Введение в компьютерный и интеллектуальный анализ данных

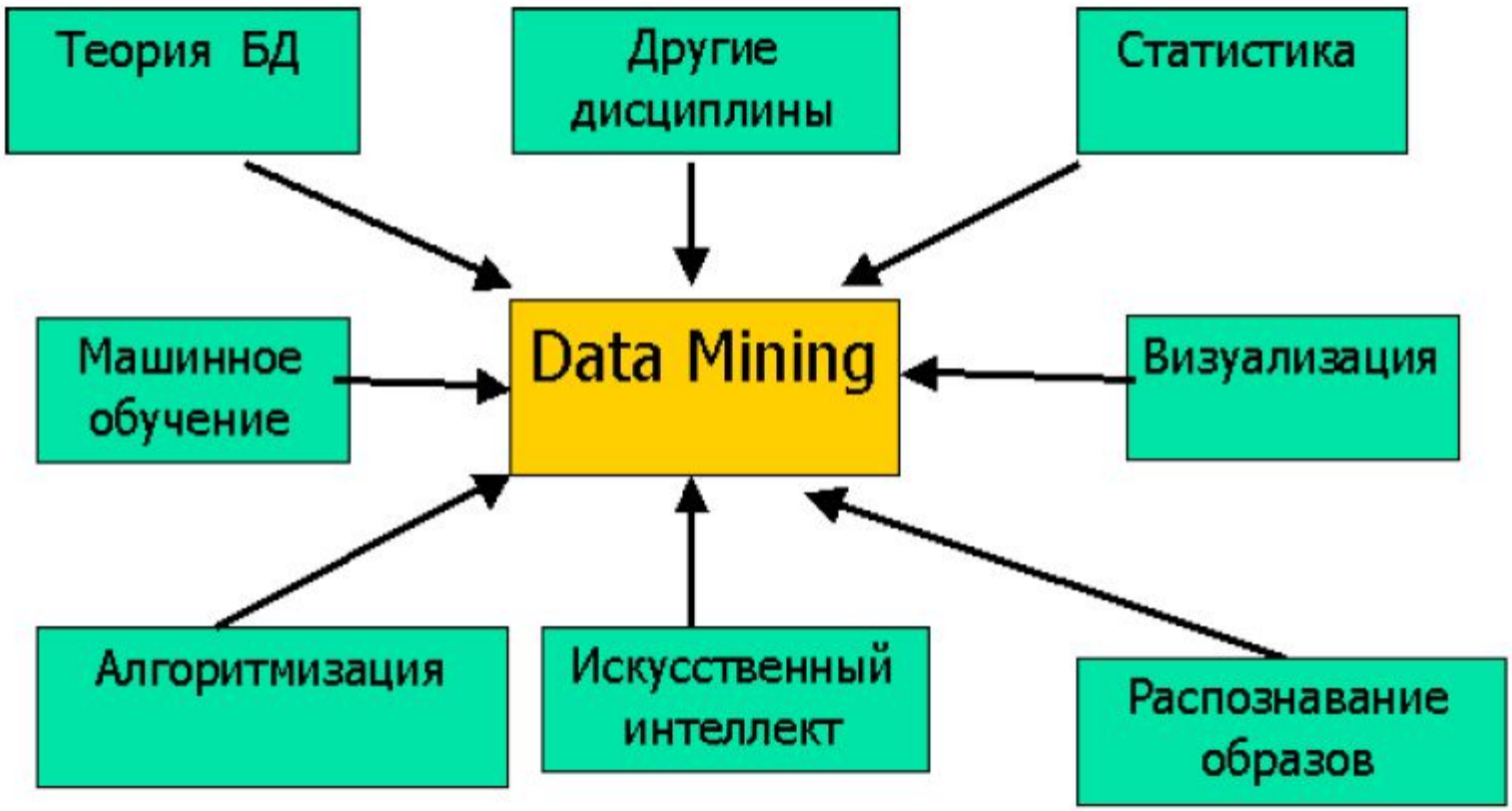
(ВКИАД)

(Data Analysis,

Data Analytics,

Data Mining)

Мультидисциплинарная область



Цели курса

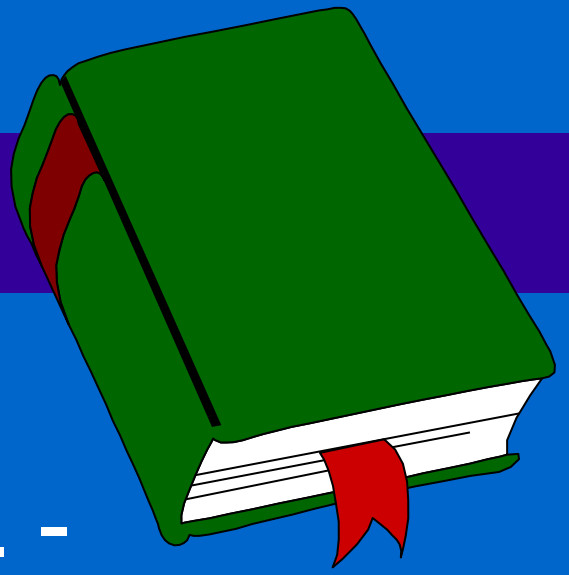
- изучение теоретических основ предварительного (домодельного) статистического анализа данных
- формирование навыков практического решения задач анализа данных

(ВКИАД)

Тема 1.

Типы статистических данных и способы их первичной обработки

Литература



- **Статистика**: учебник / Под ред. И.И.Елисеевой. - М: Изд-во Проспект, 2019.

- **Локальная сеть БГУ:**

  FPMI-STUD\subfaculty\КТС\Казаченок\ВКИАД

Развитие статистики

- Др.Китай, др.Рим, Ср.век.Европа
- Описательная статистика
  Г.Конринг (сер. XVIIв., Германия)
- Политическая арифметика
  В.Петти (сер. XVIIв., Англия)
- Математическая статистика
  Кетле, Гальтон, Пирсон, Госсет, Фишер, Митчел (XIX-XXв.)

