ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

Статистика, часть 2

Николай Вячеславович Павлов *pavlov@kafedrapik.ru*

Условные обозначения

Это самое важное, надо знать на 100%!!!

2*2=4

Быть или не быть?

Это надо решить и записать!!!

АНАЛИЗ ЭМПИРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Анализ эмпирических распределений = детальное исследование одномерных массивов данных.

Комплексный анализ **рядов распределения** включает:

1. Табличное и графическое представление ряда распределения.
2. Расчет и анализ показателей центра и структуры распределения.
3. Расчет и анализ показателей вариации.
4. Характеристику формы распределения.
5. Выравнивание эмпирического распределения и оценку его соответствия тому или иному типу теоретических распределений.

Определение и виды

Ряды распределения = упорядоченное по значению признака распределение единиц совокупности

- Атрибутивный (по качественному признаку)
- Вариационный (по количественному признаку)
  - Дискретный
  - Интервальный

Примеры = ?

Элементы ряда

Варианта

Частота

№	Категория занятости работника	Количество респондентов
1	Полная занятость	747
2	Частичная занятость	161
3	Безработный	32
4	Временно не работает	51
5	Пенсионер(ка)	231
6	Учащийся	42
7	Домохозяйка(ин)	200
8	Нет ответа	36
Всего		1500

Ранжирование = упорядочение (**Оно есть?**)

СДДН, руб. в месяц, 2013 г.	Count (частота)	Cumulative(накопленная частота) Count	Percent (частость)	Cumulative (накопленная частость) Percent
13472,00<=x<18304,00	21	21	27,27	27,27
18304,00<=x<23136,00	33	54	42,86	70,13
23136,00<=x<27968,00	12	66	15,58	85,71
27968,00<=x<32800,00	6	72	7,79	93,51
32800,00<=x<37632,00	3	75	3,90	97,40
37632,00<=x<42464,00	2	77	2,60	100,00

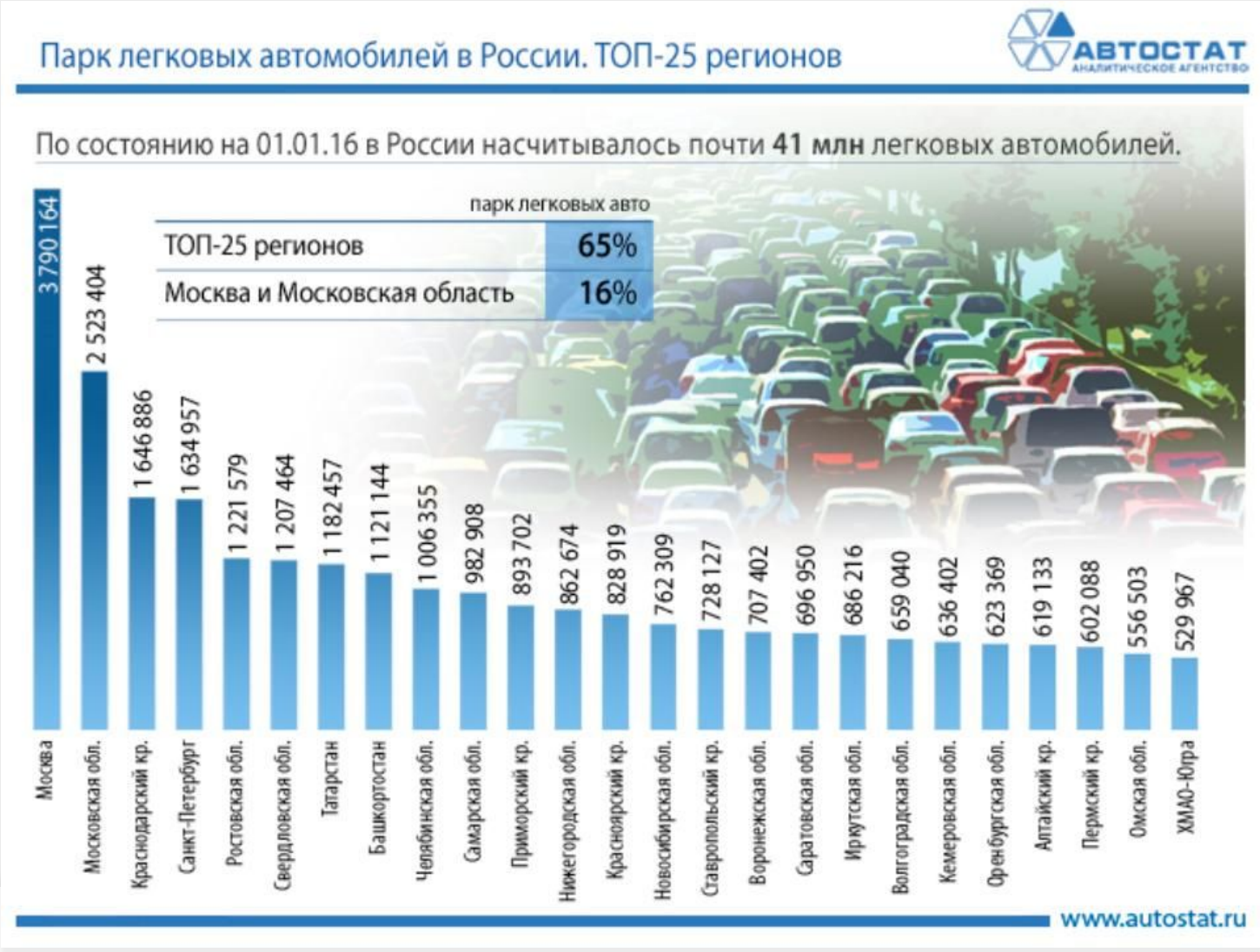
ПОЛИТЕХ

Ракеты КНДР

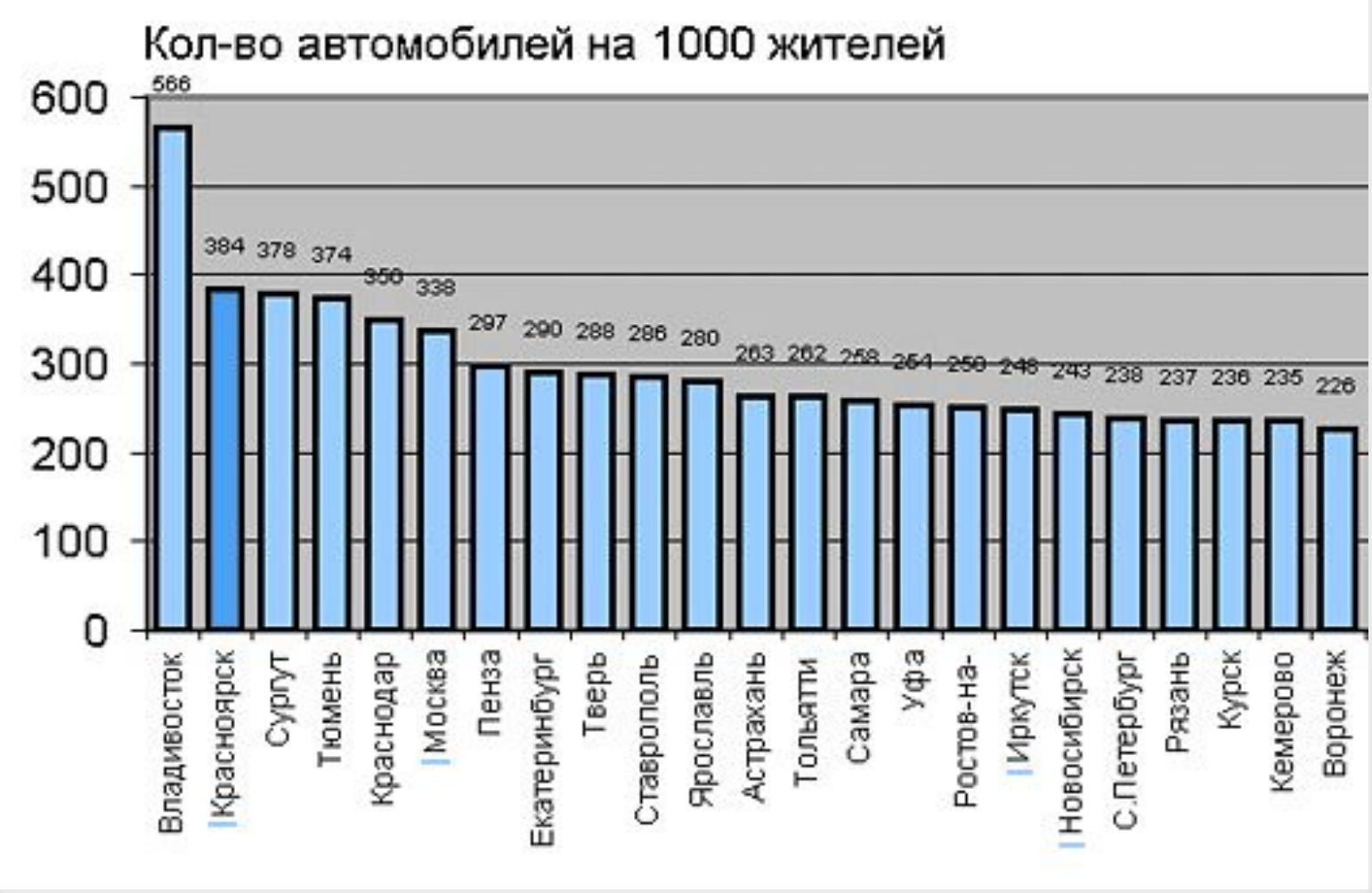
North Korea's missile arsenal

	KN-02	Scud B/C/ER	KN-11	KN-15	Nodong	Musudan	Hwasong-12	Hwasong-14
Range	75 mi	185-620	620	620	810	2,485	2800	4,970-6,215
	Could be used to hit targets in South Korea with conventional or chemical warheads	Could be used to launch a nuclear strike on U.S. military sites in Japan	If perfected, KN-11 would mark a major breakthrough in stealth, since it can be launched from the sea KN-15 is a land-based version		Has potential for nuclear strike on parts of Japan	Capable of reaching U.S. military bases in Japan and all the way out to the Pacific island of Guam	Engine thought to be locally developed North claims it can carry a large, heavy nuclear warhead	North Korea's first Intercontinental ballistic missile, launched on July 4

Парк автомобилей



Автомобилизация



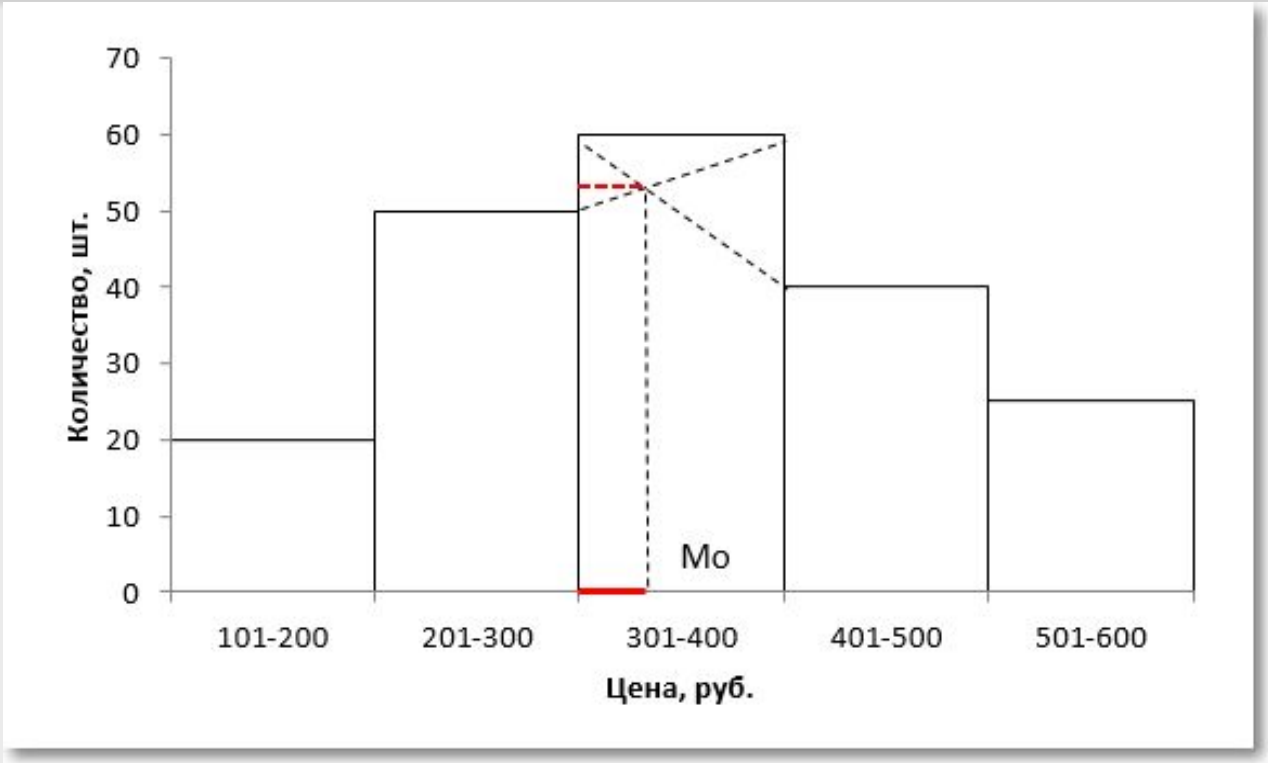
Показатели центра распределения

Показатели центра распределения

- Арифметическое среднее значение
- Мода
  - Для атрибутивного ряда (категория занятости) = ?
  - Для дискретного ряда (размер обуви) = ?
  - Для интервального ряда = ?
- Медиана
  - Для атрибутивного ряда (уровень образования) = ?
  - Для дискретного ряда (размер обуви)
  - Для интервального ряда = ?

