ОБРАБОТКА И МОДЕЛИРОВАНИЕ ДАННЫХ В MS EXCEL

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Екатерина Золотарева

ТЕМА 1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

С чем работаем

Данные – сведения, зафиксированные в определенных форматах и пригодные для дальнейшего использования

Информация – результат **обработки** данных для решения конкретных задач
Знания – проверенная информация, которая может **многократно** использоваться для принятия решений

Польза 

Какими бывают данные

Структура – объекты и признаки

Количественные признаки:

- Дискретные
- Непрерывные

Качественные:

- Номинальные
- Порядковые

Данные могут быть неточными, неполными, противоречивыми, разнородными, косвенными….
И иметь гигантские объемы

Какими бывают данные

A	C	D	E	F	G	H	I	J	K
Но	Статус кредита	Размер кре	Срок кредита	Кредит	Годовой доход	Стаж работы на	Недвижимость	Цель кредита	Ежемесячный платеж
1	погашен	445 412,00	краткосрочный	709	1 167 493,00	8 лет	в ипотеке	ремонт жилья	5 214,74
2	погашен	262 328,00	краткосрочный			10+ лет	в ипотеке	консолидация кредитов	33 295,98
3	погашен		краткосрочный	741	2 231 892,00	8 лет	в собственности	консолидация кредитов	29 200,53
4	погашен	347 666,00	долгосрочный	721	806 949,00	3 года	в собственности	консолидация кредитов	8 741,90
5	погашен	176 220,00	краткосрочный			5 лет	в аренде	консолидация кредитов	20 639,70
6	не погашен	206 602,00	краткосрочный	729	896 857,00	10+ лет	в ипотеке	консолидация кредитов	16 367,74
7	погашен	217 646,00	краткосрочный	730	1 184 194,00	< 1 года	в ипотеке	консолидация кредитов	10 855,08
8	не погашен	648 714,00	долгосрочный			< 1 года	в ипотеке	приобретение жилья	14 806,13
9	погашен	548 746,00	краткосрочный	678	2 559 110,00	2 года	в аренде	консолидация кредитов	18 660,28
10	погашен	215 952,00	краткосрочный	739	1 454 735,00	< 1 года	в аренде	консолидация кредитов	39 277,75
11	погашен		краткосрочный	728	714 628,00	3 года	в аренде	консолидация кредитов	11 851,06
12	погашен	541 970,00	краткосрочный			10+ лет	в ипотеке	ремонт жилья	23 568,55
13	погашен		краткосрочный	740	776 188,00	< 1 года	в собственности	консолидация кредитов	11 578,22
14	погашен		краткосрочный	743	1 560 907,00	4 года	в аренде	консолидация кредитов	17 560,37
15	погашен	234 124,00	краткосрочный	727	693 234,00	10+ лет	в аренде	консолидация кредитов	14 211,24
16	погашен	449 020,00	долгосрочный			9 лет	в собственности	консолидация кредитов	18 904,81
17	не погашен	653 004,00	долгосрочный			7 лет	в ипотеке	консолидация кредитов	14 537,09
18	погашен	666 204,00	долгосрочный	723	1 821 967,00	10+ лет	в ипотеке	консолидация кредитов	17 612,24

Как анализировать

Анализ данных – обнаружение неизвестных, нетривиальных, практически полезных и интерпретируемых знаний, нужных для принятия решений

Обработка

Моделирование

Теоретическая база: ТВиМС, МатАн, ЛинАл
Инструменты: MS Excel, R, Python, ML Studio

На семинаре

- Подготовка данных
- Гистограммы
- Графики и диаграммы рассеяния
- Условное форматирование

Описательная статистика (количественные данные)

Определение **центра** распределения

- Среднее
- Мода
- Медиана

Среднее значение

Среднее - сумма элементов, поделенная на их количество, например:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

Среднее значение

Функции в MS Excel:

- СУММ(), СУММЕСЛИ(),
- СЧЁТ(), СЧЁТЗ(),СЧЁТЕСЛИ()
- СРЗНАЧ()

Данные о зарплатах: среднее значение

Company	Salary ($mln)	Company	Salary ($mln)
Boeing	21	Delta Airlines	13
Whirlpool	13	Chrysler	1
Bank of America	12	Coca-Cola	22
Sherwin-Williams	11	DuPont	13
Bristol-Myers	16	Motorola	47
General M			9
Hillshire B			16
Wal-Mart	21	Exxon	35
Apple Computers	4	CBS	60
Goodyear	13	AT&T	21
Teledyne	3	Philip Morris	25