Термин «статистика»

- STATUS (лат.) – состояние дел
- «Статистика» – (Готфрид Ахенваль, XVIII век)

Современное значение:

- Отрасль деятельности
- Научная дисциплина
- Цифровой материал

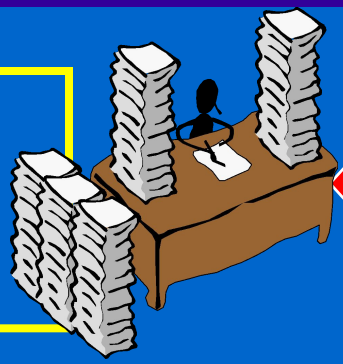
Статистика как…

- Отрасль деятельности
  - Государственная статистика
  - Ведомственная статистика
  - Муниципальная статистика, …
- Научная дисциплина
  - Описательная статистика
  - Экономическая статистика
  - Математическая статистика, …

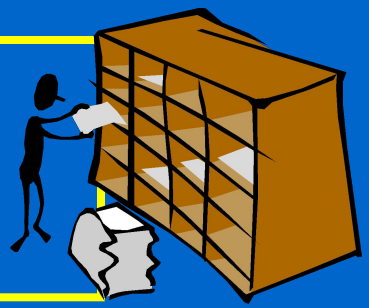
Статистическое исследование

Объекты статистического наблюдения

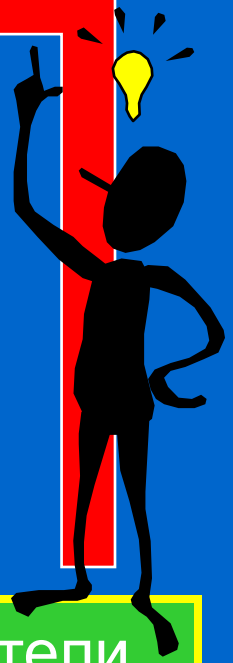
1 Сбор первичной информации



2 Сводка и обработка данных



3 Анализ и интерпретация результатов



Потребители статистических данных

Категории статистики

1. Статистическая совокупность
2. Единица совокупности
3. Признак
4. Статистический показатель
5. Система статистических показателей

Методы статистики

- Статистическое наблюдение
- Метод группировок
- Метод статистических показателей

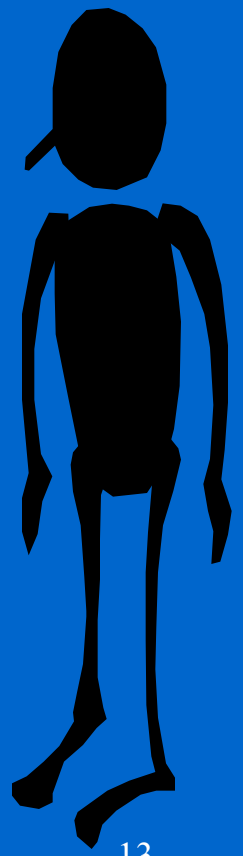
Статистическая совокупность

- совокупность изучаемых социально-экономических объектов или явлений, имеющих общую качественную основу, но отличающихся друг от друга отдельными признаками.

Единица совокупности

- первичный элемент статистической совокупности, являющийся носителем признаков, подлежащих регистрации.

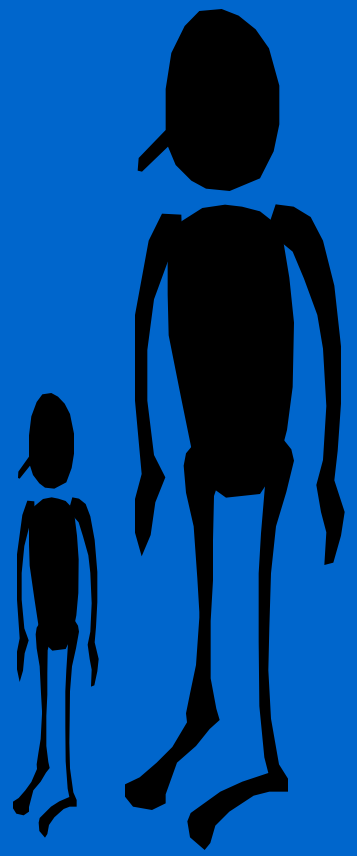
  – Единица совокупности рассматривается как неделимый элемент



Признак

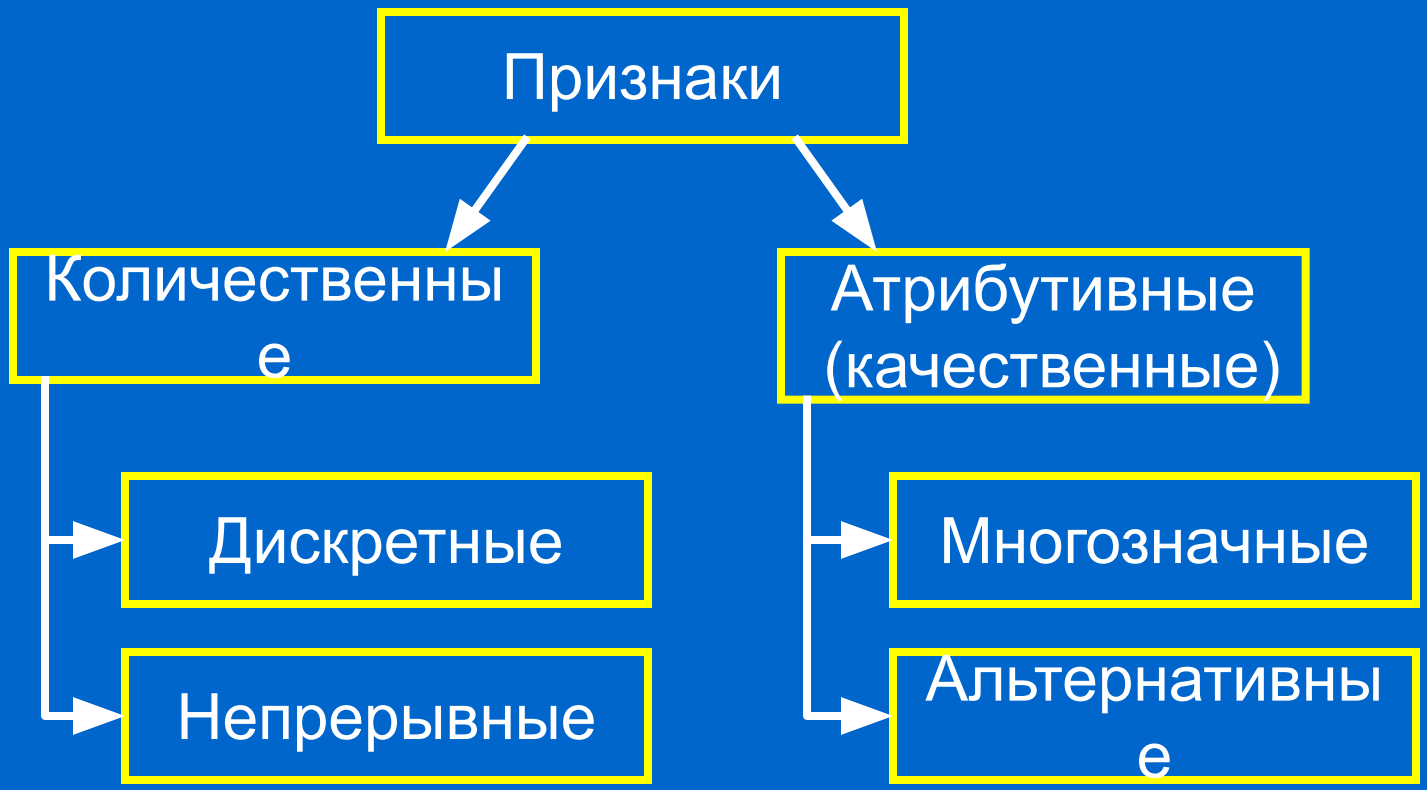
- показатель, характеризующий индивидуальную особенность единицы совокупности, рассматриваемый как случайная величина