Мода интервального ряда

Цена, руб.	Количество, шт.
101-200	20
201-300	50
301-400	60
401-500	40
501-600	25



Mo – мода,
x_0 – значение начала модального интервала,
h – размер модального интервала,
f_{Mo} – частота модального интервала,
f_{Mo-1} – частота интервала, находящего перед модальным,
f_{Mo1} – частота интервала, находящего после модального.

$$Mo = x_0 + h\frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo1})}$$

$$Mo = 301 + 100\frac{60 - 50}{(60 - 50) + (60 - 40)} = 301 + 34{,}3 = 334{,}3 \text{руб.}$$

Что не так с

Медиана интервального ряда

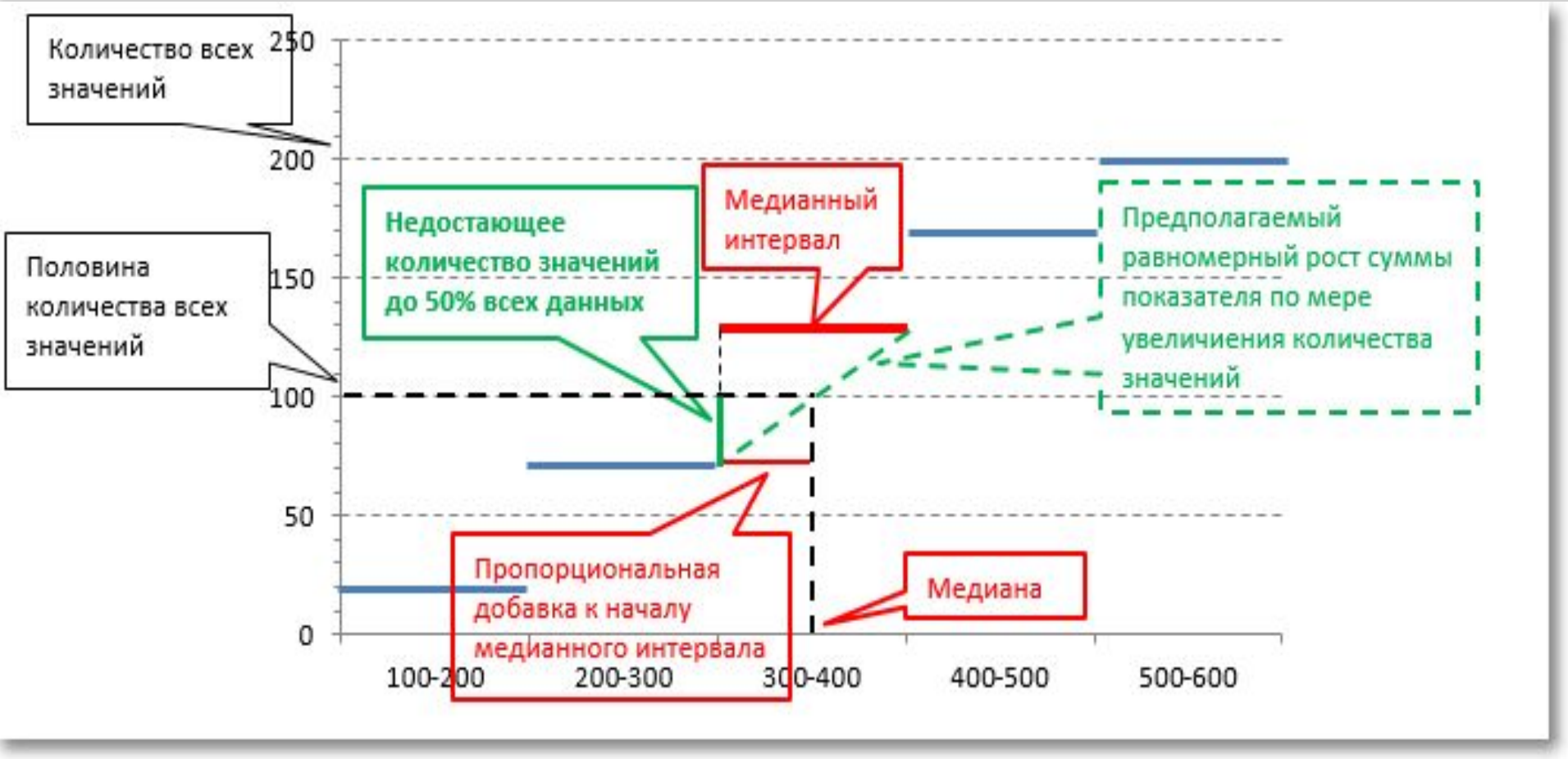
Кумулята = нарастающий итог

Цена, руб.	Количество, шт.	Накопленная частота, шт.	Накопленная доля
100-200	20	20	10,0%
200-300	50	70	35,0%
300-400	60	130	***65,0%***
400-500	40	170	85,0%
500-600	30	200	100,0%
Итого	**200**		

1. Интервал, в котором середина =медианный интервал. **Где он?**
2. В нем ищем единственное значение.

Как понимать

Медиана интервального ряда



$$Me = x_{Me} + i_{Me} \frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}} = 300 + 100 * \frac{\frac{200}{2} - 70}{60} = 350 \text{ руб.}$$

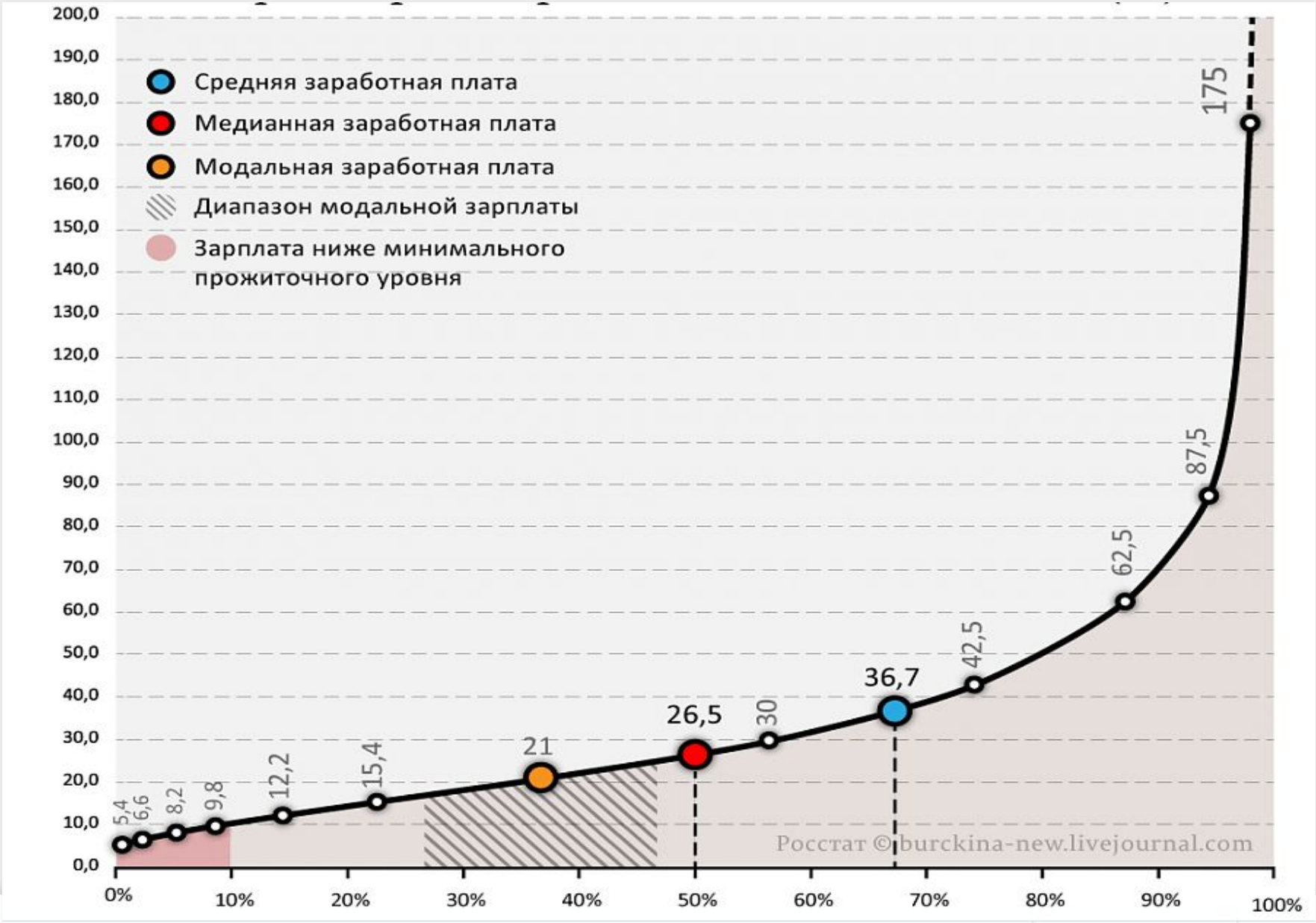
Медиана интервального ряда

$$Me = x_{Me} + i_{Me}\frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}} = 300 + 100 * \frac{\frac{200}{2} - 70}{60} = 350 \text{ руб.}$$

где

- x_{Me} — нижняя граница медианного интервала;
- i_{Me} — ширина медианного интервала;
- $\sum f/2$ — количество всех значений, деленное на 2 (два);
- $S_{(Me-1)}$ — суммарное количество наблюдений, которое было накоплено до начала медианного интервала, т.е. накопленная частота предмедианного интервала;
- f_{Me} — число наблюдений в медианном интервале.

