Описательная статистика в Excel

1 Надстройка «Анализ данных»

Средство Excel 2013 **Анализ данных** является надстройкой, которую перед первым использованием необходимо установить. Выполним для этого последовательность действий: **Файл → Параметры → Надстройки → Имя → Пакет анализа → Перейти →** откроется меню **Надстройки → Доступные настройки →** выберите **Пакет анализа → ОК**. В результате на вкладке **Данные** в группе **Анализ** появится пиктограмма, которая обеспечит доступ к **Инструментам анализа.**

Данная надстройка предназначена для выполнения базовых операций статистического анализа данных. Используется она и при проведении инженерных расчетов. При запуске надстройки открывается диалоговое окно, в котором можно выбрать необходимый инструмент анализа (рис. 1а и 1.б).

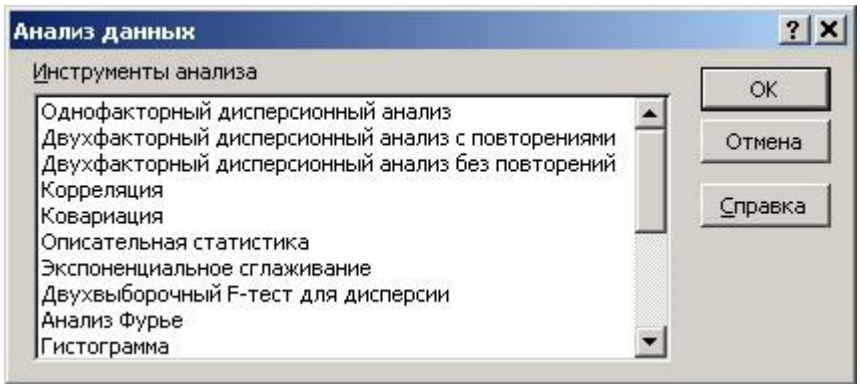

Рис 1а. Диалоговое окно Анализ данных

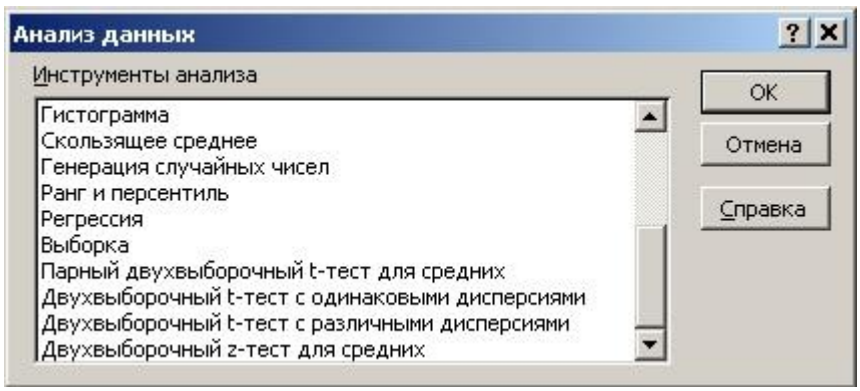

Рис 1б. Диалоговое окно Анализ данных (продолжение)

Всего инструментов анализа в настоящее время 19. По назначению их можно разбить на несколько групп:

- инструменты описательной статистики и построение гистограмм;
- генератор случайных чисел для различных распределений и инструменты для создания случайной выборки;
- инструменты для подсчета рангов и персентилей;
- инструменты для вычисления корреляций и матриц ковариации;
- скользящие средние и инструменты однопараметрического экспоненциального сглаживания;
- инструменты множественной линейной регрессии;
- инструменты дисперсионного анализа, включая однофакторный, двухфакторный без повторений и двухфакторный сбалансированный с повторениями;
- парные двухвыборочные t-тесты с одинаковой и разной дисперсией;
- z-тест для средних и F-тест для дисперсий;
- быстрое преобразование Фурье.