  – Значение признака - измеренный индивидуальный показатель



Классификация признаков

По типу значений (измерений)



Типовые измерительные шкалы

Тип шкалы

Качественные (атрибутивные)

- Шкала наименований
- Порядковая шкала

Количественные

- Интервальная шкала
- Шкала отношений

Шкала наименований

= номинальная = классификационная

Примеры:

- имя, пол, семейство, класс, номер игрока …

Обработка таблиц наблюдений:

- Неупорядоченный список класса эквивалентных объектов

2 Порядковая шкала

= ранговая = ординальная

Примеры:

- ранг служащего, балльные шкалы (сила ветра, оценка на экзамене, магнитуда землетрясения, твердость минерала) …

Обработка таблиц наблюдений:

- Упорядочение объектов
- Ранг (порядковый номер) объекта

3 Интервальная шкала

= шкала разностей

Примеры:

- температура °C, °F, летоисчисление, высота над уровнем моря …

Обработка таблиц наблюдений:

- Взятие интервалов – разностей

Шкала отношений

= метрическая

Примеры:

- длина, высота, вес, скорость, светимость …

Обработка таблиц наблюдений:

- Арифметические операции

Статистический показатель

- количественно-качественная обобщающая характеристика какого-либо свойства группы (части) единиц совокупности или совокупности в целом

- Стат.данные – совокупность значений стат.показателей

Типы показателей

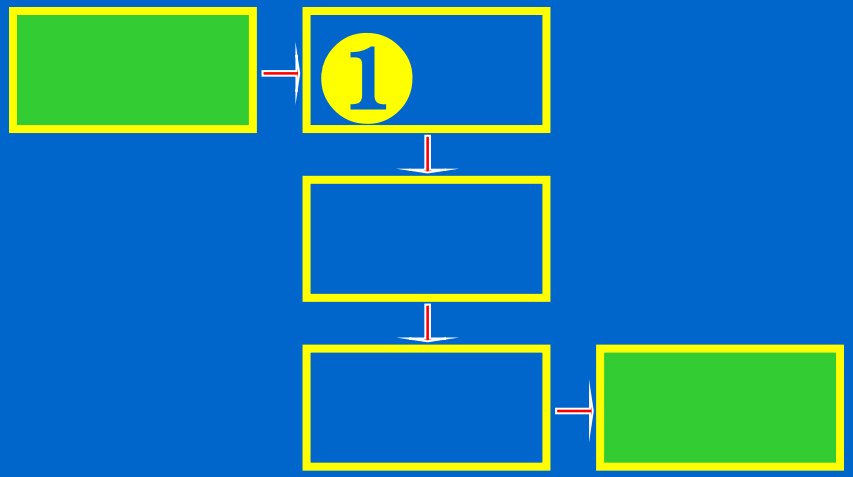
- Первичные (объемные)
- Вторичные (производные)

- Индивидуальные (единичные)
- Сводные (групповые, суммарные)

Система статистических показателей

- совокупность взаимосвязанных показателей, отражающая существующие между явлениями взаимосвязи

- Сист. стат. показателей фиксирует:
  - Множество показателей
  - Классификацию единиц

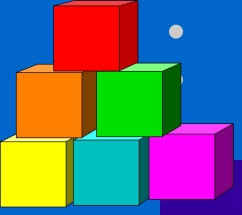


Статистическое наблюдение

- Определение
- Формы и виды
- Программа
- Точность наблюдения

Статистическое наблюдение

- планомерный, научно организованный сбор информации о массовых общественных явлениях путем регистрации заранее намеченных признаков с целью получения обобщающих характеристик



Виды стат. наблюдения

По охвату единиц совокупности:

- Сплошное: все единицы
- Несплошное: часть единиц
  - Метод основного массива: наиболее «крупные» единицы
  - Выборочное: механический или случайный отбор единиц

Выборочный метод

- Генеральная совокупность (исследуемая стат. совокупность)
- Выборочная совокупность (отобранные единицы, «выборка»)
  - Представительность выборки (репрезентативность) - близость свойств генеральной и выборочной совокупностей

Формирование выборки

1. Выясняется состав совокупности (*N*)
2. Определяется объем выборки (*n*)
3. Осуществляется отбор:
   - Индивидуальный
     - Механический
     - Случайный
     - и т.д.

Механический отбор

- отбор каждой (*N/n*)-ой единицы
  $k_i = k_1 + [\ (i-1)\ N/n\ ]$ $i=1..n$

Статистическое наблюдение

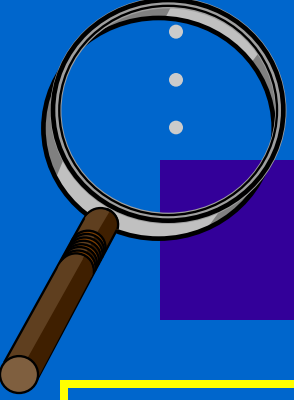
Точность наблюдения

Ошибки (погрешности)

- различия между показателями выборочной и генеральной совокупностей

Измеряется с помощью

- Абсолютная ошибка (разность)
- Относительная ошибка (отношение, %)



Ошибки выборки

Оценка	Число студентов		
	Ген.совок	Выборка 1	Выборка 2
2 3 4 5	100 300 520 80	9 27 54 10	12 29 52 7
Итого	1000	100	100
Среднее	3,58	3,65	3,54
Доля «4 и 5»	0,6	0,64	0,59

Ряды динамики

Ряды динамики – статистические данные, отображающие развитие во времени изучаемого явления.

Их также называют **динамическими рядами**, **временными рядами**.