Заработная плата в РФ 2016



ПОЛИТЕХ

Forbes

25 САМЫХ ДОРОГИХ РУКОВОДИТЕЛЕЙ КОМПАНИЙ – 2016

№	CEO	Компания	Оценка вознаграждения, $ млн		Суммарное вознаграждение ключевых менеджеров, $ млн		Изменение, %	Выручка, $ млрд (2015 год)
			в 2015 году	в 2014 году	в 2015 году	в 2014 году		
1	Алексей Миллер	Газпром	17,7	27	78,3	113,8	–31	140,4
2	Игорь Сечин	Роснефть	13	17,5	53,1[3]	181	–71	82,7
–	Андрей Костин	ВТБ		21	99,5[4]	152,8	–35	19,6
3	Герман Греф	Сбербанк	11	13,5	47,8	73,2	–35	22,4
4	Дмитрий Разумов	Онэксим	10	15	н/д	н/д	н/д	н/д
5	Иван Стрешинский	USM Advisors	10	15	н/д	н/д	н/д	н/д
6	Владислав Соловьев	UC Rusal	7,4[6]	4,4[6]	23	16,2	42	8,7
7	Михаил Шамолин	АФК Система	7,4[5]	11,6[5]	75,3	114,1	–34	11,6
8	Михаил Задорнов	ВТБ24	6,5	8,5	14,7	35	–58	5
9	Андрей Акимов	Газпромбанк	6,3	8	29,3	38	–22	6,5
10	Рубен Аганбегян	Открытие	6	11	49,2	59	–16	5,1
11	Жан-Ив Шарлье[1]	Vimpelcom	5,7	н/д	68,6	48	43	9,6
12	Максим Соков	En+	5	7	н/д	н/д	н/д	н/д
13	Алексей Марей	Альфа-Банк	4,5	4,5	22,3	18,3	22	4,1
14	Гульжан Молдажанова	Базовый элемент	4	6	н/д	н/д	н/д	н/д
15	Александр Дюков	Газпром нефть	3,8	7	25,5	36,8	–31	23,9
16	Наиль Маганов	Татнефть	3,6	5	29,3	41	–28	9
17	Павел Грачев	Полюс	3,5	н/д	14,1	3,2	341	2,2
18	Борис Ковальчук	Интер РАО	3,5[6]	н/д	13,3	15,7	–15	13,1
19	Андрей Варичев	Металлоинвест	3,4	н/д	23	38,4	–40	4,4
20	Алексей Москов	Ренова	3	4,5	н/д	н/д	н/д	н/д
21	Николай Токарев	Транснефть	3	н/д	26,1	40,6	–36	13,3
22	Андрей Дубовсков	МТС	2,9	7	17,8	44,6	–60	7
23	Игорь Шехтерман[2]	X5 Retail Group	2,8[6]	н/д	18,5	5,9	213	13,2
24	Дмитрий Конов	Сибур	2,6[7]	5,5	13,5	37,9	–64	6,2

[1] С марта 2015-го. [2] С ноября 2015-го. [3] Вознаграждение членов правления. [4] Данные консолидированной отчетности по Группе ВТБ.
[5] Данные компании. [6] Данные из отчета компании. [7] Без учета выплат на основе акций.

Источник: данные компаний, расчеты Forbes, оценки экспертов.
Для банков в качестве выручки взята сумма процентных и комиссионных доходов

Показатели структуры распределения

Показатели структуры распределения

- Медиана
- Кварт’или
- Дец’или

Децильный коэффициент - соотношение средних доходов 10 % самых богатых жителей государства к такому же проценту беднейших.
РФ 2007 – 16,7; 2016 – 15,7

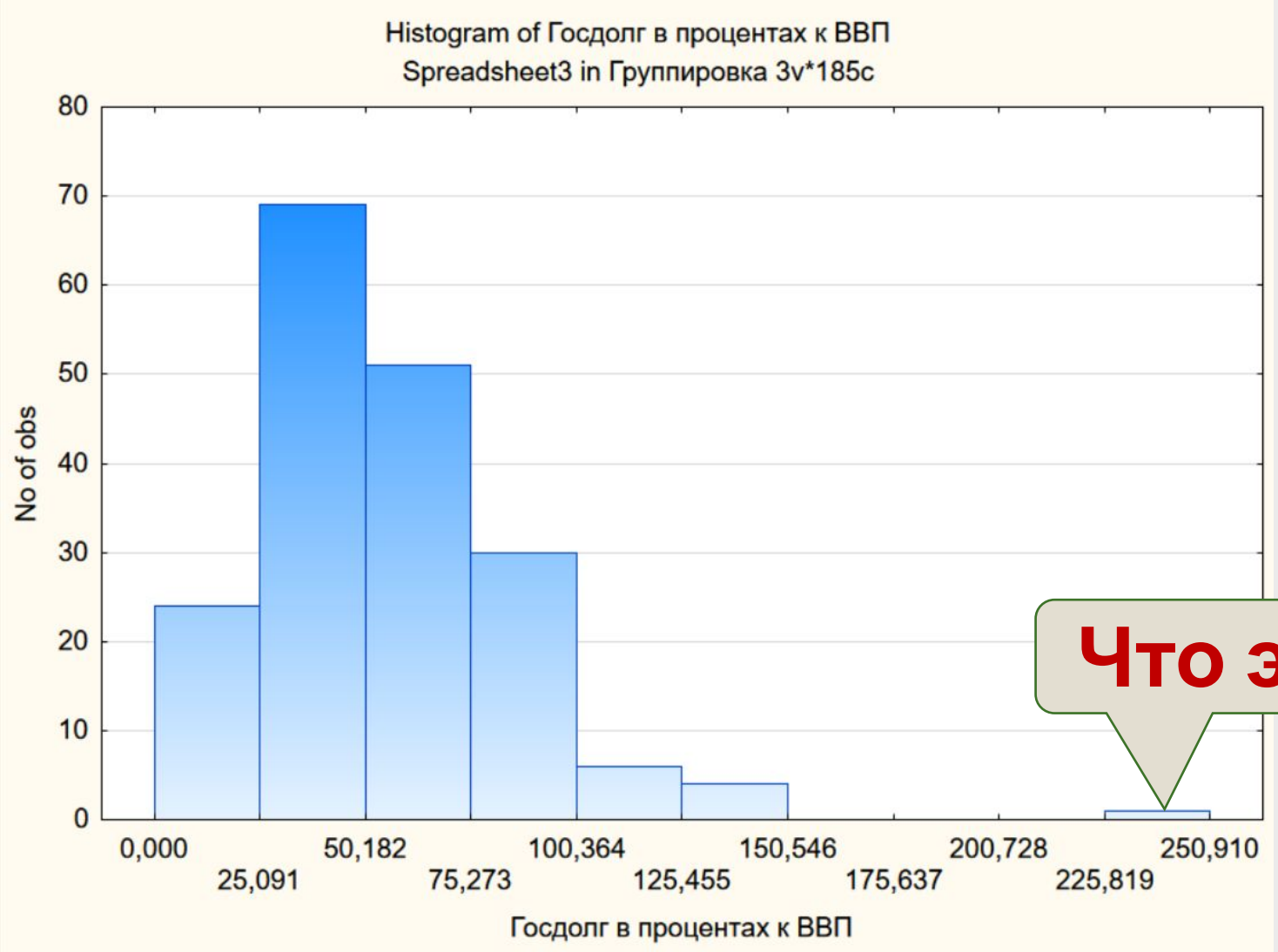
- **Перцентили**

Перцентили

- Это характеристики данных, которые выражают ранги элементов в виде процентов (от 0 до 100%), а не в числах.
- Наименьшему значению признака соответствует нулевой перцентиль, наибольшему – 100-й.
- Перцентили – это показатели, разбивающие ранжированный ряд данных на определенное число частей.

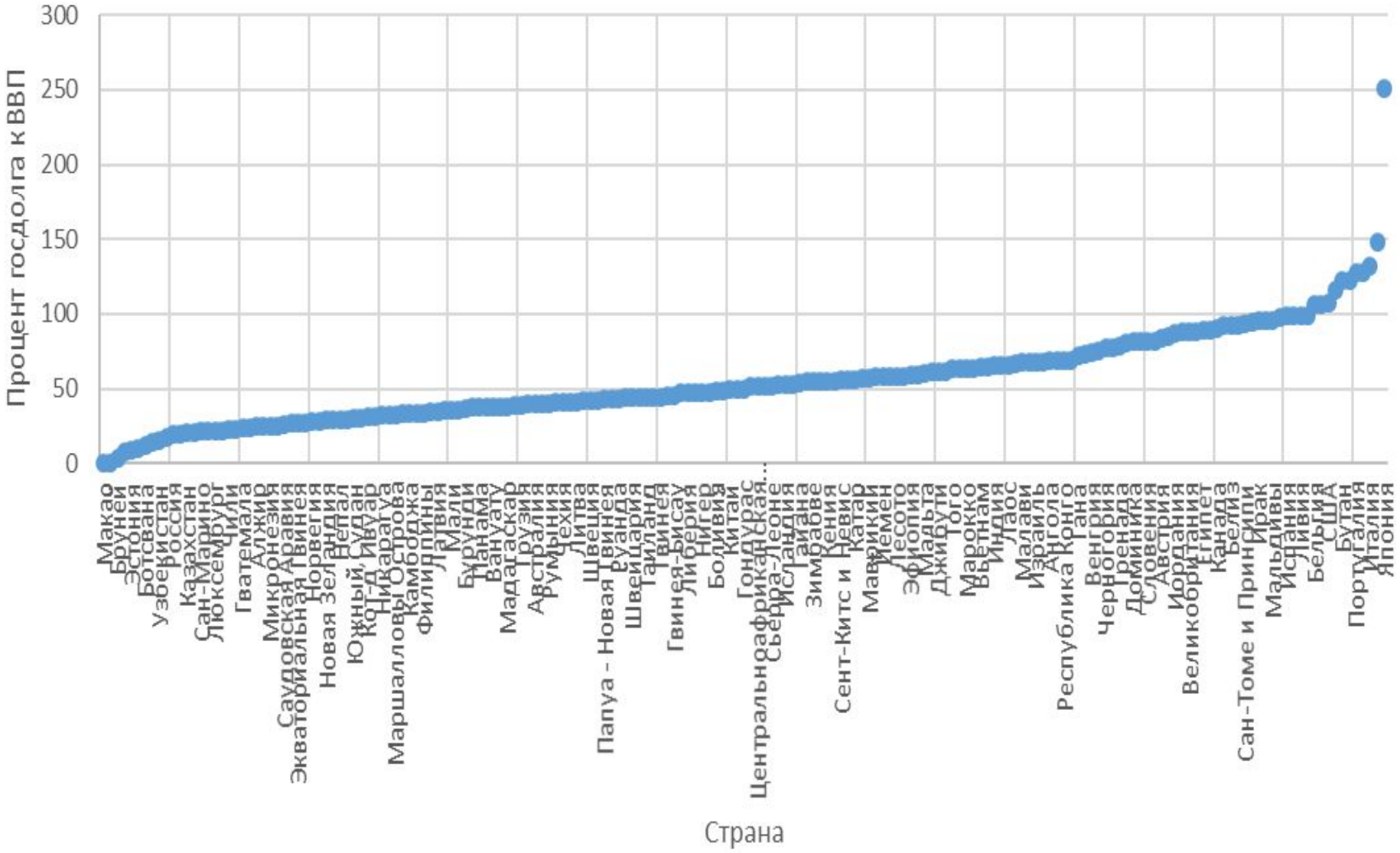
Выбросы

Было: 10 групп Что это?



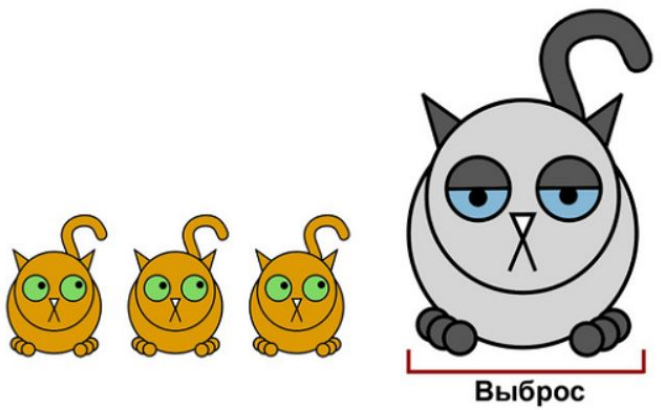
ПОЛИТЕХ

Выбросы видны в ранжированном ряду



Выбросы

- Это единицы совокупности, значения признака которых резко отличаются в меньшую или большую сторону от основной массы значений призна