2 Описательная статистика

Первичный анализ скалярных экспериментальных данных начинается с вычисления описательных статистик. Добавив к этому графические характеристики, получим некоторые основания для выводов о характере распределения данных исследуемой совокупности. К тому же, базовый анализ дает основу для дальнейшего проведения более сложного анализа данных.

Из множества инструментов надстройки **Анализ данных** будем использовать **Описательную статистику,** для получения числовых характеристик, и **Гистограмму** — для графических. Заметим, что наряду с этим можно использовать также встроенные **Статистические функции**, которые дублируют возможности надстройки.

Рассмотрим работу с описательной статистикой на примере.

Пример. Имеются некоторые данные о стоимости новогодних туров (рис. 2). Каждый из столбцов можно рассматривать как отдельный признак или переменную. Требуется провести анализ данных о продолжительности туров.

Исходные данные содержат несколько переменных, характеризующих тур. **Название фирмы**, **Страна**, **Транспорт** — качественные переменные, которые относятся к номинальной шкале. **Отель** — качественная переменная, которую можно отнести к порядковой шкале, так как количество звездочек отражает уровень обслуживания в отеле. **Количество дней** и **Стоимость** — количественные данные, которые относятся к метрической шкале.

	A	B	C	D	E	F	G
1	Новогодние туры						
2	№	Название фирмы	Страна	Кол-во дней	Стоимость	Отель	Транспорт
3	1	Нева	Италия	11	447 €	***	Поезд
4	2	Нева	Германия	7	374 €	****	Самолет
5	3	Нева	Польша	7	199 €	***	Поезд
6	4	Нева	Франция	10	581 €	****	Автобус
7	5	Нева	Швейцария	8	1 240 €	***	Самолет
8	6	Одиссея	Норвегия	12	645 €	***	Паром
9	7	Одиссея	Испания	8	796 €	****	Самолет
10	8	Одиссея	Италия	12	430 €	****	Поезд
11	9	Одиссея	Польша	9	265 €	****	Поезд
12	10	Олимпика	Финляндия	3	175 €	***	Автобус
13	11	Олимпика	Швеция	5	800 €	***	Поезд
14	12	Олимпика	Мальдивы	14	2 890 €	*****	Самолет
15	13	Олимпика	Германия	12	700 €	*****	Поезд
16	14	Олимпика	Португалия	8	1 460 €	*****	Самолет

Рис 2. Таблица исходных данных

Вычислим основные описательные статистики для переменной **Количество дней**, которая является числовой переменной, принимающей дискретные значения. Для этого используем инструмент **Описательная статистика**, входящий в **Пакет анализа**.

Для перехода к описательной статистике выполните: **Данные → Анализ → Анализ данных → Описательная статистика → ОК**. В открывшемся диалоговом окне **Описательная статистика** (рис.3) укажите **Входной интервал**, диапазон **D2:D16**, выберите **Группирование по столбцам**, установите **Метки в первой строке**, так как входной интервал содержит наименование столбца. Для **Выходного интервала** достаточно указать одну, первую, ячейку на текущем листе, как альтернативу можно выбрать **Новый рабочий лист** или **Новую рабочую**

книгу. И наконец, укажите хотя бы одну из выводимых статистик: **Итоговая статистика**, **Уровень надежности**, **К-ый наименьший**, **К-ый наибольший**.

В большинстве случаев достаточно выбрать **Итоговую статистику**, которая рассчитывает основные числовые характеристики исследуемой совокупности. Три последних значения рассчитывают, только когда они действительно нужны.