Среднее равно 17,9 $mln

Мода

Мода – наиболее часто встречающееся значение

Данные	42	33	42	47	42	47

Мода равна 42 - это самый «популярный» элемент

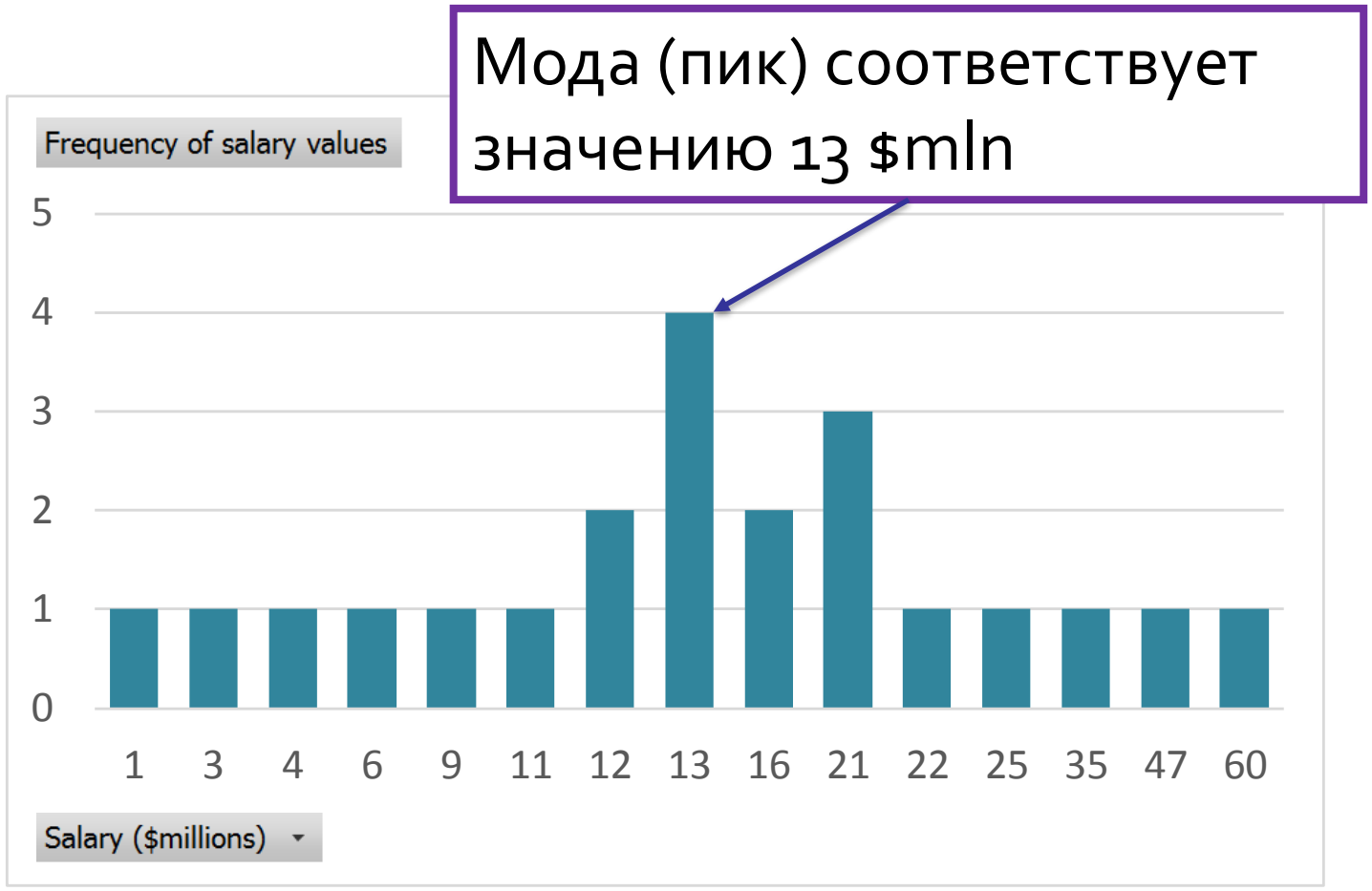
На гистограмме – класс с наибольшей частотой

Данные о зарплатах: мода

Company	Salary ($mln)	Company	Salary ($mln)
Chrysler	1	Goodyear	13
Teledyne	3	Bristol-Myers	16
Apple Computers	4	Honeywell	16
Hillshire Brands	6	Wal-Mart	21
Marriot			21
Sherwin-V			21
Bank of America	12	Coca-Cola	22
General Mills	12	Philip Morris	25
Delta Airlines	13	Exxon	35
DuPont	13	Motorola	47
Whirlpool	13	CBS	60

Мода равна 13 $mln

Мода на графике



Мода

Функции в MS Excel:

- МОДА.ОДН()
- МОДА.НСК()

Мода может быть не единственна, а может и не существовать

Медиана

Медиана – серединный элемент данных, отсортированных по возрастанию или убыванию

Данные	46	54	42	45	32
Сортировка	32	42	45	46	54

Медиана – третье число, то есть 45

- Преимущества: на медиану не влияют очень большие или очень маленькие значения
- Если заменить 54 на 5000, среднее возрастет до 1033, но медиана не изменится

Медиана

В общем случае, медиана – это элемент с порядковым номером (n+1)/2 (актуально для четных n)