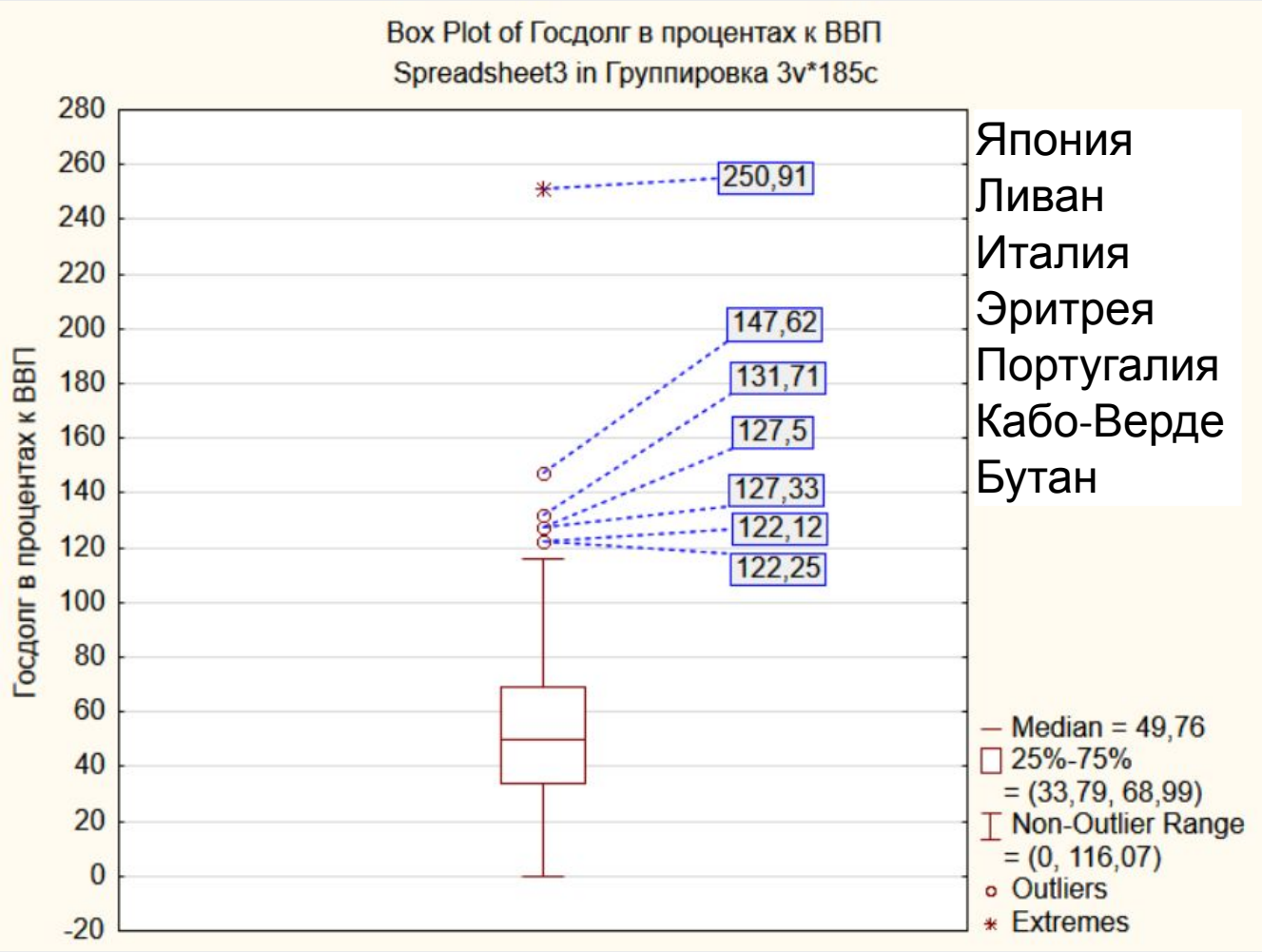
- Данные единицы не подчиняются общей закономерности распределения, поэтому **анализируются отдельно.**

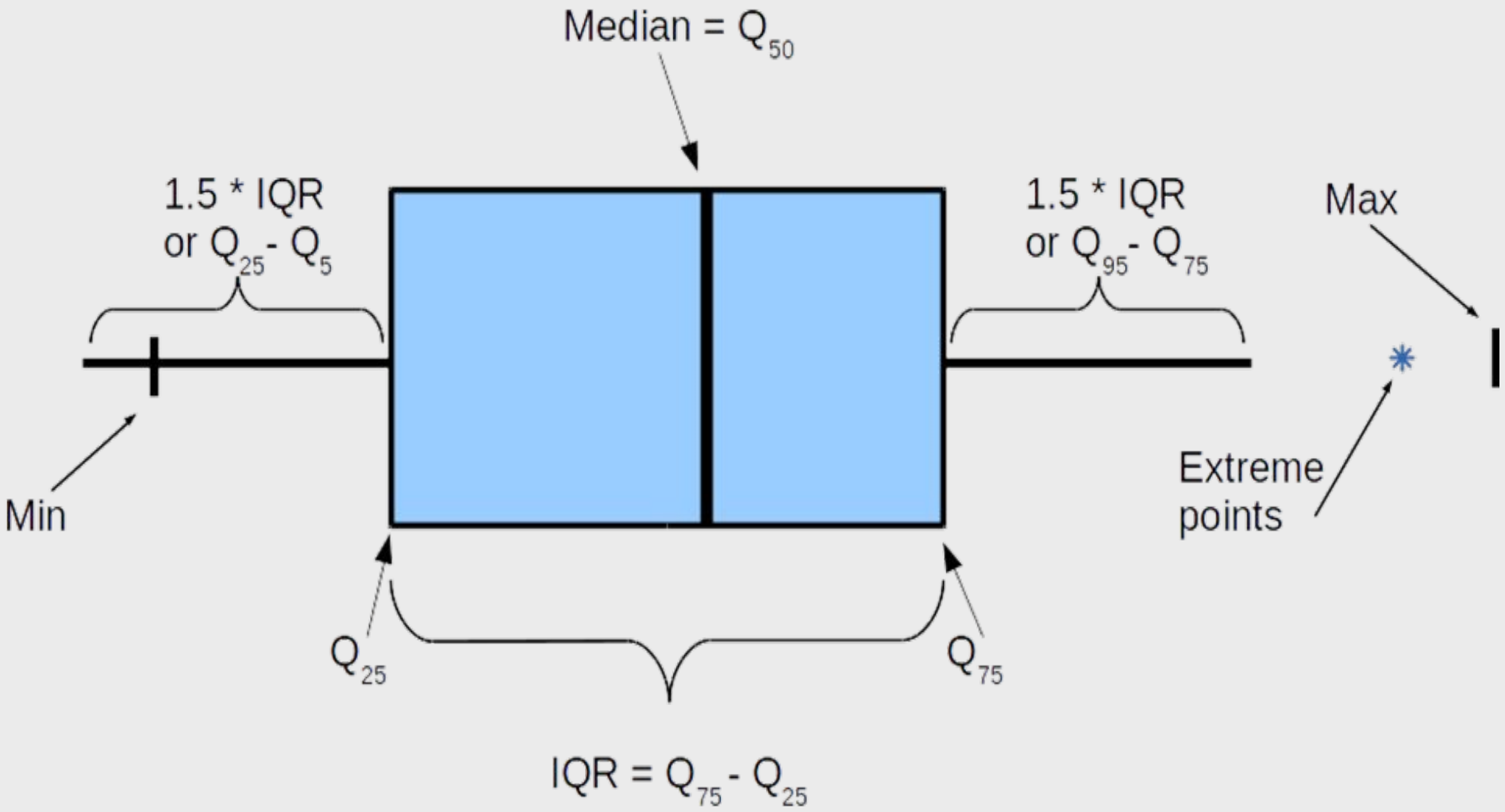
Метод Тьюки

- Границы ящика – 1-й (снизу) и 3-й квартили
- Ширина ящика = интерквартиль-ный размах
- Ус = полтора интерквартиль-ных размаха от ящика

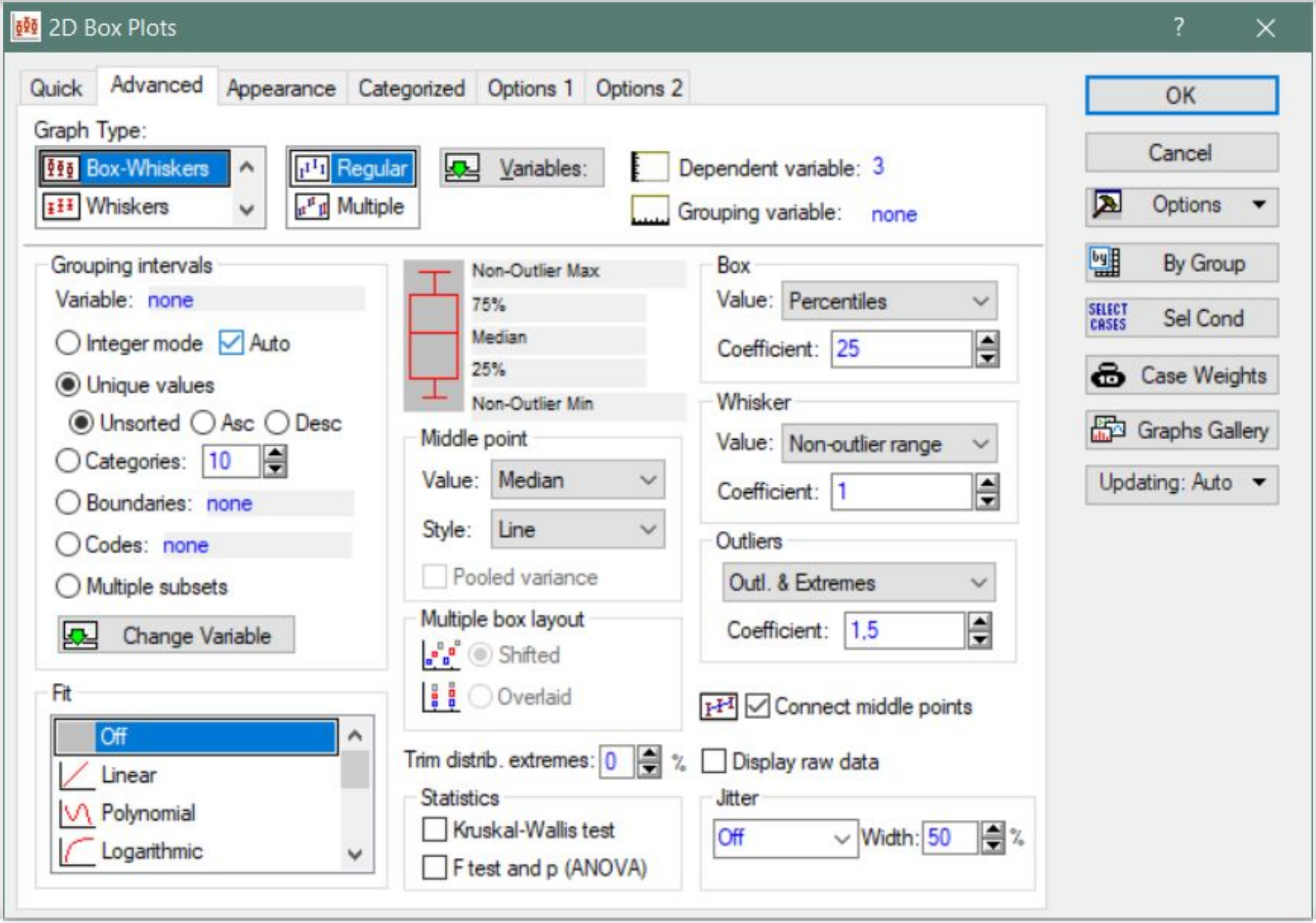
Почему медиана не посередине?

Почему разные усы???