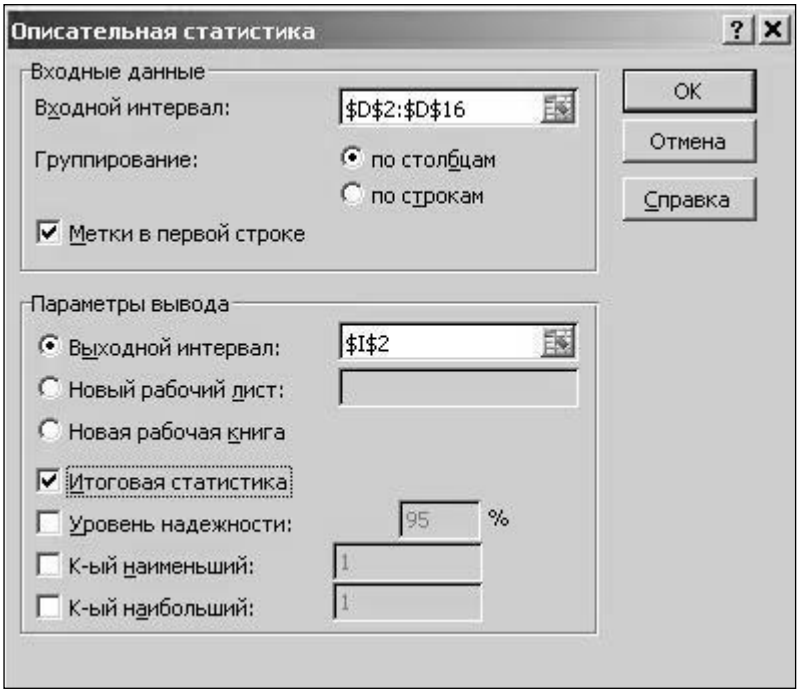

Рис 3. Диалоговое окно ***Описательной статистики***

Описательная статистика вычисляет 16 значений, из них 13 относятся к **Итоговой статистике,** еще 3 определяют доверительный интервал и два выборочных значения.

Отметим главное — **Описательная статистка** надстройки **Анализ данных** предназначена для вычислений статистических характеристик, или статистик, одномерной *выборки* или нескольких выборок.

В литературе по статистике часто используют термин «генеральная совокупность». Обычно имеется в виду, что это множество всех доступных для наблюдения данных в противоположность «выборки» — которая подразумевает, что исследуется лишь часть данных выбранных из генеральной совокупности (может быть с помощью случайного отбора).

Обычно числовые характеристики генеральной совокупности называют *параметрами*, а числовые характеристики выборки — *статистиками*, или

выборочными характеристиками, которые являются оценками параметров генеральной совокупности. Для более полного понимания *выборочного метода* следует обратиться к специальной литературе.

Результаты расчетов **Итоговой статистики** для переменной **Количество дней** приведены на рис.4. На этом же рисунке приведены альтернативные расчеты этих числовых характеристик с использованием встроенных функций категории **Статистические**. Аргументом статистических функций является диапазон исходных данных, в данном случае **D3:D16.**

Таким образом, практически все расчеты **Описательной статистики** дублируются **Статистическими** функциями. Остальные характеристики можно посчитать, используя формулы. Для того чтобы на рабочем листе Excel отобразились не результаты, а формулы, следует выполнить: **Формулы → Зависимости формул → Показать формулы.**

Отметим некоторое отличие в применении инструментов **Анализа данных** и использованием статистических функций. При изменении значений исходных данных формулы пересчитываются, в то время как результаты, полученные с помощью инструментов **Анализа данных** не изменяются. Чтобы обновить результаты, потребуется вызывать **Анализ данных** снова.