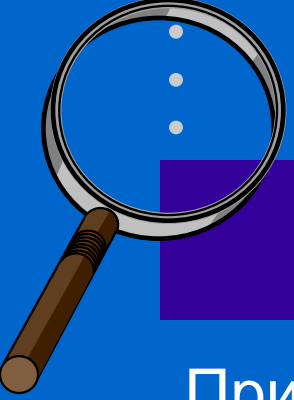
Пример. Производство изделий «А» в 2009-2015гг.

Год	2009	2010	2011	2012	2013	2014	2015
Производство	30,1	34,9	44,3	27,0	31,0	34,5	47,0

Вариационный ряд

Если ряд распределения построен по *количественному* признаку, то такой ряд называют **вариационным.**

Построить вариационный ряд - значит *упорядочит*ь количественное распределение единиц совокупности по значениям признака, а затем подсчитать числа единиц совокупности с этими значениями (построить групповую таблицу).



Пример вариационных рядов

Пример 1.

В магазине продана мужская обувь следующих размеров:
38, 41, 41, 38, 43, 39, 39, 42, 42, 39, 42, 39, 40, 40, 40, 39, 39.

Дискретный вариационный ряд:

Размер обуви	38	39	40	41	42	43
Кол-во пар	2	6	3	2	3	1

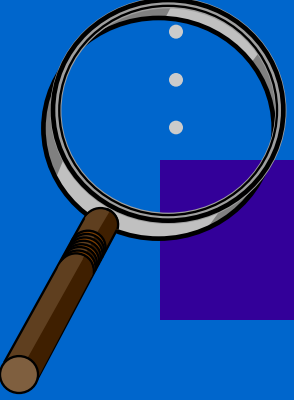
Интервальный вариационный ряд:

Размеры обуви	38-39	40-41	42-43
Кол-во пар	8	5	43

Атрибутивный ряд

Если за основу группировки взят *качественный* признак, то такой ряд распределения называют **атрибутивным**

(распределение по видам труда, по полу, по профессии, по религиозному признаку, национальной принадлежности и т.д.).



Пример атрибутивного ряда

Образование рабочих	Количество рабочих	
	абсолютное	в %
Высшее	20	15,4
Неполное высшее	25	19,2
Среднее специальное	35	26,9
Среднее	50	38,5
ИТОГО	**130**	**100**

Статистическая группировка

Формально-математический способ предполагает использование формулы **Стерджесса:**

$$k = 1 + [\log_2 n]$$

где k — *число групп;*

n — *число единиц* совокупности.

Применение группировки (шаг 1)

Пример 2.

Построить интервальный вариационный ряд распределения по первичным данным о размере прибыли 20 коммерческих банков за год (млрд. руб.)

3.7 4.3 6.7 5.6 5.1 8.1 4.6 5.7 6.4 5.9 5.2 6.2 6.3 7.2 7.9
5.8 4.9 7.6 7.0 6.9

РЕШЕНИЕ (6 шагов)

1. Упорядочиваем ряд:

3.7 3.7 4.6 4.9 5.1 5.2 5.6 5.7 5.8 5.9 6.2 6.3 6.4 6.7 6.9
7.0 7.2 7.6 7.9 8.1

Применение группировки (шаги 2-4)

2. Вычисляем размах:

$$R = X_{\max} - X_{\min} = 8.1 - 3.7 = 4.4$$

3. Вычисляем количество групп:

$$k = 1 + [\log_2 20] = 5$$

4. Вычисляем величину интервала:

$$H = R / k = 4.4 / 5 = 0.88 \sim 0.9$$

Применение группировки (шаги 5-6)

5. Вычисляем границы интервалов:

[3.7;4.6), [4.6;5.5), [5.5;6.4), [6.4;7.3), [7.3;8.2]

6. Подсчитаем количество вариант, попавших в каждый интервал, и запишем в таблицу:

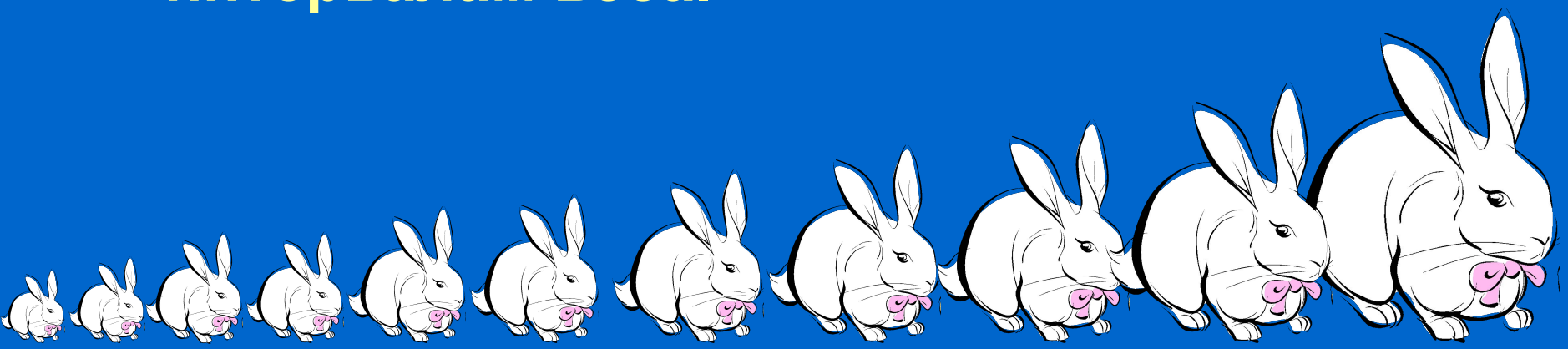
x_i *(размер прибыли)*	[3.7;4.6)	[4.6;5.5)	[5.5;6.4)	[6.4;7.3)	[7.3;8.2]
m_i *(кол-во банков)*	2	4	6	5	3