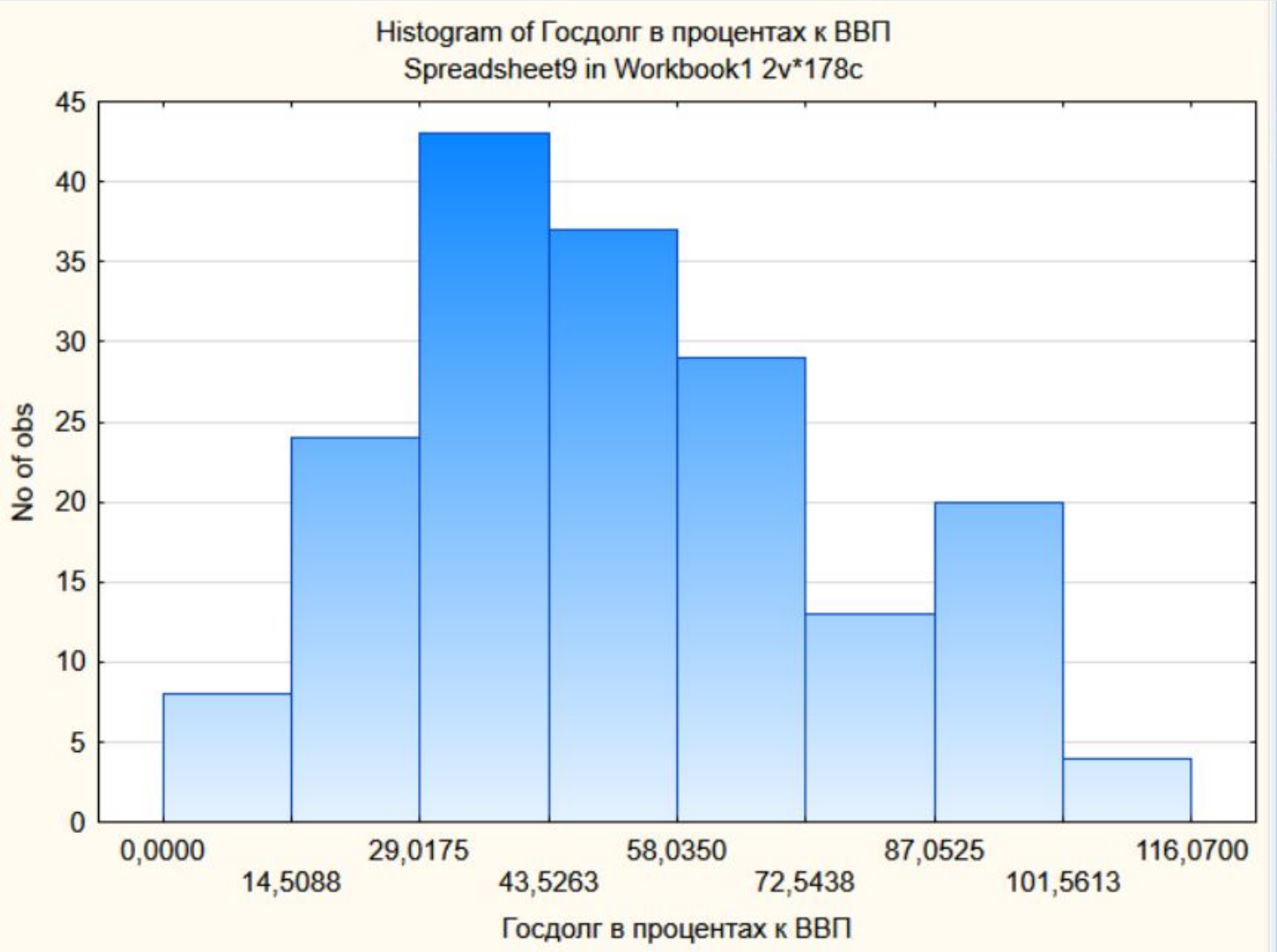



Median = Q_{50}
1.5 * IQR
or Q_{25} - Q_5
1.5 * IQR
or Q_{95} - Q_{75}
Max
Min
Extreme
points
Q_{25}
Q_{75}
IQR = Q_{75} - Q_{25}

Метод Тьюки Statistica 12

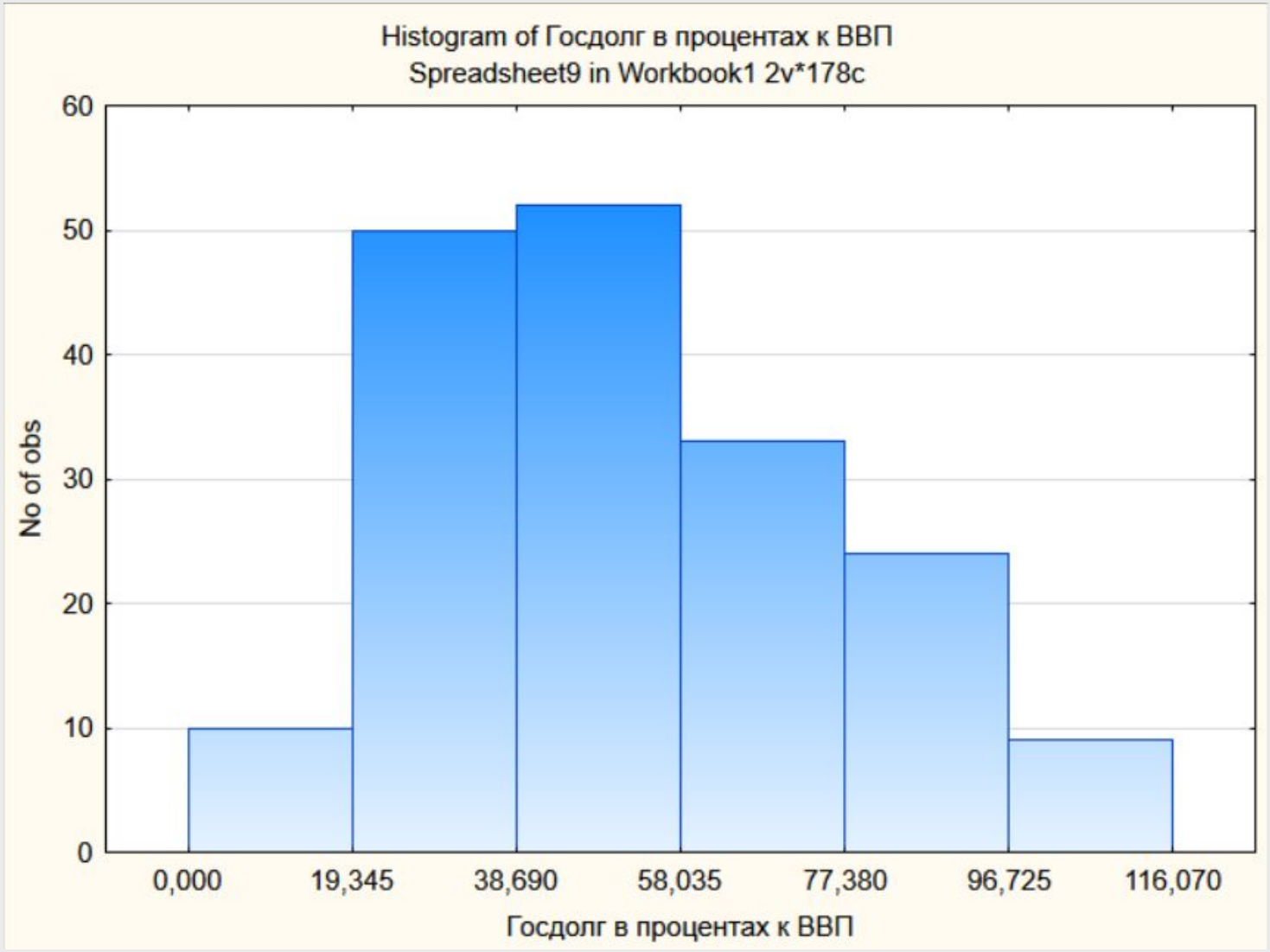


Правильная группировка



Двухмодальное распределение

Правильная группировка



Одномодальное распределение

ПОЛИТЕХ

Что делать с выбросами?

Источник: данные ФТС, аналитика IndexBox

Рисунок 1. Динамика импорта роз в РФ в 2011 г.. млн. штук

Источник: Росстат

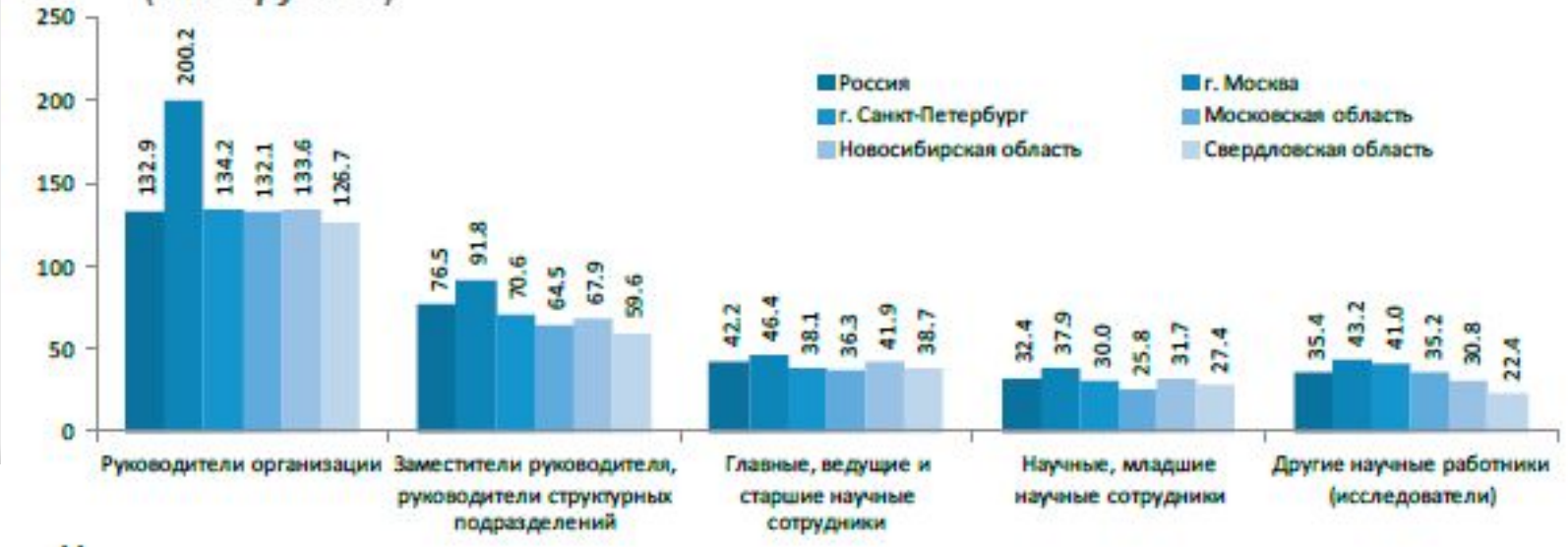
Показатели вариации

Что это и зачем

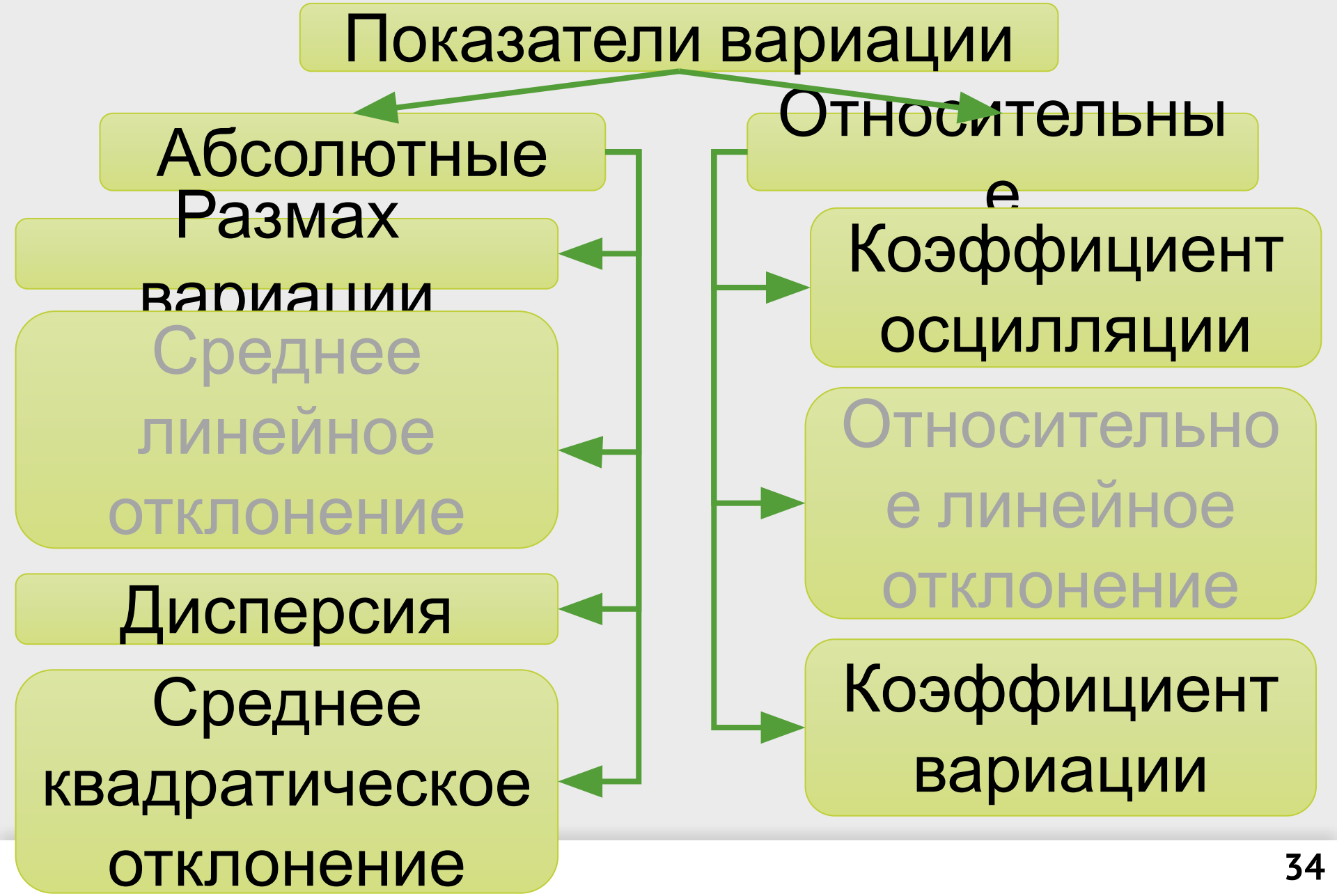
Вариация = различия в индивидуальных признаках единиц совокупности.