Данные	46	54	42	45	32	57
Сортировка	32	42	45	46	54	57

n = 6, (n+1)/2 = 3.5

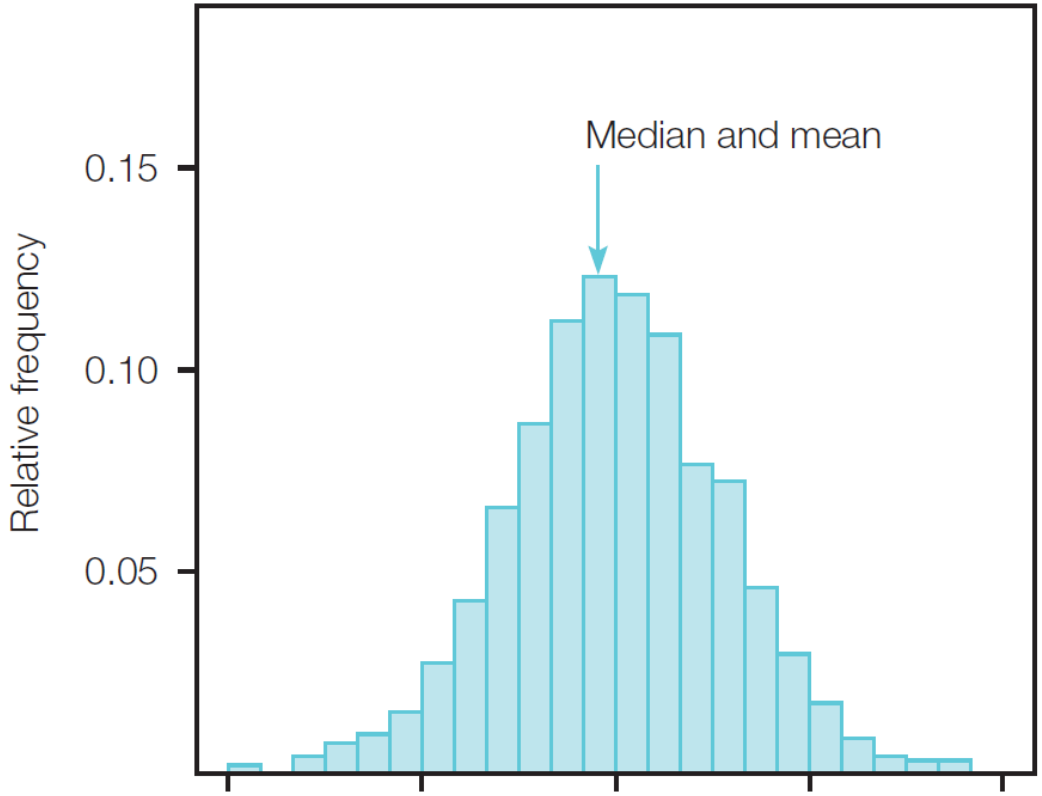
Медиана - среднее между 45 и 46, то есть 45,5

Данные о зарплатах: медиана

Company	Salary ($mln)	Company	Salary ($mln)
Chrysler	1	Goodyear	13
Teledyne	3	Bristol-Myers	16
Apple Computers	4	Honeywell	16
Hillshire Brands	6	Wal-Mart	21
Marriot			21
Sherwin-W			21
Bank of America	12	Coca-Cola	22
General Mills	12	Philip Morris	25
Delta Airlines	13	Exxon	35
DuPont	13	Motorola	47
Whirlpool	13	CBS	60

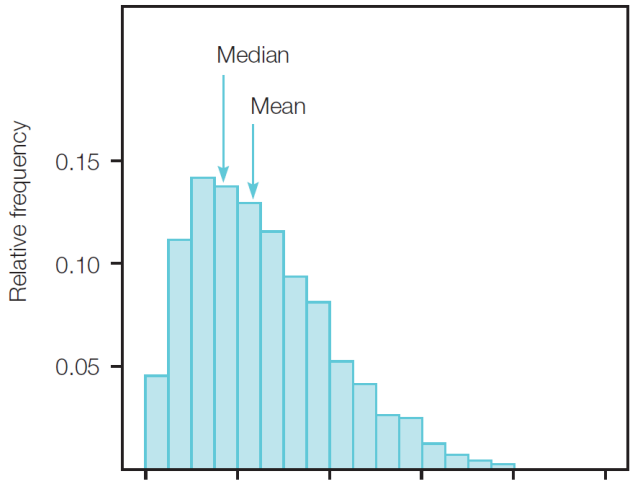
Медиана равна 13 $mln

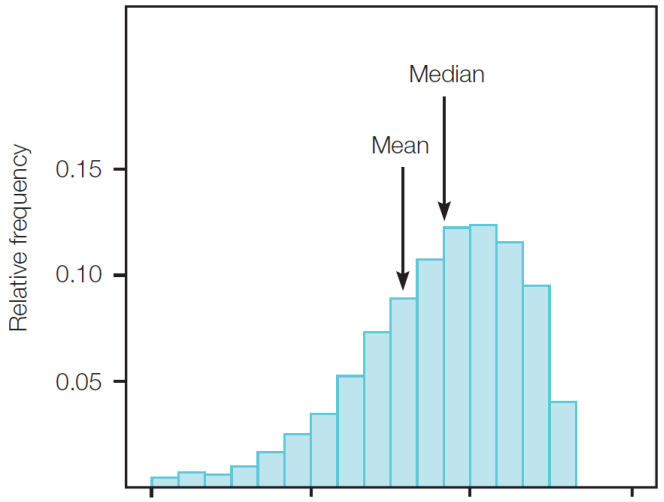
У симметричных распределений среднее и медиана равны