Непараметрическое описание распределений

Пример: Взвешиваем *N* кроликов

Пример: Упорядочение кроликов

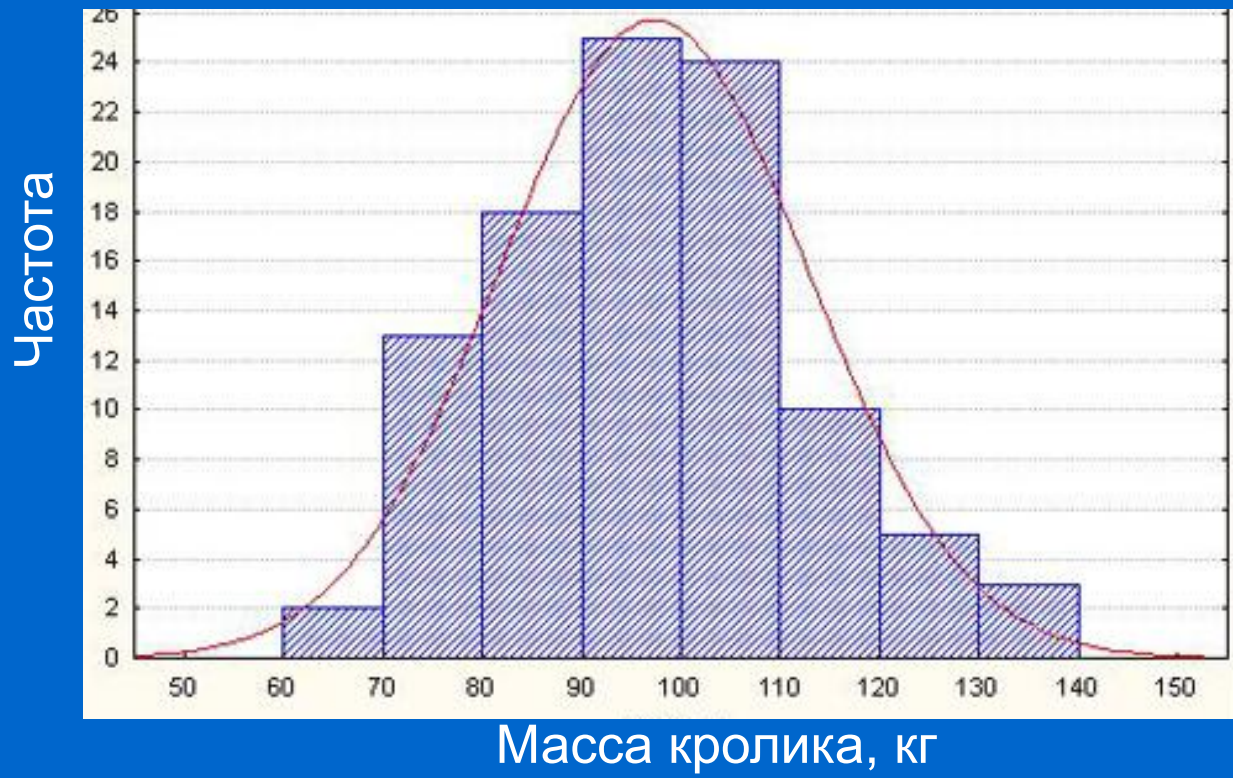
1. Упорядочим кроликов по возрастанию веса (значения переменной);
2. Разобьём их на группы по равным интервалам веса.



Частотное распределение переменной
(Плотность распределения вероятностей ?)

Частота – то, сколько раз встретилось данное значение переменной

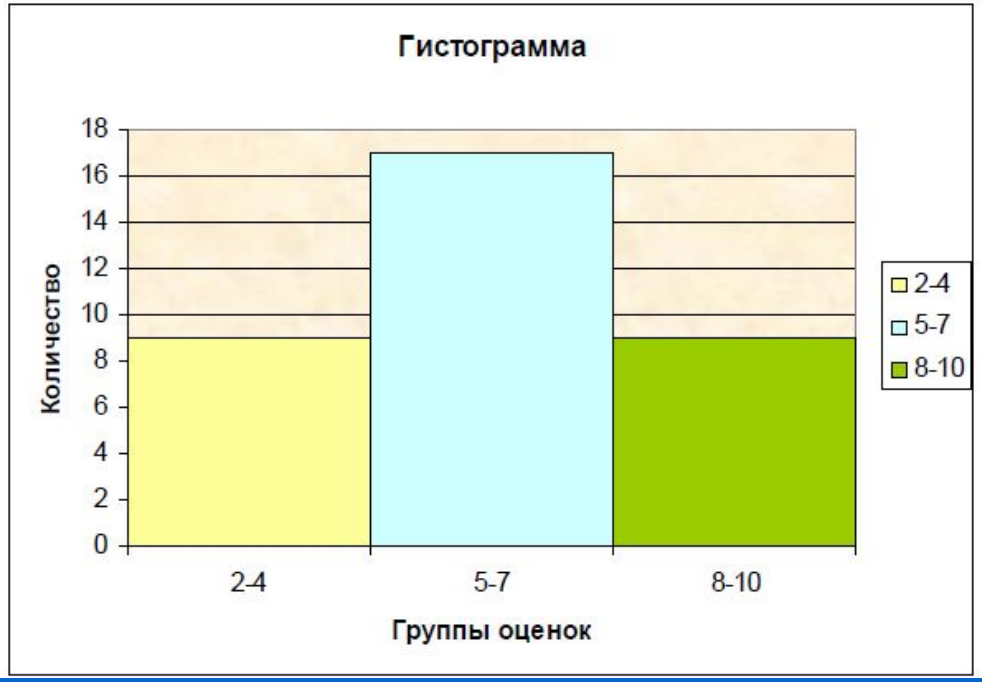
Гистограмма – графическое представление частотного распределения, разбитого по интервалам, где высота столбика отражает **ЧАСТОТУ**



Интервалы должны быть одного размера.

Другой пример гистограммы

$[x_i, x_{i+1})$	2-4	5-7	8-10
m_i	9	17	9



Для интервальных вариационных рядов

Описание частотного распределения

Три **ОСНОВНЫЕ ХАРАКТЕРИСТИКИ**:

1. «**Середина**» распределения;
2. «**Ширина**» распределения;
3. **Форма** распределения

Это относится
не только к количественным данным,
но и к качественным

Варианты «Середины» распределения

Все значения могут служить оценками.

Среднее значение в выборке – наиболее эффективная оценка.

Медиана (квартиль?)

Медиана – значение, которое делит распределение пополам (его площадь в т.ч.): половина значений больше медианы, половина – не больше.

Имеет смысл не только для *количественных* переменных, но и для *ранговых*! (*не для качественных*).