- Малая вариация => среднее = типичное
- Большая вариация => среднее ≠ типичное
- И вообще интересно, насколько варьирует заработная плата в России и других странах

Рис. 1. Среднемесячная начисленная заработная плата в организациях, выполняющих исследования и разработки, по категориям персонала: январь–март 2017 г. (*тыс. рублей*)



Показатели вариации



Абсолютные показатели вариации

ПОЛИТЕХ

Размах вариации

$$R = X_{max} - X_{min}$$

X_{max}, X_{min} – максимальное и минимальное значения признака в изучаемой совокупности

Зависит от двух измерений, поэтому неустойчив **Это как так?**

Пример: размах зарплаты по СПб

Среднее линейное отклонение

$$\overline{d} = \frac{\sum \left| x_i - \overline{x} \right| * f_i}{\sum f_i}$$

$\overline{x}$ среднее значение признака в совокупности;

x_i индивидуальные значения признака;

f_i - вес или частота (частость).

Формула попроще, без взвешивания = ?

Дисперсия

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Физического смысла нет, но часто используется

Среднее квадратическое отклонение

= стандартное = типовое отклонение

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Правило Бьеномэ-Чебышева

- Независимо от формы распределения, процент наблюдений, лежащих на расстоянии, не превышающем ***k*** стандартных отклонений от среднего значения, не меньше:

$$\left[1-(1/k)^2\right]\times 100\%$$

- для ***k=2***:

$$\left[1-(1/2)^2\right]\times 100\% = 75\%$$

$$(\overline{X} \pm 2\sigma) \rightarrow 75\% наблюдений$$

Относительные показатели вариации

Коэффициент осцилляции

$$K_R = \frac{R}{\overline{x}}$$

$\boldsymbol{R}$ – размах вариации,

$\overline{x}$ – среднее значение признака в совокупности.

Относительное линейное отклонение

$$K_{\bar{d}} = \frac{\bar{d}}{\bar{x}}$$

$\bar{d}$ - среднее линейное отклонение,

$\bar{x}$ среднее значение признака в совокупности.

Коэффициент вариации

$$V = \frac{\sigma}{\overline{x}} \cdot 100$$

$\overline{x}$ - среднее значение признака в совокупности;

σ - среднее квадратическое (стандартное) отклонение.

Пример коэффициента вариации

- Средняя заработная плата 50 тыс, **СКО = 5 тыс**.
- Прогноз ВВП РФ на следующий год (мой личный) 75 трлн. руб = 75 000 000 млн. руб. **СКО = 5 млн**.

Какая оценка более точна?

- Вариация заработной платы = 5/50*100 = 10%
- Вариация ВВП = 5/75 000 000*100 = 0,000007%

Характеристики формы распределения

Коэффициент асимметрии Пирсона

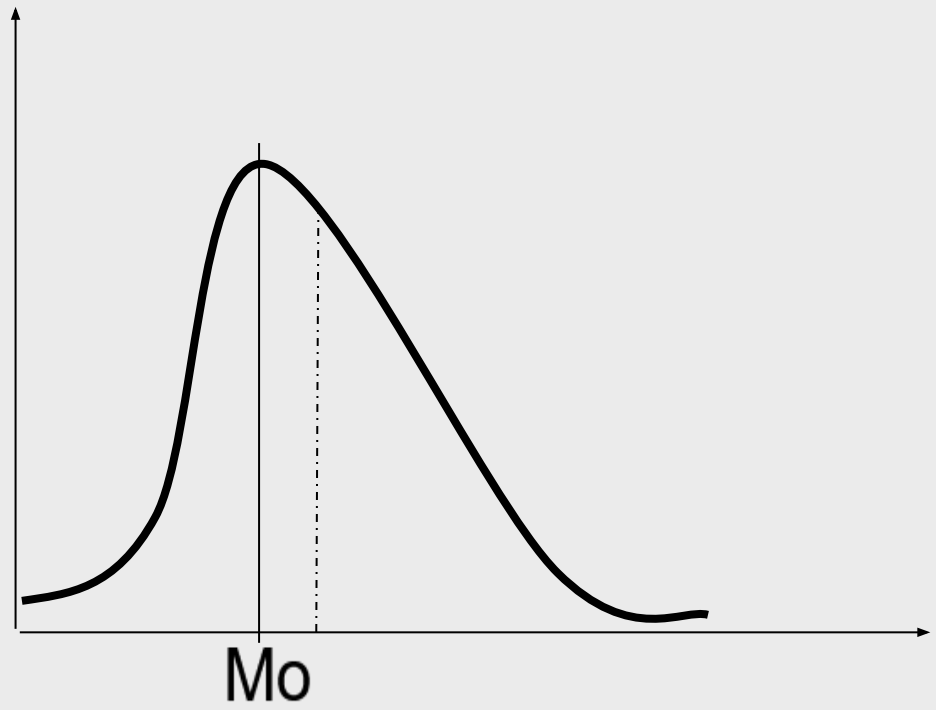
$$A_s = \frac{\bar{x} - (M_0)}{\sigma}$$

Мо – мода,

σ – среднее квадратическое (стандартное) отклонение.

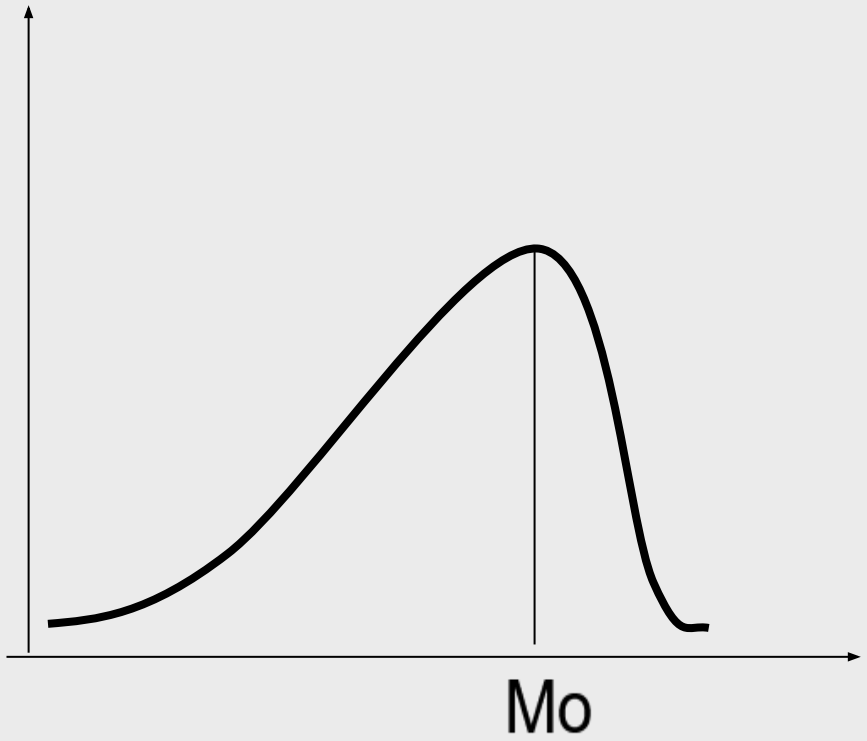
Асимметрия

Асимметрия



Правосторонняя,

$$\bar{x} > Mo$$



Левосторонняя,

$$\bar{x} < Mo$$

Распределение населения по доходам

Что произошло?

Сдвиги в плотности распределения населения России по месячному душевому доходу за 1990–2008 годы

График 5

▲ Доля населения с доходом в окрестности ±500 руб. от данного дохода (%)

► Доходы (тыс. руб. в мес. в ценах 2008 г.)

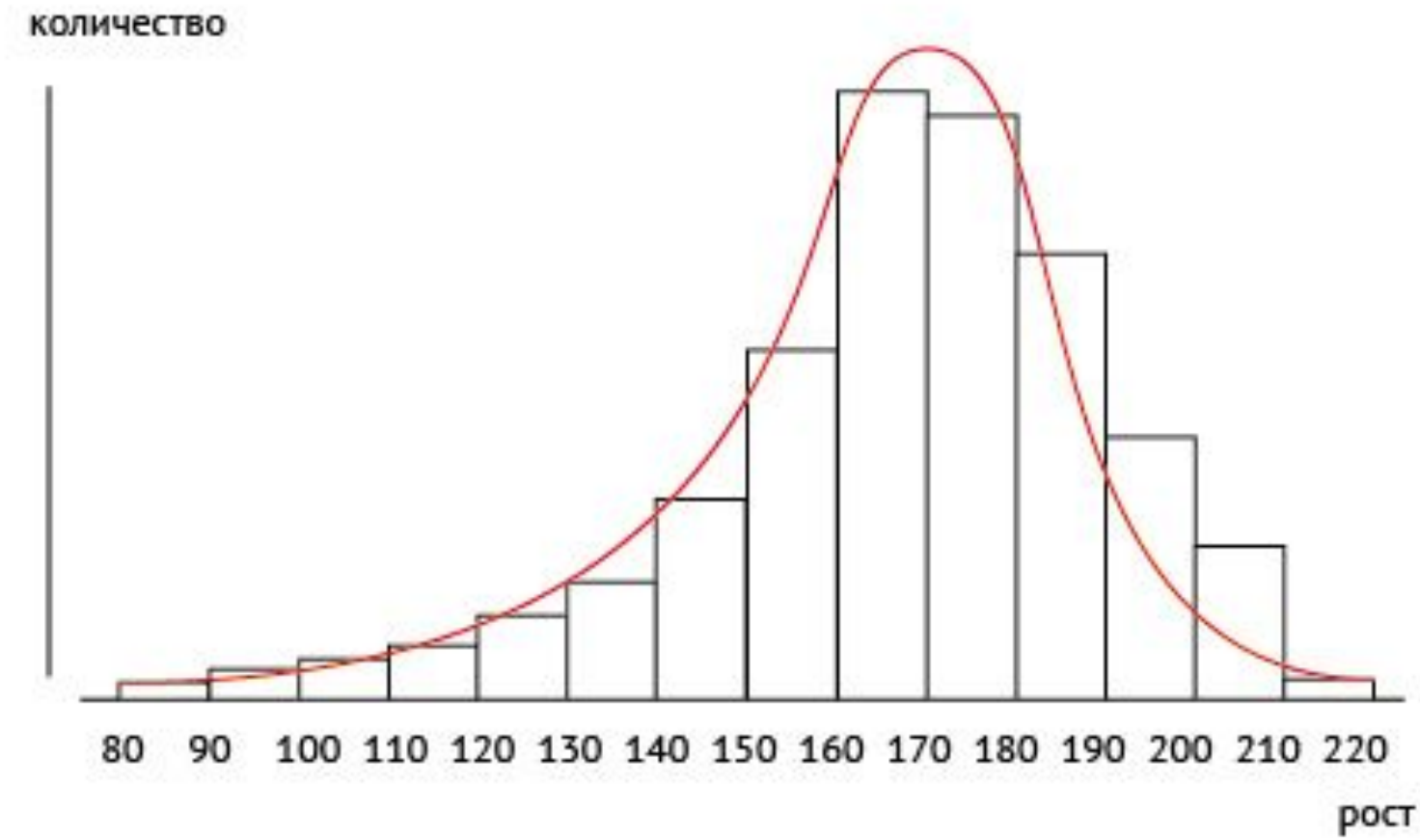
■ 1990 г.

■ 2008 г.