У скошенных распределений среднее и медиана не равны

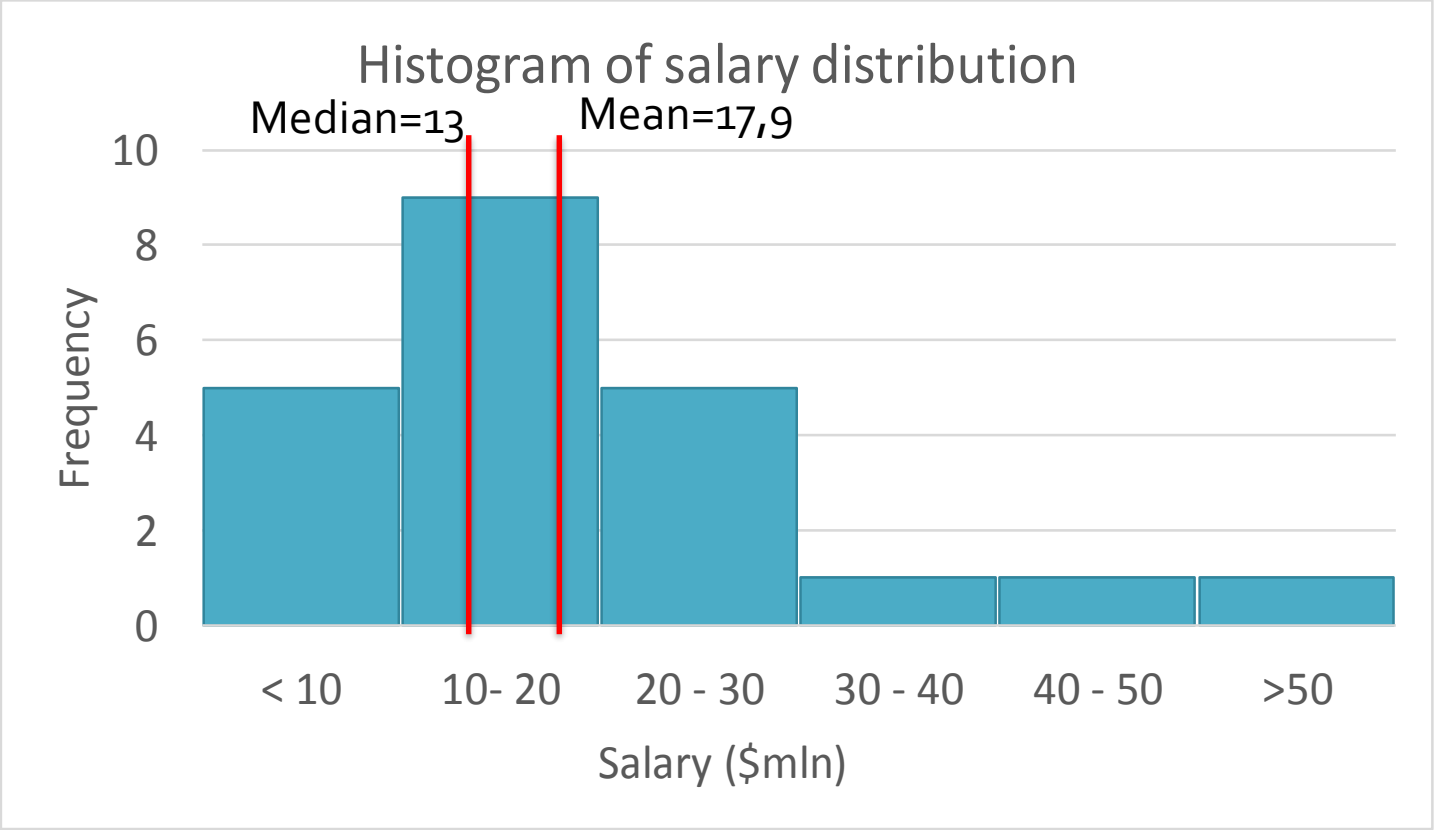
Внизу: Несколько больших значений ‘тянут’ среднее вверх, но это не влияет на медиану – положительная асимметрия (вправо)





Вверху : Несколько больших значений ‘тянут’ среднее вверх, но это не влияет на медиану – отрицательная асимметрия (влево)

Среднее и медиана (Salary)



Медиана < среднего - положительная асимметрия

Медиана

Функции в MS Excel:

- МЕДИАНА()
- КВАРТИЛЬ.ВКЛ(), КВАРТИЛЬ.ИСКЛ()

Описательная статистика (количественные данные)

Измерение **разброса** данных

- Размах
- Квантили/квартили
- Дисперсия,стандартное отклонение

Размах

Размах – разность между максимумом и минимумом

R = Xmax - Xmin

- Пример: Данные: 0, 48, 49, 51, 52, 100

  Размах = 100 (среднее 50, медиана 50)

Недостаток – часть информации теряется

- Пример: Данные: 0, 1, 1, 99, 99, 100

Размах по-прежнему 100, но разброс вокруг центра больше

Данные о зарплатах: размах

Company	Salary ($mln)	Company	Salary ($mln)
Chrysler	1	Goodyear	13
Teledyne	3	Bristol-Myers	16
Apple Computers	4	Honeywell	16
Hillshire Brands	6	Wal-Mart	21
Marriot			21
Sherwin-W			21
Bank of America	12	Coca-Cola	22
General Mills	12	Philip Morris	25
Delta Airlines	13	Exxon	35
DuPont	13	Motorola	47
Whirlpool	13	CBS	60

Размах равен 59 $mln

Размах

Функции в MS Excel:

- МИН(), МИНА(), МИНЕСЛИ()
- МАКС(), МАКСА(), МАКСЕСЛИ()

Квартили

Медина делит данные на 2 равные части:

- Половина (50%) наблюдений лежит ниже медианы
- Половина (50%) наблюдений лежит выше медианы

Пример:

1 1 3 3 **5** 7 7 9 9

n = 9, медиана – 5^{oe} число= 5

Квартили

Аналогичным образом модно разбить данные на **4** равные части. Границы называются **квартилями**

- Первая, или нижняя, квартиль - это значение с порядковым номером 0.25(n+1):

  25% наблюдений лежат ниже нее, 75 % - выше

- Третья, или верхняя, квартиль - это значение с порядковым номером 0.75(n+1):

  75% наблюдений лежат ниже нее, 25 % - выше

- Вторая квартиль – это? и есть **медиана**

Квартили - пример

1 1 | 3 3 **5** 7 7 | 9 9

n = 9, медиана = 5

- Нижняя квартиль = (n+1)/4 = 2.5, то есть посередине между 1 и 3, значит 2
- Верхняя квартиль = 3(n+1)/4 – 7.5, то есть посередине между 7 и 9, значит 8

Межквартильный размах

Межквартильный размах - разница между третьей и первой квартилями

Inter-quartile range (IQR) = Q3-Q1

1 1 | 3 3 **5** 7 7 | 9 9

Q1= 2, Q3= 8

IQR = 8 – 2 =6

Сравните!