Медиана 1

Если дискретный ряд содержит *нечетное* количество вариант, то находится та единственная варианта, справа и слева от которой находится одинаковое число вариант:

$$Me = x_{\frac{n+1}{2}}$$

Пример: 1 2 3 3 4 5 5 6 7 7 8 , n=11

$$Me = x_{\frac{11+1}{2}} = x_6 = 5$$

Медиана 2

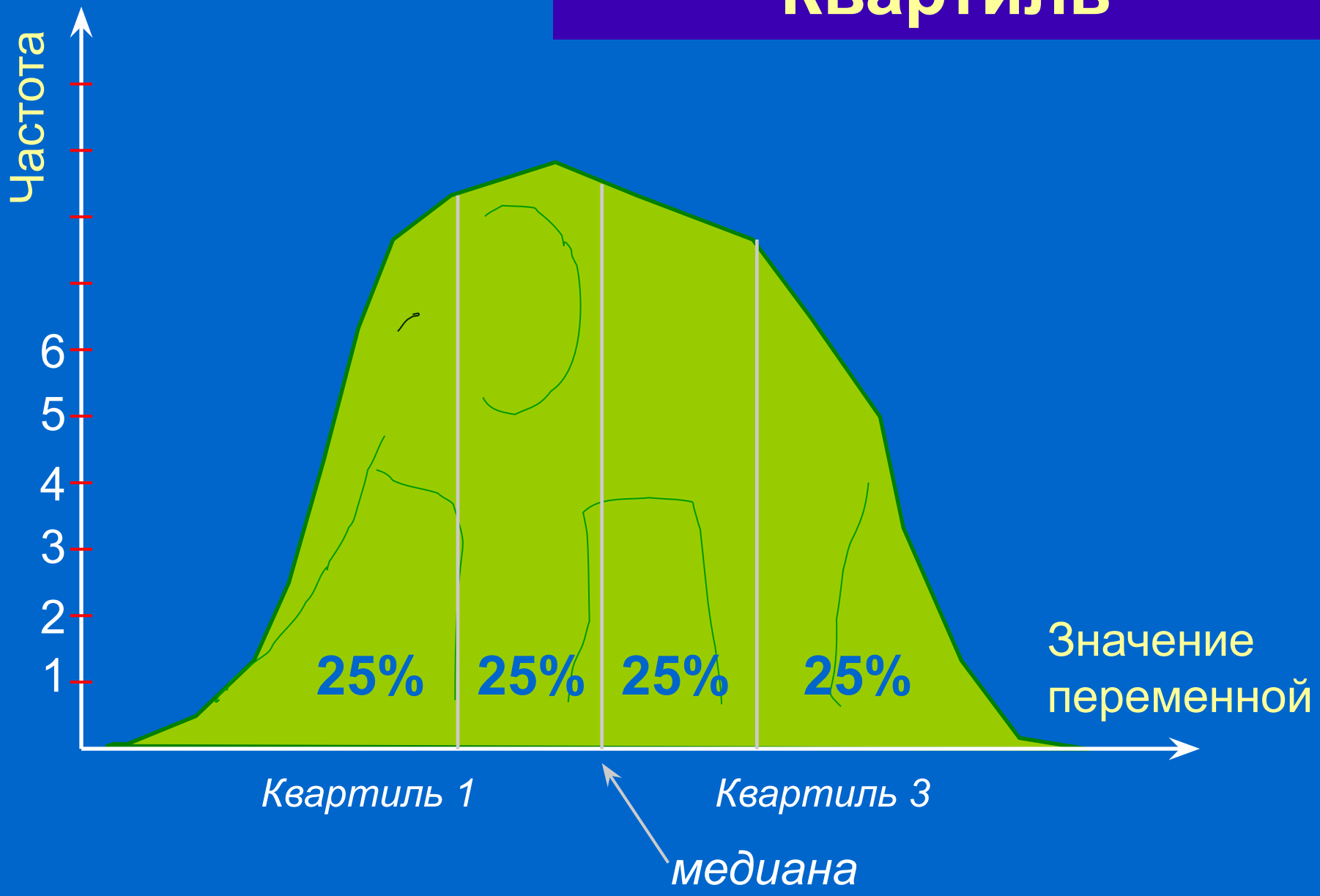
Если дискретный ряд содержит четное количество вариант, то находятся две варианты, справа и слева от которых располагается одинаковое количество вариант. *Me* равна средней арифметической из двух значений:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$$

Пример: 1 2 3 3 4 5 5 6 7 7 8 9, n=12

$$Me = \frac{x_{\frac{12}{2}} + x_{\frac{14}{2}}}{2} = \frac{x_6 + x_7}{2} = 5$$

Квартиль


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

Интерквартильный размах

Квартили (quartiles) делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана).

1-я квартиль = 25% процентиль

3-я квартиль = 75% процентиль

Интерквартильный размах – разность между третьей и первой квартилями.

Деление распределения на части

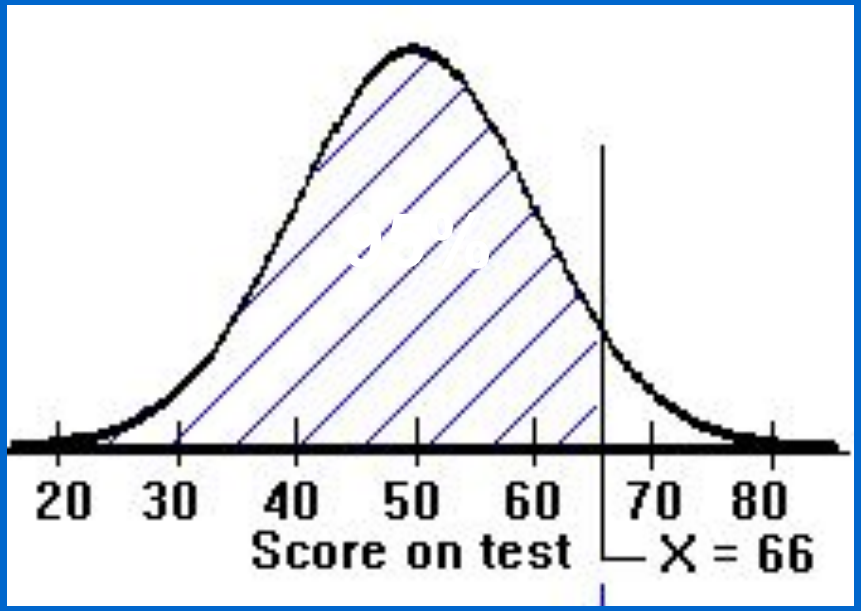
Распределение можно поделить не только на ДВЕ равные части, но и на:

- ✔ ЧЕТЫРЕ (значения, стоящие на границах - квартили);
- ✔ ВОСЕМЬ (... октили);
- ✔ СТО (... процентили);
- ✔ N (квантили порядка 1/N).