Источник: Росстат

Распределение населения по величине среднедушевых денежных доходов

Рост человека

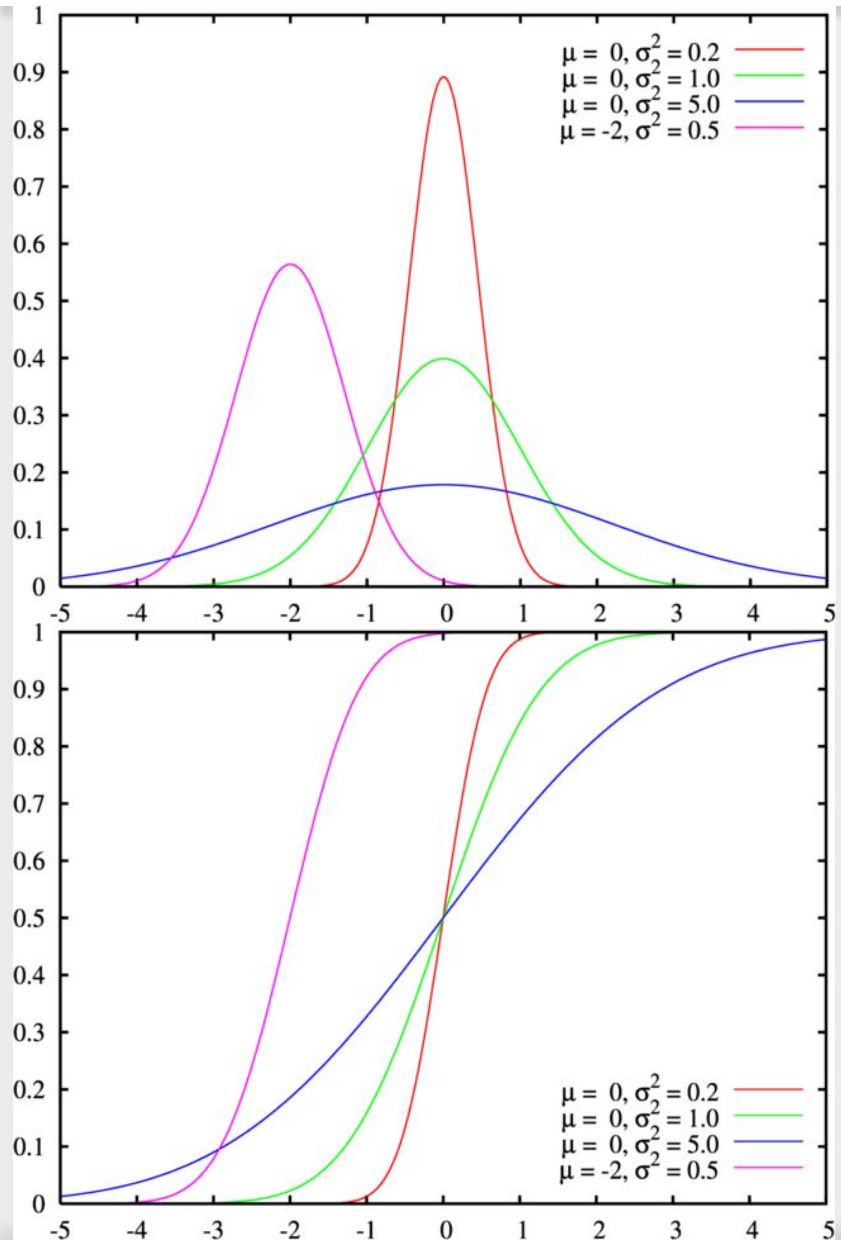


Рост человека

Законы распределения вероятностей

- Нормальное
- Логарифмически нормальное
- Пуассона
- Биноминальное
- … … … …

Нормальное распределение

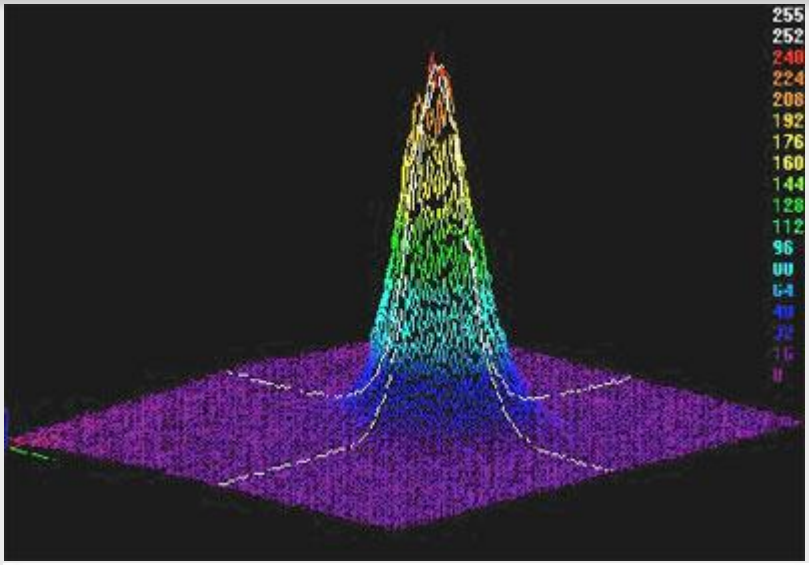


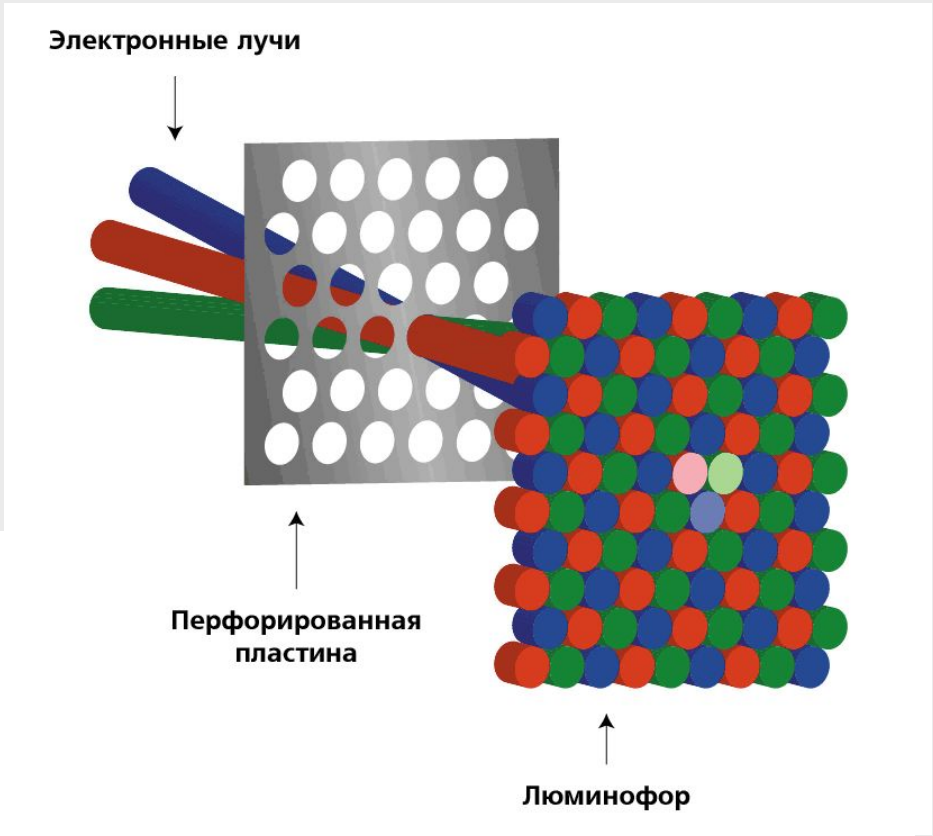
- Плотность распределения

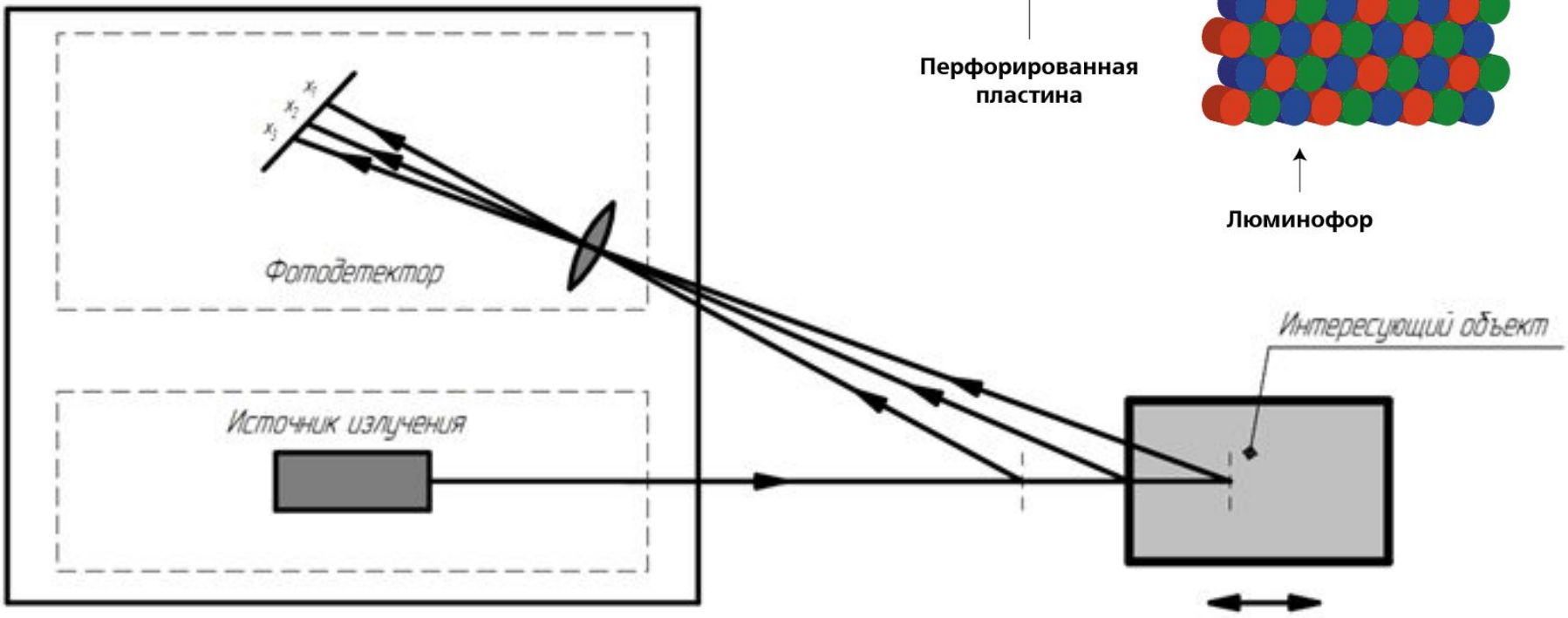
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **Сумма независимых одинаково распределенных случайных величин**
- Давление крови (?)
- Отклонения при стрельбе
- Лазерный луч (ниже)