1 1 | 5 5 **5** 5 5 | 9 9

Среднее (5), Медиана (5), Размах (8) те же, но IQR другой (7-3=4 vs 8-2=6)

Данные о зарплатах: квартили

	Company	Salary ($mln)		Company	Salary ($mln)
1	Chrysler	1	12	Goodyear	13
2	Teledyne	3	13	Bristol-Myers	16
3	Apple Computers	4	14	Honeywell	16
4	Hillshire Brands	6	15	Wal-Mart	21
5	Marriot	9	16	AT&T	21
6	Sherwin-Williams	11	17	Boeing	21
7	Bank of America	12	18	Coca-Cola	22
8	General Mills	12	19	Philip Morris	25
9	Delta Airlines	13	20	Exxon	35
10	DuPont	13	21	Motorola	47
11	Whirlpool	13	22	CBS	60

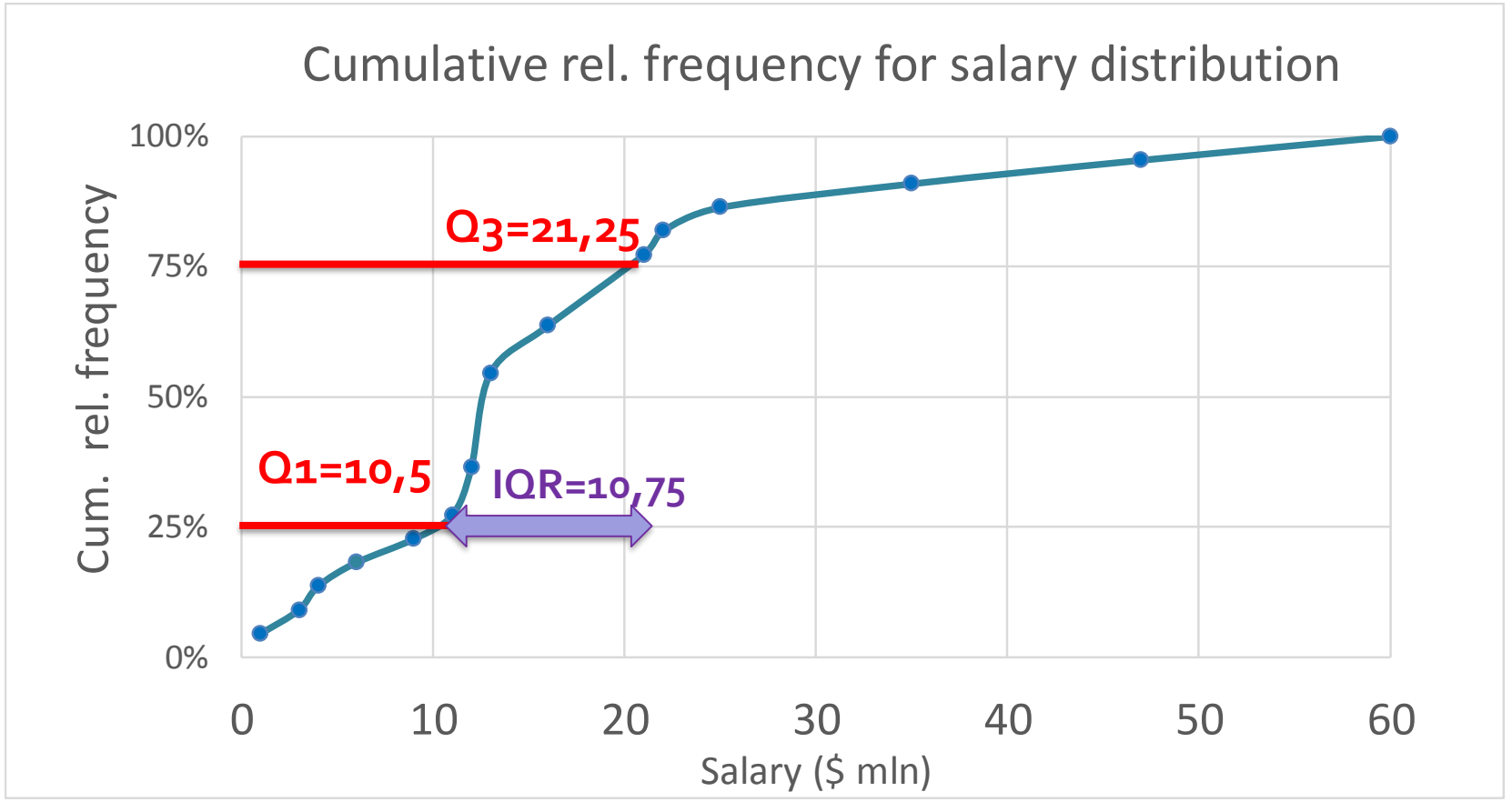
Q1=10,5

Q2=13

Q3=21,25

IQR= 10,75

График накопленных частот



Итог из 5 чисел

Итог из 5 чисел:

{Xmin, Q1, Q2,Q3, Xmax}

Для набора данных о зарплатах:

{1, 10.5, 13,21.25, 60}

Квартили

Функции в MS Excel:

- КВАРТИЛЬ.ВКЛ(), КВАРТИЛЬ.ИСКЛ()
- В старых версиях: КВАРТИЛЬ(), ПЕРСЕНТИЛЬ()

Выбросы

"Малые" выбросы:

Значения, лежащие ниже **Q1-1,5 IQR**

"Большие" выбросы:

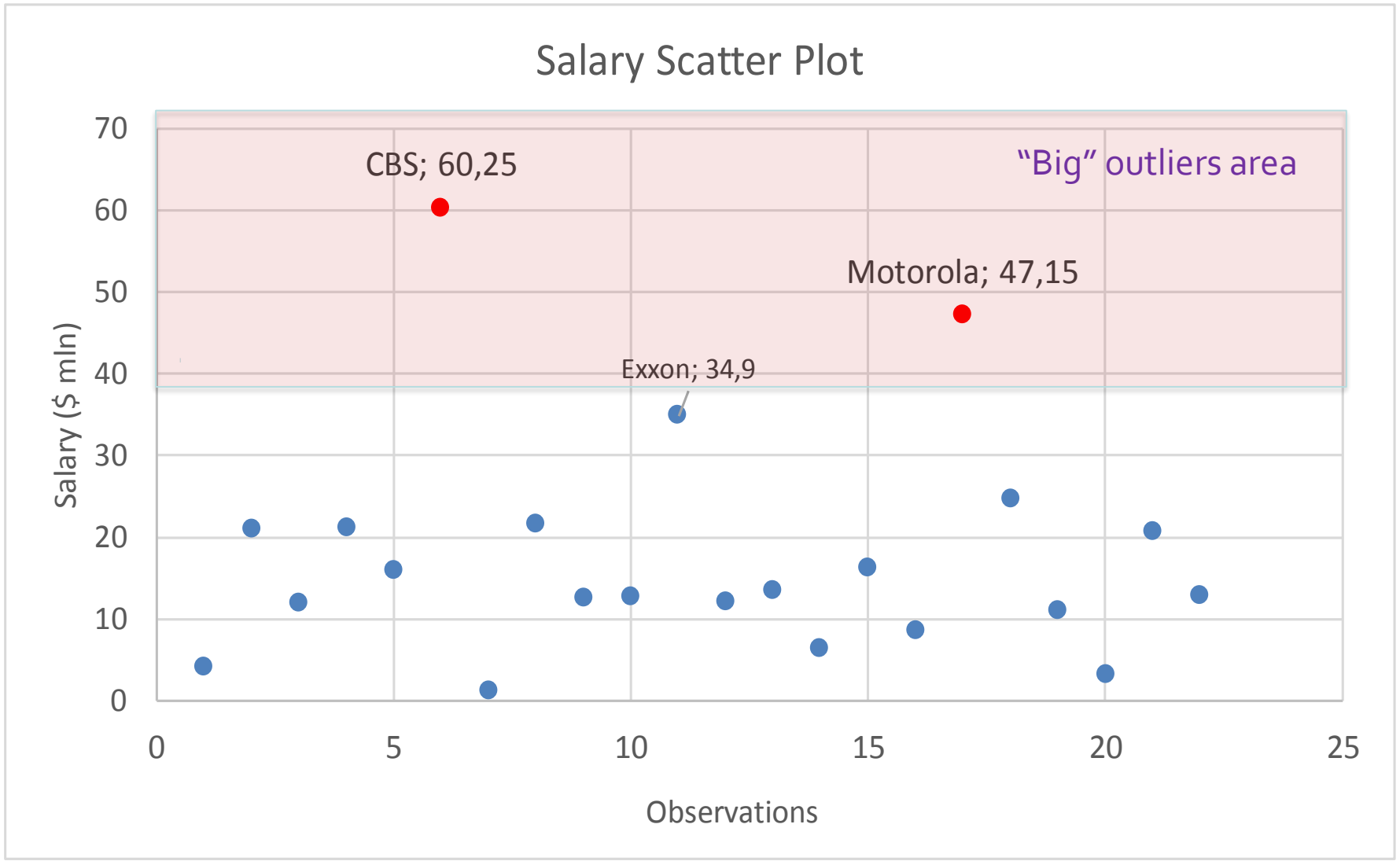
Значения, лежащие выше **Q3+1,5 IQR**

Для данных о зарплатах

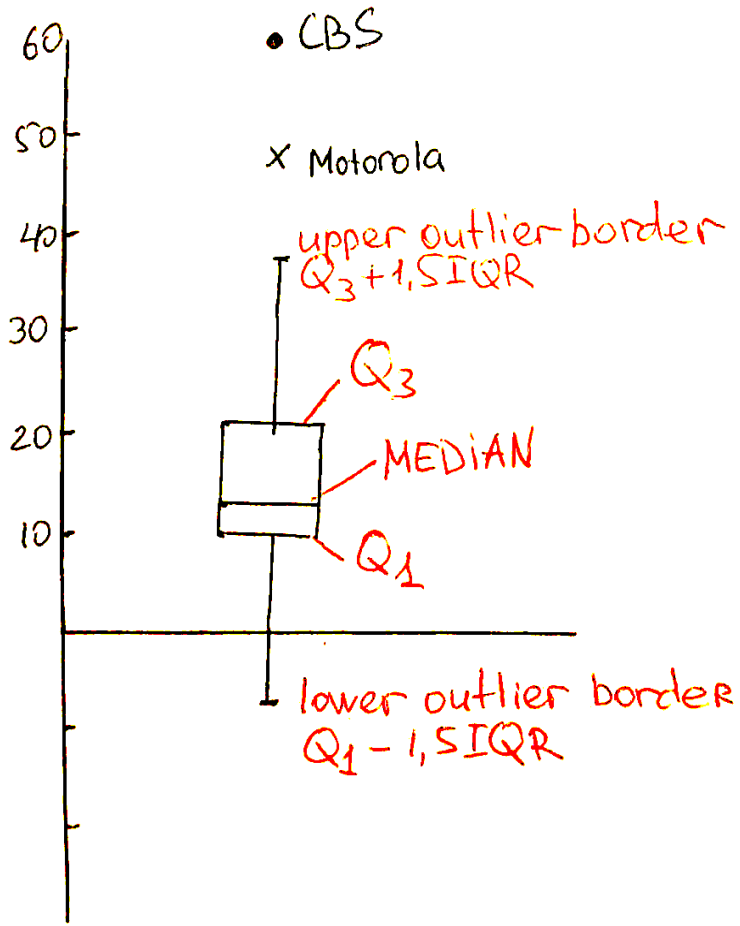
Q1-1,5 IQR=10,5-1,5*10,75=-5,625

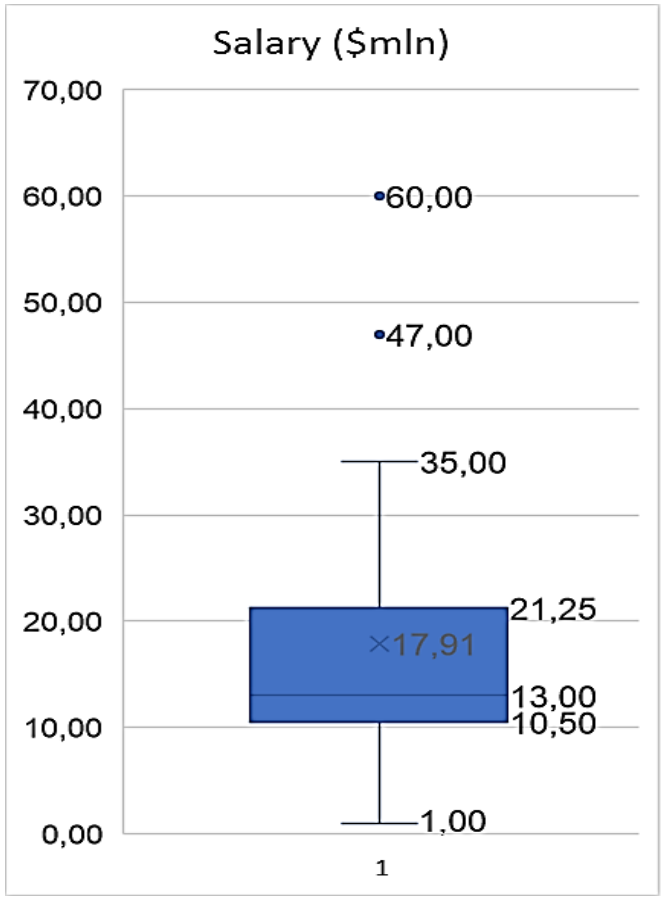
Q3+1,5 IQR=21,25+1,5*10,75=37,375

