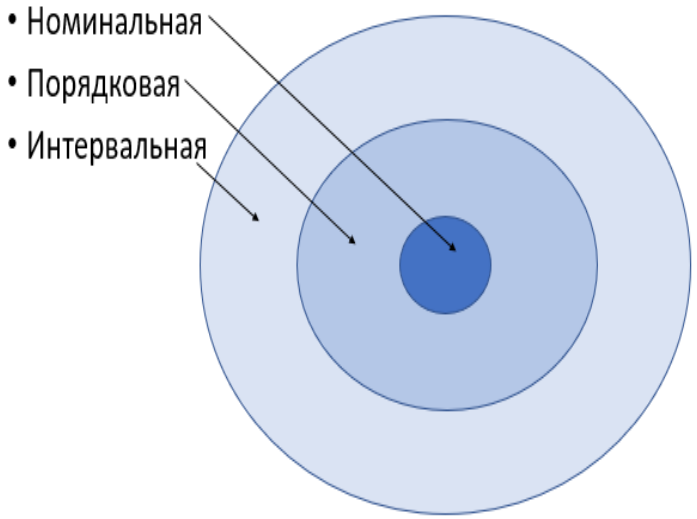
$$MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - m(X)|$$

Межквартильный размах

- **Межквартильный размах** — это разница между 1-м и 3-м квартилями, т.е. между 25-м и 75-м процентилями. В него входят центральные 50% наблюдений в упорядоченном наборе, где 25% наблюдений находятся ниже центральной точки и 25% — выше.



Виды данных по объёму содержащейся в них информации



Типы шкал	**Меры центра**			**Меры вариативности**		
	Мода	Медиана	Среднее	Размах	MAD	СКО
Номинальные						
Порядковые						
Интервальные						
Интервальные с особенностями						

Формирование выборок

Выборка должна быть

- Репрезентативной
- Репрезентативны только случайные выборки

Случайные выборки

- Простая случайная выборка
- Механическая выборка
- Стратифицированная
- Гнездовая или кластерная

Неслучайные выборки

Могут использоваться:

- Метод снежного кома
- Квотная выборка

Необходимо избегать

- Доступная

Ошибки выборки

Ошибка выборки — отклонение средних характеристик выборочной совокупности от средних характеристик генеральной совокупности.

Предельная ошибка выборки

Предельная ошибка — максимально возможное расхождение средних значений выборки и генеральной совокупности с заданной вероятностью.

$$P\left\{\left|\bar{X}-\mu\right|<\Delta_{\bar{X}}\right\}=p$$

$$\Delta_{\bar{x}}=t\sqrt{\frac{s^2\left(1-\frac{n}{N}\right)}{n}}$$

s^2 - дисперсия признака в выборочной совокупности

n- число единиц в выборке

N- объем генеральной совокупности

t- коэффициент доверия Стьюдента

Доверительный интервал — интервал, в который попадает неизвестный параметр с заданной вероятностью.

$$\bar{X}-\Delta_{\bar{x}}<\mu<\bar{X}+\Delta_{\bar{x}}$$

Объем выборки можно получить из формулы предельной ошибки

$$n=\frac{t^2s^2N}{\Delta^2N+t^2s^2}$$

Пропущенные наблюдения

- Выявить причину в случае наличия систематических ошибок
- Исключить пропущенные наблюдения
- Заменить функцией от соседей
- Заменить похожими значениями

Неопределенные данные

- Ответы на вопрос «не знаю»
- Для порядковых шкал – заменить самым старшим рангом.

!Порядковая шкала превратится в номинальную!