I	J	K
Кол-во дней		**Статистические функции**
Среднее	9	=СРЗНАЧ(D3:D16)
Стандартная ошибка	0,81199794294115	
Медиана	8,5	=МЕДИАНА(D3:D16)
Мода	8	=МОДА(D3:D16)
Стандартное отклонение	3,038218101251	=СТАНДОТКЛОН(D3:D16)
Дисперсия выборки	9,23076923076923	=ДИСП(D3:D16)
Эксцесс	-0,290530303030302	=ЭКСЦЕСС(D3:D16)
Асимметричность	-0,268797907013457	=СКОС(D3:D16)
Интервал	11	
Минимум	3	=МИН(D3:D16)
Максимум	14	=МАКС(D3:D16)
Сумма	126	=СУММ(D3:D16)
Счет	14	=СЧЁТ(D3:D16)

Рис 4. Итоговая статистика и Статистические функции

Числовые характеристики **Итоговой статистики** описывают средние, вариацию и форму распределения, всего 13 параметров:

– *среднее*, или выборочное среднее, вычисляется как среднее арифметическое наблюдаемых значений выборки;

– *медиана* определяется как значение, находящееся в середине распределения, полученного из исходного путем упорядочивания по возрастанию;

– *мода* равна наиболее часто встречающемуся значению.

Кроме того, выделяют две величины, характеризующие изменчивость, или разброс, значений распределения относительно среднего:

– *дисперсию выборки*, или выборочную дисперсию, равную сумме квадратов отклонений каждого значения от среднего, деленной на $(N-1)$, где N — число значений в распределении, или объем выборки;

– *стандартное отклонение*, или выборочное среднеквадратическое отклонение, равное квадратному корню из выборочной дисперсии.

Дополнительными мерами изменчивости являются 4 простые характеристики, отражающие границы распределения данных и его размах:

– *минимум* равен наименьшему из выборочных значений;

– *максимум* равен наибольшему из выборочных значений;

– *интервал* составляет разность между максимумом и минимумом, этот параметр называют также *размахом*.

Если набор данных рассматривается как множество независимых реализаций случайной величины, то возникает вопрос, что можно сказать о функции распределения этой величины на основании выборки. Очень часто распределение оказывается нормальным или близким к нему.

Для отражения близости формы распределения к нормальному виду существует две основные характеристики:

– *эксцесс*, или выборочный коэффициент эксцесса, который является мерой «сглаженности» распределения;

– *асимметричность*, или выборочный коэффициент асимметрии, показывает, в какую сторону относительно среднего сдвинуто большинство значений выборки.

И наконец, *сумма* равна сумме всех выборочных значений, *счет* вычисляет объем выборки, *стандартная ошибка* равна выборочному стандартному отклонению, деленному на квадратный корень из объема выборки.

При необходимости можно вычислить три дополнительные характеристики (рис. 5). Результаты расчетов этих характеристик приведены на рис. 6.

К-ый наибольший выдает К-тое выборочное значение, если бы выборка была отсортирована по убыванию. В рассматриваемом примере сортировка по убыванию имеет вид 14, 12, 12, 12, 11, 10 и т.д., третье значение равно 12. **К-ый наименьший** выдает К-тое выборочное значение, если бы выборка была отсортирована по возрастанию, это значение равно 5.