Выбросы


Salary Scatter Plot
"Big" outliers area
CBS; 60,25
Motorola; 47,15
Exxon; 34,9
Salary ($ mln)
Observations
0
10
20
30
40
50
60
70
0
5
10
15
20
25

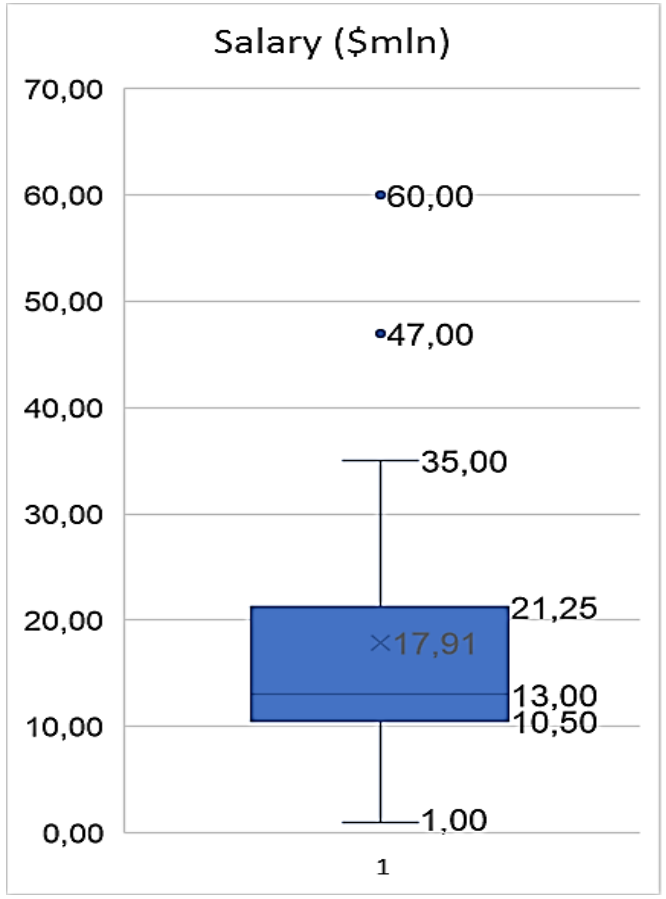
Ящик с усами (box-whiskers plot)