Процентили, пример

95% процентиль – значение переменной,

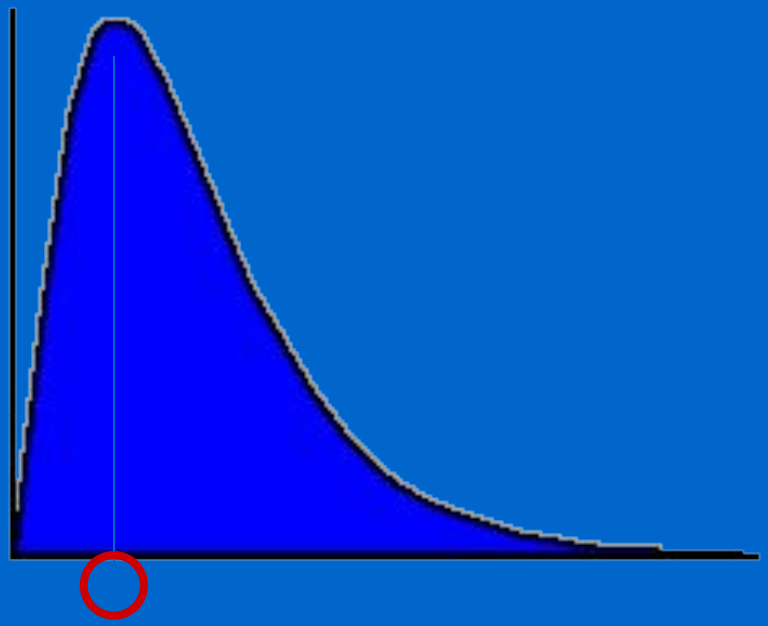
левее которого находится 95% значений переменной



Мода

Мода – наиболее часто встречающееся значение

Существует не только для *количественных*, но и для *ранговых*, и для *качественных* переменных



Мода может быть не единственной

Мода

Мода — это варианта, которая имеет наибольшую частоту. Она соответствует определенному значению признака.

Соглашения о существовании моды:

Если все варианты наблюдаются с одинаковой частотой, то говорят, что вариационный ряд *не имеет моды*.

Если две или более *соседние варианты* имеют наибольшие частоты, равные между собой, то мода равна *средней арифметической* этих вариант.

Если равные варианты, имеющие наибольшие частоты, расположены *не по соседству*, то принято говорить, что признак имеет *две и более моды* (бимодальный, полимодальный признаки и т.д.)

Пример полигона частот

Оценки	Количество
	0
3	5
4	6
5	7
6	9
7	5
8	4
9	2
10	1
	0

Для дискретных вариационных рядов

Пример данных для кумуляты

Оценки	Количество	**Накопл. абс. частоты**	Накопл. отн. частоты	в %
3	5	5	0,13	13%
4	6	11	0,28	28%
5	7	18	0,46	46%
6	9	27	0,69	69%
7	5	32	0,82	82%
8	4	36	0,92	92%
9	2	38	0,97	97%
10	1	39	1,00	100%

Пример кумуляты

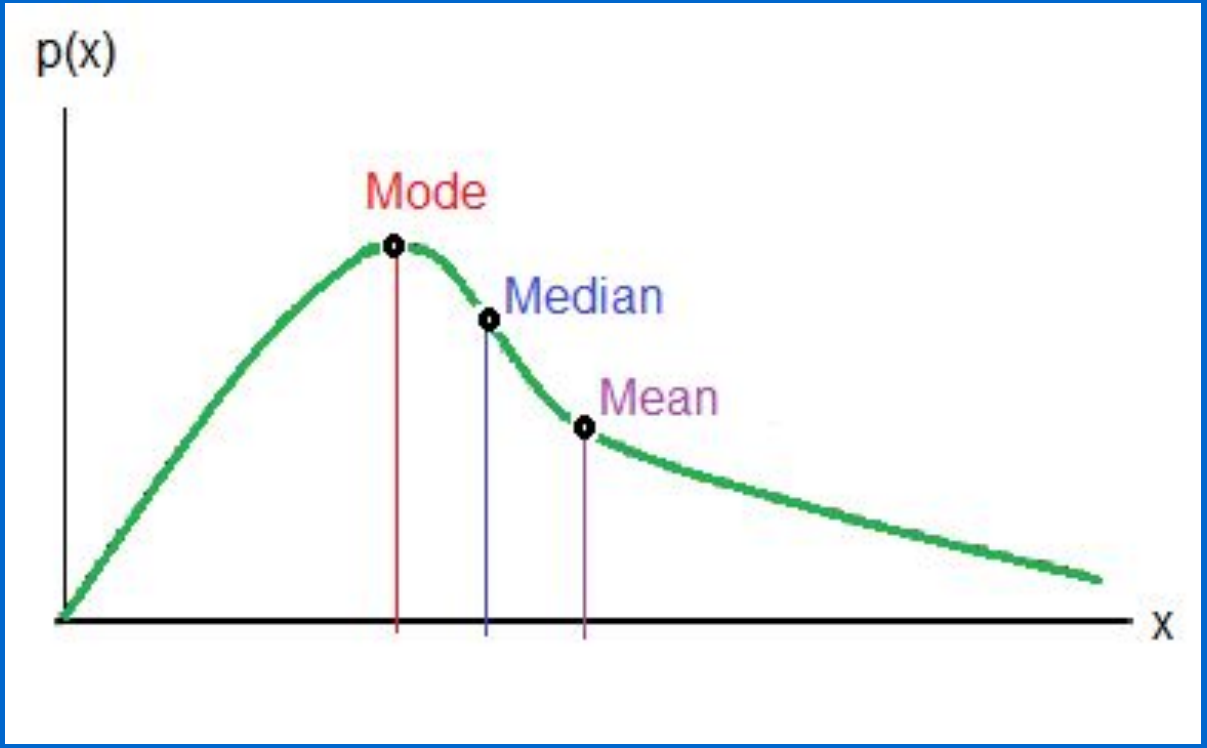
(Функция распределения вероятностей ?)