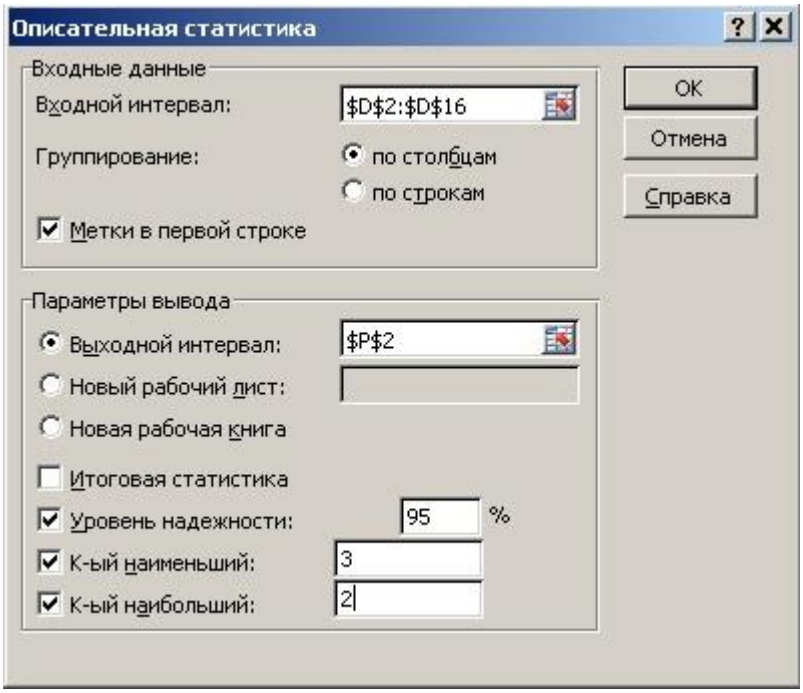

Рис 5. Описательная статистика, дополнительные параметры

P	Q
Кол-во дней	
Наибольший(3)	12
Наименьший(2)	5
Уровень надежности(95,0%)	1,7542149

Рис 6. Результаты расчетов дополнительных параметров

Задав **Уровень надежности**, например 95%, получим значение для построения доверительного интервала для неизвестного математического ожидания

генеральной средней с доверительным уровнем 95%. Доверительный интервал строится как выборочное среднее плюс-минус полученное значение. Обратим внимание, что граница здесь вычисляется с помощью распределения Стьюдента, что требует достаточного количества наблюдений на каждую степень свободы.

Таким образом, к вычислению доверительных интервалов нужно относиться с осторожностью, особенно при малых выборках. Использование функции расчета доверительного интервала без понимания статистического смысла может привести к ошибкам. Начинающим исследователям посоветуем обратиться к специальной литературе.

Например, для рассматриваемого примера, полученный доверительный интервал не несет смыслового содержания.

Итак, на этапе проведения описательной статистики, исследуемый ряд данных может быть как генеральной совокупностью, так и выборкой. Если для генеральной совокупности вычисляются значения параметров распределения, то для выборки находят оценки этих параметров. Рассмотрим ниже подробнее вычисление некоторых числовых характеристик в пакете Excel.

2.1. Дисперсия

Описательная статистика, реализованная в **Пакете анализа**, рассчитывает *выборочные* характеристики. То есть *среднее* здесь — это выборочная оценка математического ожидания генеральной совокупности, из которой извлечена выборка, *дисперсия выборки* — выборочная оценка дисперсии генеральной совокупности, то есть несмещенная оценка, *стандартное отклонение* — оценка среднеквадратического отклонения на основе несмещенной оценки дисперсии. Описательная статистика при вычислении среднего и выборочной дисперсии использует, соответственно, следующие формулы:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N},\ S^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N-1},$$

где N — объем выборки, x_i — i-ый элемент выборки.

Описательная статистика в Excel предназначена для работы именно с выборками. Одна из статистик так и называется, *дисперсия выборки*. Поэтому, если расчеты проводятся непосредственно для генеральной совокупности, для вычисления *дисперсии* нужно использовать специальную встроенную статистическую функцию. В последних версиях Excel 2010/2013 на это различие обратили особое внимание, добавив функции с соответствующими названиями:

- ДИСПР.Г вычисляет дисперсию для генеральной совокупности по формуле (μ — математическое ожидание):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N};$$

- ДИСПР.В оценивает дисперсию генеральной совокупности по выборке, используя формулу:

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}.$$

При этом оставлены еще две дублирующие функции ДИСПР и ДИСП для обеспечения совместимости, чтобы не пришлось переделывать многочисленные расчеты, выполненные в более ранних версиях.

Соответственно, при вычислении *среднеквадратического отклонения*, или *среднего квадратического отклонения* (СКВО), для генеральной совокупности следует использовать статистическую функцию СТАНДОТКЛОН.Г, а для его оценки по выборке нужно использовать СТАНДОТКЛОН.В.