О лазерном луче



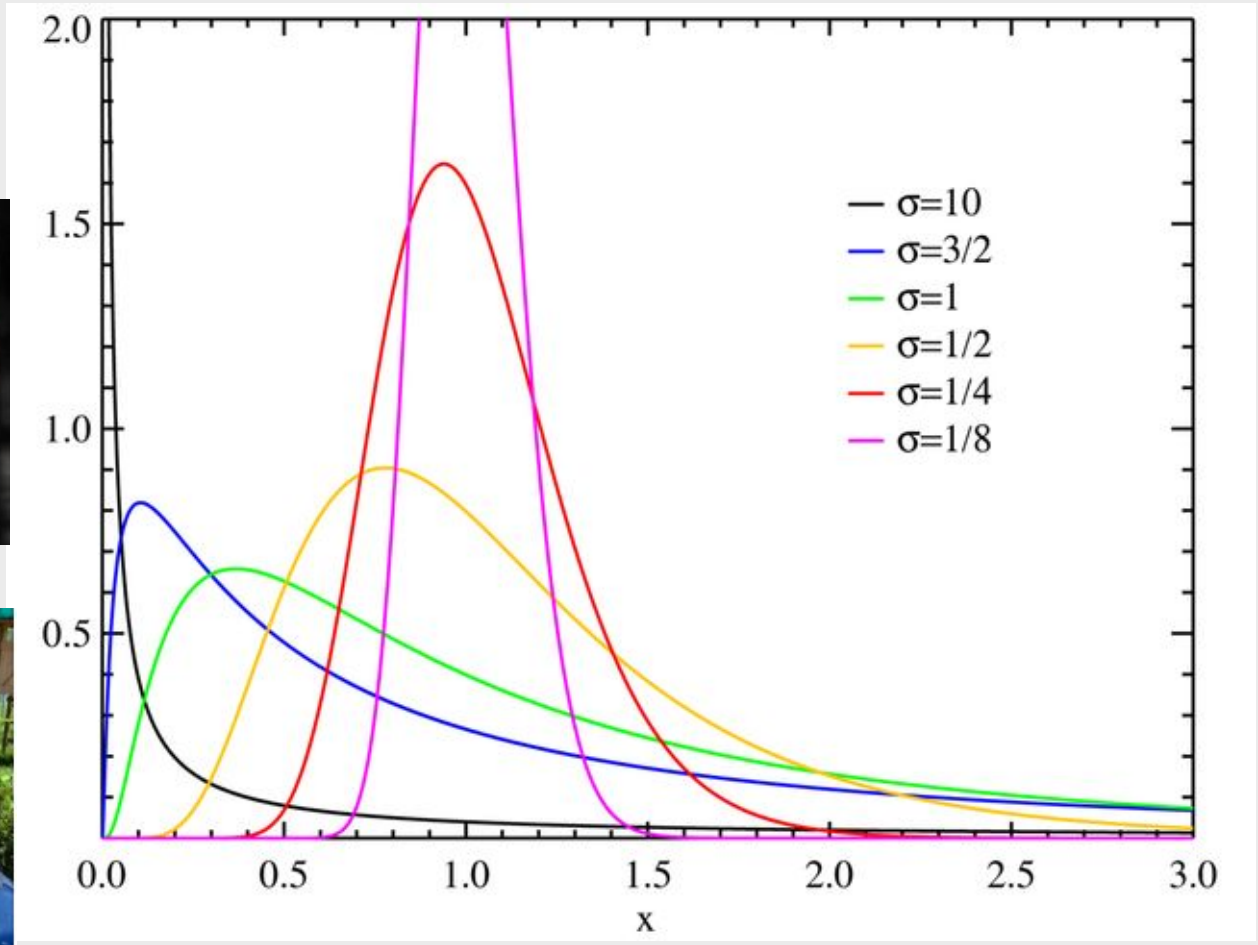




Логарифмически нормальное

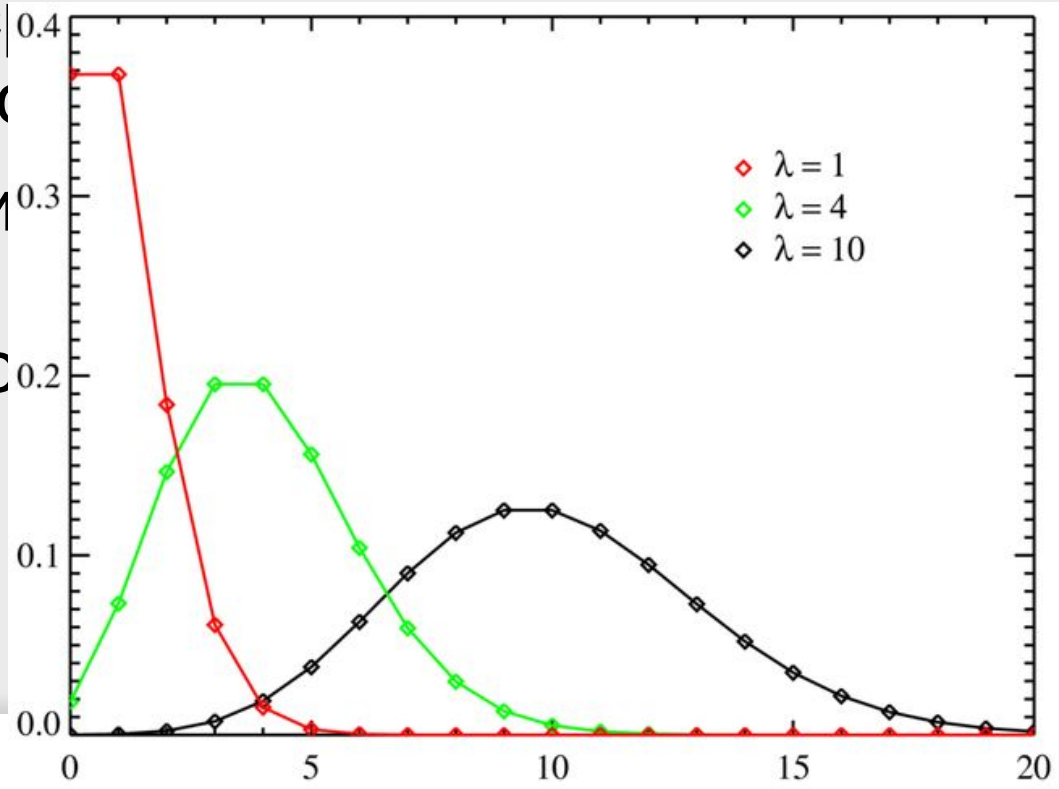
- Логарифм величины имеет нормальное распределение

- Размер градин



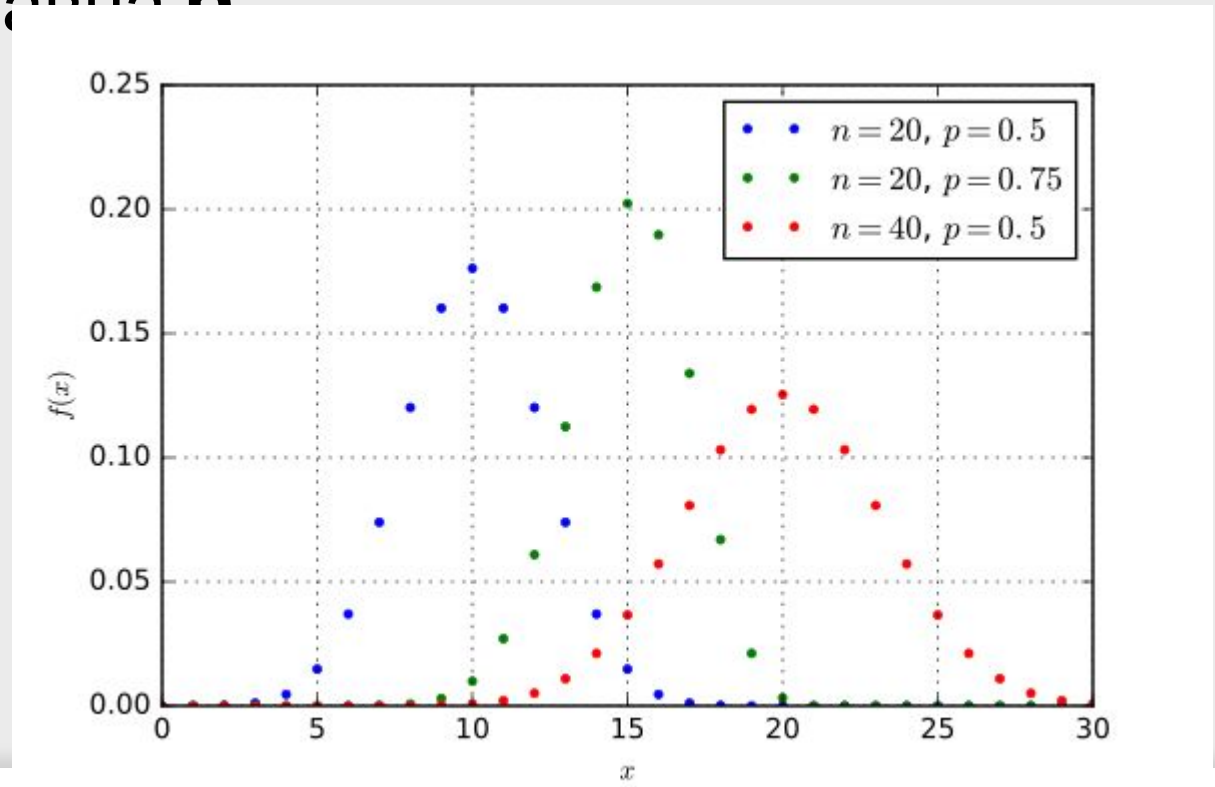
Распределение Пуассона

- Вероятностное распределение дискретного типа.
- Моделирует **число событий, произошедших за фиксированное время**, при условии, что данные события происходят с некоторой фиксированной с
  независимо друг
- Используется при
  моделировании
  систем массового
  обслуживания

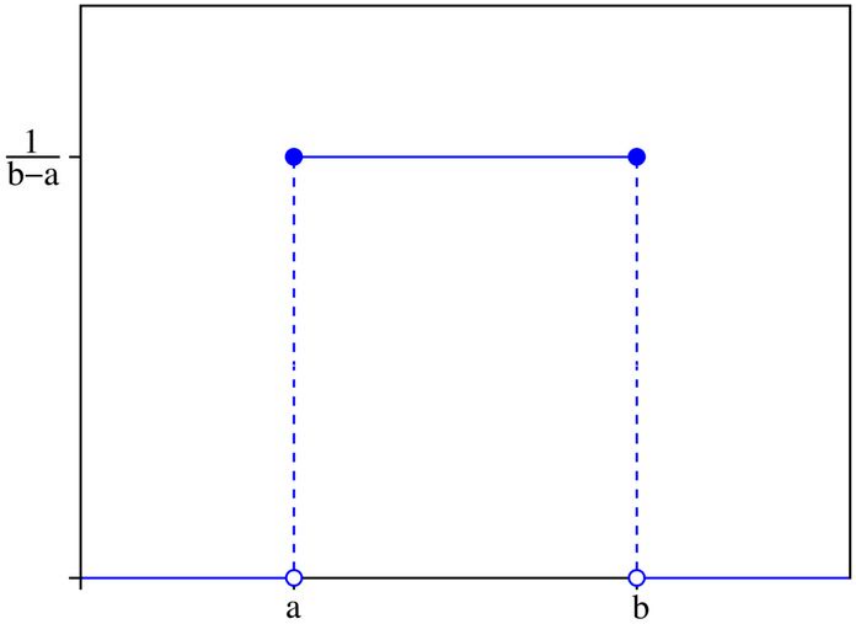


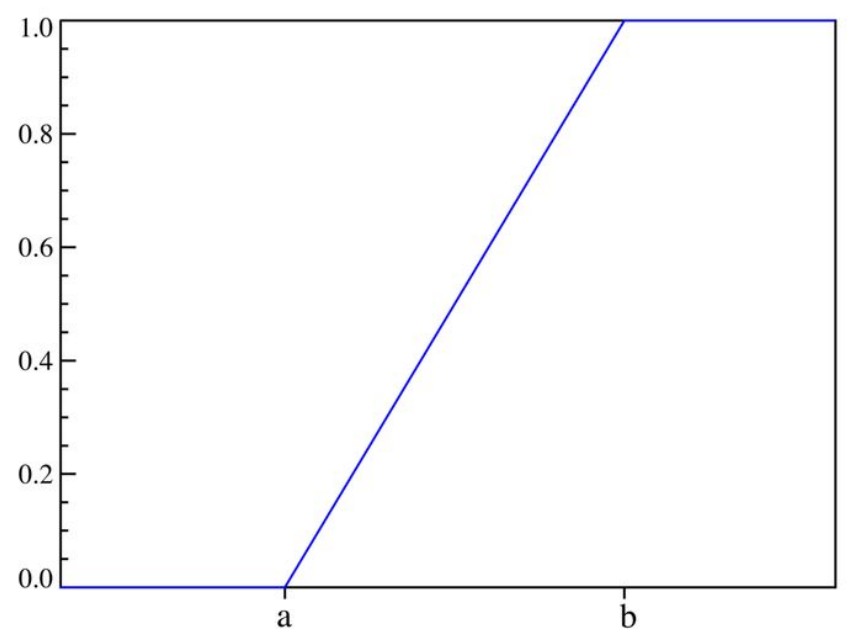
Биноминальное распределение

- распределение **количества «успехов»** в последовательности из **n** независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна **p**



Равномерное распределение





«Генерация случайных чисел слишком важна, чтобы оставлять её на волю случая.» *Роберт Кавью*

«Всякий, кто питает слабость к арифметическим методам получения случайных чисел, грешен вне всяких сомнений.» *Джон фон Нейман*

=слчис()

Как получить нормальное распределение с помощью этой

Законы распределения вероятностей

- Реальность всегда не идеальна
- Требуется проверить близость реальных данных теоретическому распределению
- Эта область = проверка гипотез, будет ниже.

Конец части 2