Для дискретных и интервальных вариационных рядов

Пример: «Середина» распределения

Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения

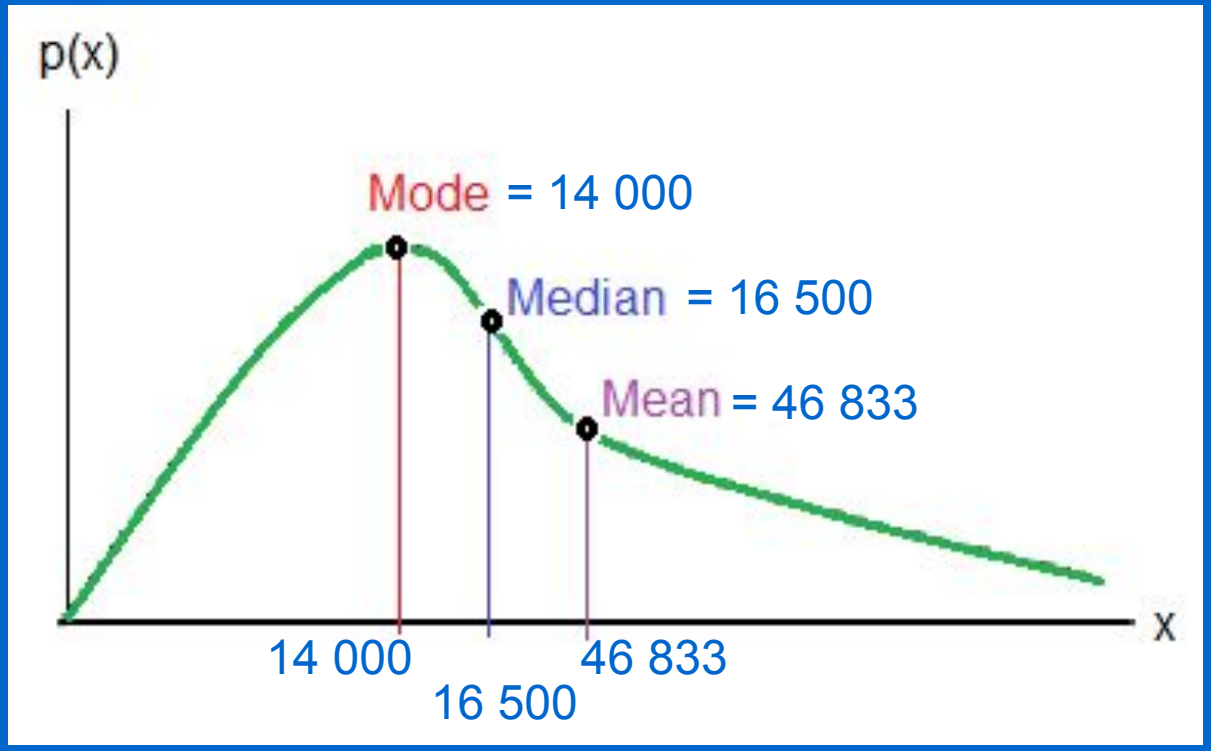
ЗАРПЛАТА, $	ЧАСТОТА
200 000	1
20 000	1
19 000	1
14 000	3



Пример: «Середина» распределения

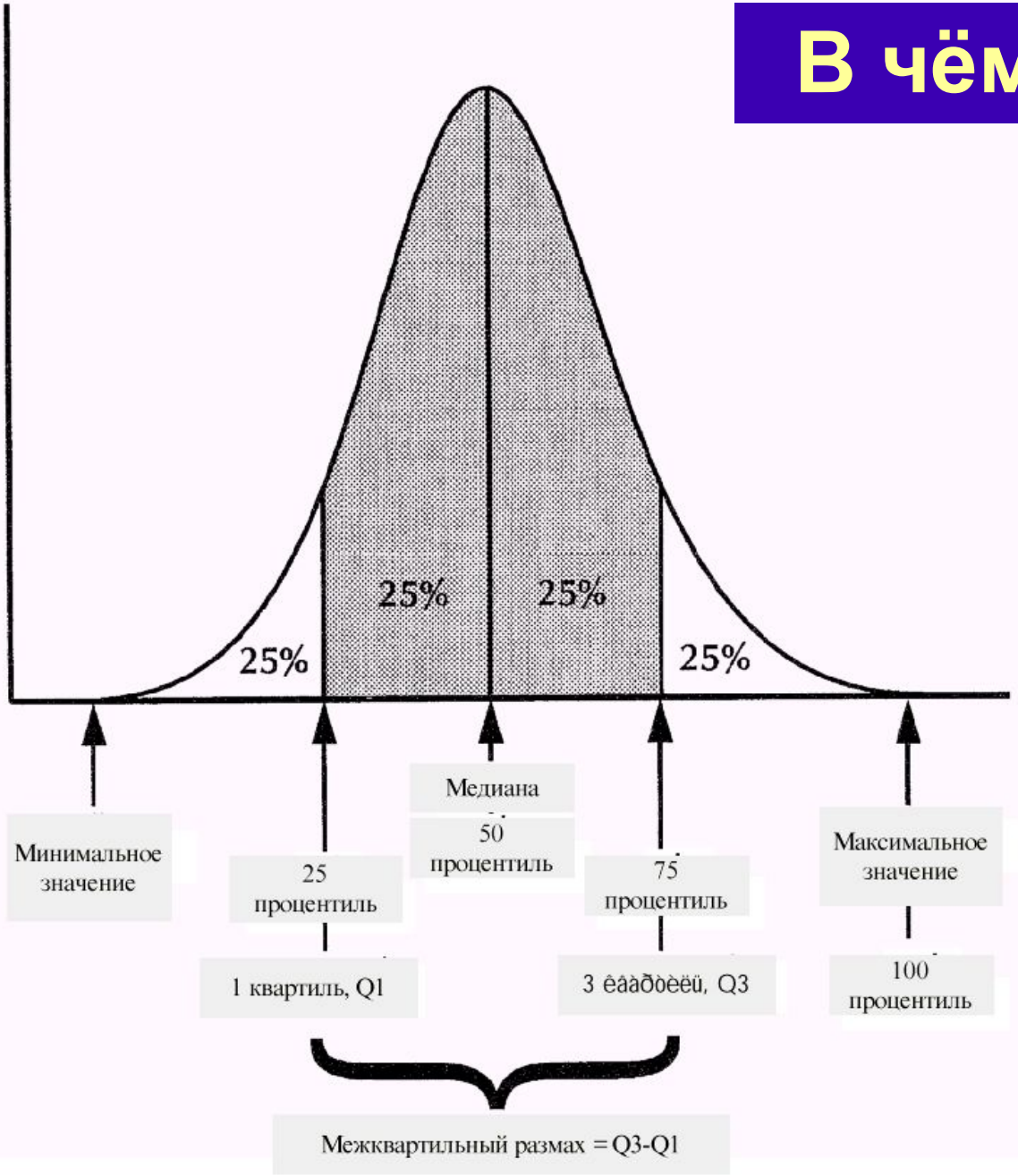
Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения

ЗАРПЛАТА, $	ЧАСТОТА
14 000	1
14 000	1
14 000	1
19 000	1
20 000	1
200 000	1



К появлению перекоса чувствительнее всего среднее значение

В чём ошибка?


25%
25%
25%
25%
Минимальное значение
25 процентиль
1 квартиль, Q1
Медиана
50 процентиль
75 процентиль
3 êâàðòèëü, Q3
Максимальное значение
100 процентиль
Межквартильный размах = Q3-Q1