Иногда дисперсию выборки считают по формуле:

$$\tilde{S}^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}.$$

При этом называют S^2 *исправленной выборочной дисперсией*, которая является несмещенной оценкой дисперсии генеральной совокупности, а $\tilde{S}^2$ — *выборочной дисперсией*, которая является смещенной оценкой дисперсии генеральной совокупности.

Если объем выборки велик, то разница между значениями S^2 и $\tilde{S}^2$ не велика. При малых N это различие существенно. Вычислим оба значения для рассмотренного выше примера: ДИСП.В(D3:D16)=ДИСП(D3:D16)=9,23077, а ДИСП.Г(D3:D16)=ДИСПР(D3:D16)=8,57143.

В **Описательной статистике** надстройки **Анализ данных** дисперсия вычисляется с использованием функции ДИСП.В, то есть вычисляется исправленная выборочная дисперсия S^2. Соответственно, для вычисления среднеквадратического отклонения используется функция СТАНДОТКЛОН.В.

2.3. Мода и медиана

Мода — наиболее часто встречающееся значение во множестве наблюдений. Если такое значение только одно, распределение называется *унимодальным*, а если несколько — *полимодальным*. Изучаемая случайная величина может не иметь моды, в этом случае Excel выдает сообщение об ошибке #Н/Д.

Для вычисления моды в Excel есть несколько встроенных функций:

- МОДА.ОДН и МОД вычисляют моду для унимодального распределения, и выдают только одно значение моды, даже если распределение полимодально;
- МОДА.НСК вычисляет моду для полимодального распределения и возвращает вертикальный массив наиболее часто встречающихся значений в указанном диапазоне, то есть несколько значений моды.

Заметим, что при вычислении моды с помощью **Описательной статистики** используется функция МОДА.ОДН, то есть выдаются только одно значение моды, меньшее по значению. Так в рассмотренном выше примере (см. рис. 2) расчетное значение моды равно 8 (см. рис. 4) , хотя числа 12 и 8 встречаются по три раза.

При вычислении моды рекомендуется сначала использовать функцию МОДА.НСК. Применение этой функции имеет свои особенности. Формулу **=МОДА.НСК(диапазон)** необходимо ввести как формулу массива.

Замечание. Ряд функций в Excel необходимо вводить как *формулу массива*, так как они возвращают несколько значений, которые выводится в диапазон ячеек,

или массив. Для ввода формулы массива, выделите диапазон, в который будет сохранен результат, введите формулу и нажмите комбинацию клавиш **Ctrl+Shift+Enter**.

Так как, заранее не известно, имеет ли исследуемая совокупность моду, а если имеет, то одну или несколько, то диапазон для вывода может содержать несколько ячеек. Найдем моду для вышеуказанного примера. Выделите диапазон **F3:F6** и введите формулу **=МОДА.НСК(D3:D16)**, где диапазон **D3:D16** задает исходные данные. Эта формула отобразится также в строке формул (рис. 7).

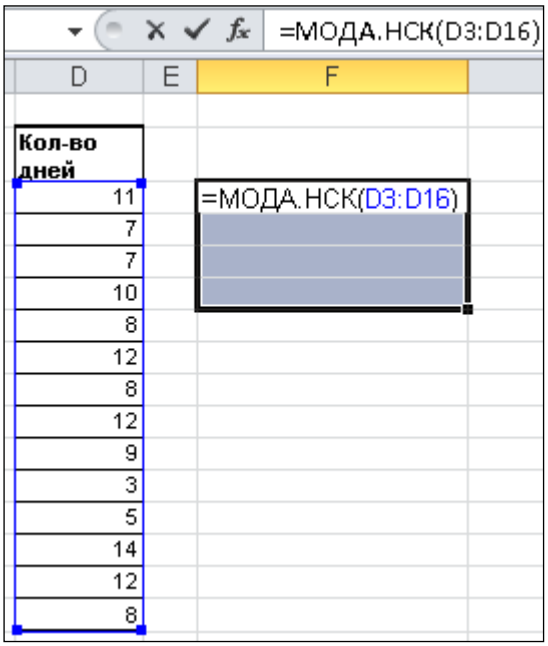

Рис 7. Вычисление моды

Теперь нажмите одновременно комбинацию клавиш **Ctrl+Shift+Enter**, формула введется во все выделенные ячейки как формула массива. Отобразятся два значения моды, в остальные ячейки появится сообщение «Нет данных» (рис. 8).