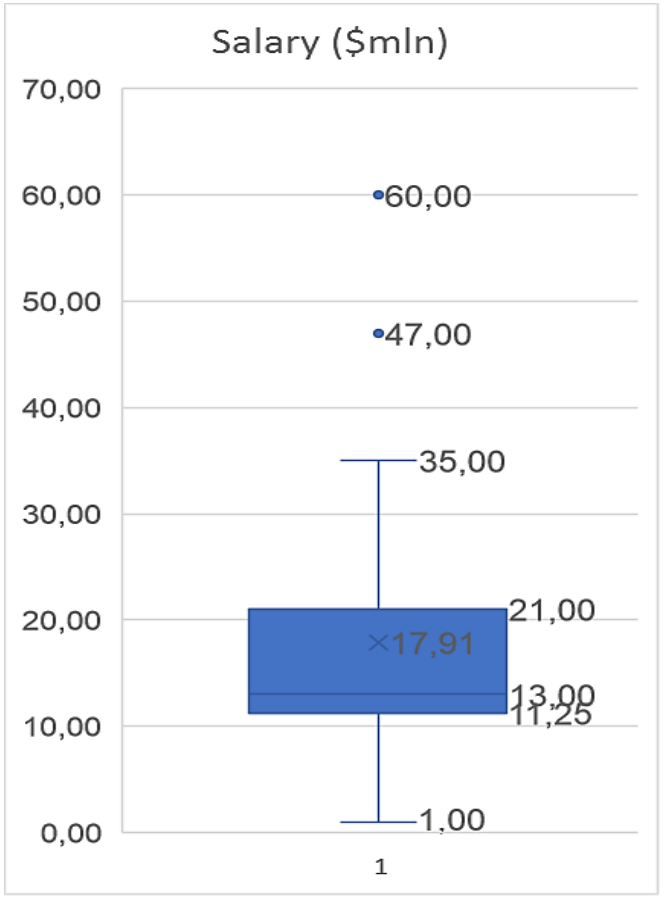




Выбросы сверху – положительная асимметрия

Ящик с усами (box-whiskers plot)





Слева – эксклюзивная медиана

Справа – инклюзивная медиана

Стандартное отклонение

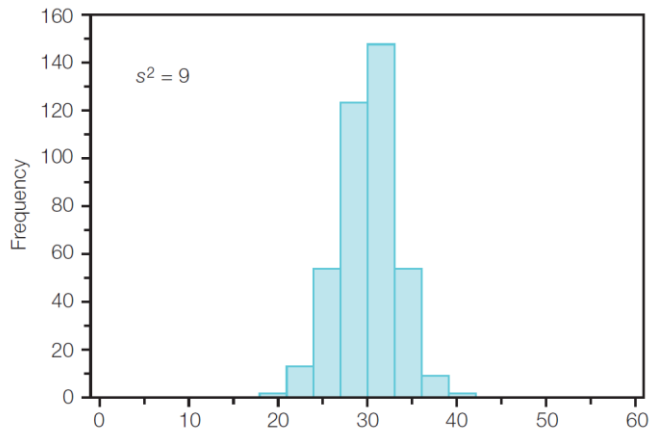
Стандартное отклонение – это мера разброса, которая использует все данные.

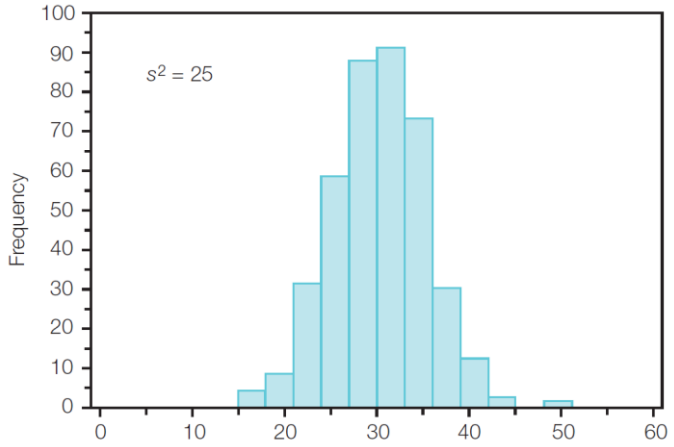
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

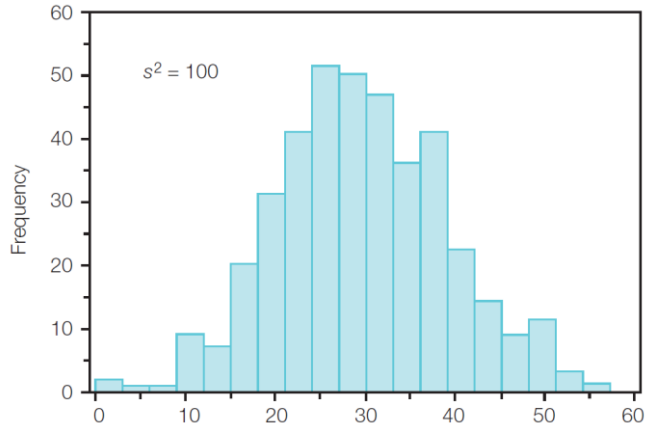
Дисперсия

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

Чем больше разброс, тем больше s







Стандартное отклонение

Функции в MS Excel:

- СУММ(), СУММКВ(), СУММКВРАЗН(), СРЗНАЧ()
- СТАНД.ОТКЛОН.В()

Данные о зарплатах: стандартное отклонение

Company	Salary ($mln)		Company	Salary ($mln)
Chrysler	1		Goodyear	13
Teledyne	3		Bristol-Myers	16
Apple Computers	4		Honeywell	16
Hillshire Brands	6		Wal-Mart	21
Marriot				21
Sherwin-V				21
Bank of America	12		Coca-Cola	22
General Mills	12		Philip Morris	25
Delta Airlines	13		Exxon	35
DuPont	13		Motorola	47
Whirlpool	13		CBS	60

s= 14,06 $mln

Данные о зарплатах: описательная статистика

Salary ($mln)	
Среднее	17,90909
Стандартная ошибка	2,997047
Медиана	13
Мода	13
Стандартное отклоне	14,0574
Дисперсия выборки	197,6104
Эксцесс	3,249631
Асимметричность	1,701409
Интервал	59
Минимум	1
Максимум	60
Сумма	394
Счет	22

Визуализация качественных признаков – на семинаре

- Сводные таблицы и диаграммы
- Таблицы сопряженности
- Иерархия признаков