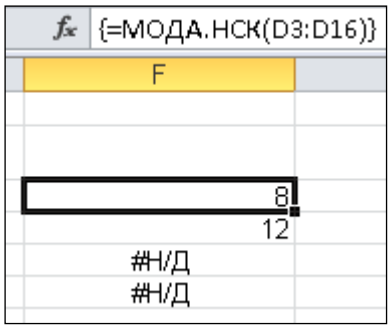

Рис 8. Результаты вычисления моды

Если известны все значения признака, то для нахождения моды не требуется проводить дополнительные расчеты, значением моды является конкретное значение признака. Расчет моды для не сгруппированных данных состоит в определении наиболее часто встречающегося значения. Для дискретного ряда распределения мода соответствует значению признака имеющего наибольшую частоту. Моду для интервального ряда распределения определяются по специальной формуле, в этом случае ее значение вычисляется приближенно.

Медиана — такое значение признака, которое делит ранжированный ряд на две равные части, со значениями признака меньше медианы и со значениями признака больше медианы. Для нахождения медианы исходный ряд предварительно упорядочивают по возрастанию (ранжируют).

Для вычисления медианы в Excel есть встроенная функция **МЕДИАНА(диапазон)**, причем исходный ряд не требуется предварительно упорядочивать.

Если известны все значения признака, ряд не сгруппирован, то для нахождения медианы не требуется проводить дополнительные расчеты. Все сводится к нахождению порядкового номера медианы. Если данные содержат нечетное число значений, то медиана есть центральное значение. Если же данные содержат четное число значений, то медиана находится как среднее арифметическое двух центральных значений. Значением моды является конкретное значение признака.

Для интервальных рядов медиана рассчитывается по специальной формуле.

Мода и медиана называются структурными средними. Кроме того, часто используют понятие «пяти базовых показателей», в которые входят — минимальное значение, 1 квартиль Q_1, 2 квартиль Q_2, 3 квартиль Q_3 и максимальное значение. Квартили — это значения признака, делящие ранжированную совокупность на четыре равновеликие части, 2 квартиль совпадает с медианой.

2.4. Асимметрия и эксцесс

Асимметрия и эксцесс являются мерами формы распределения данных. *Асимметрия* является мерой несимметричности данных. *Эксцесс* характеризует

относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением.

При положительной асимметрии, As>0, значения распределения скучены в области малых значений, при этом мода будет меньше медианы, а медиана меньше среднего. То есть более половины значений будут меньше среднего. При отрицательной асимметрии, As<0 , значения распределения скучены в области высоких значений. В этом случае среднее будет меньше медианы, а медиана меньше моды. То есть более половины значений будут больше среднего.

Положительный эксцесс характерен для остроконечного распределения, а отрицательный — для относительно сглаженного распределения.

Для нормального распределения асимметрия и эксцесс равны нулю. Это свойство нормального распределения часто используют для приближенной проверки нормальности исследуемой совокупности данных. Дополнительно к этому строят для наглядности гистограмму. Визуальное представление гистограммы называют формой распределения.

Средство **Генерация случайных чисел** можно рассматривать как для организации случайных чисел, так и для иллюстративного поведения случайных величин с заданным законом распределения.

Воспользуемся средством **Генерация случайных чисел**, которое позволяет сформировать совокупность данных, имеющих заданный закон распределения. Сформирует нормальное и равномерное распределение, каждое объемом 20 0 , вычислим для них описательную статистику и воспользуемся инструментом