Определение среднего значения, вариации и формы распределения. Описательные статистики

Способы представления числовых и категорийных данных в виде таблиц и диаграмм являются существенной, но не основной частью анализа данных. Ведущая роль принадлежит методам исследования числовых данных и их свойств. В этой заметке рассмотрены способы определения среднего значения, вариации и формы распределения генеральной совокупности.[1]

В большинстве случаев данные концентрируются вокруг некоей центральной точки. Таким образом, чтобы описать любой набор данных, достаточно указать средне значение. Рассмотрим последовательно три числовые характеристики, которые используются для оценки среднего значения распределения: среднее арифметическое, медиана и мода.

Среднее арифметическое

Среднее арифметическое (часто называемое просто средним) — наиболее распространенная оценка среднего значения распределения. Она является результатом деления суммы всех наблюдаемых числовых величин на их количество. Для выборки, состоящей из чисел $X_1, X_2, \ldots, X_n$, выборочное среднее (обозначаемое символом $\bar{X}$) равно $\bar{X} = (X_1 + X_2 + \ldots + X_n) / n$, *или*

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

где $\bar{X}$ — выборочное среднее, n — объем выборки, X_i – i-й элемент выборки.

Рассмотрим вычисление среднего арифметического значения пятилетней среднегодовой доходности 15 взаимных фондов с очень высоким уровнем риска (рис. 1).

Фонд	Доходность
Amer. Century GiftTrust Inv.	$X_1 = -2,8$
AXP Strategy Aggressive A	$X_2 = 5,5$
Berger Small Company Growth Inv	$X_3 = 8,3$
Consulting Group Small Cap Growth	$X_4 = 4,3$
Fidelity Aggressive Growth	$X_5 = 5,9$
Invesco Growth Inv	$X_6 = -0,7$
Janus Enterprise	$X_7 = 6,5$
Janus Venture	$X_8 = 9,8$
John Hancock Small Cap Growth A	$X_9 = 7,6$
MS Mid Cap Equity Tr. B	$X_{10} = 9,6$
PBHG Growth	$X_{11} = -1,2$
Putnam OTC Emerging Growth A	$X_{12} = -6,1$
RS Emerging Growth A	$X_{13} = 18,5$
Rydex OTC Inv	$X_{14} = 13,1$
Van Kampen Aggressive Growth A	$X_{15} = 12,9$

Рис. 1. Среднегодовая доходность 15 взаимных фондов с очень высоким уровнем риска

Выборочное среднее вычисляется следующим образом:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{-2,8 + 5,5 + \ldots + 12,9}{15} = \frac{91,2}{15} = 6,08$$

Это хороший доход, особенно по сравнению с 3–4% дохода, который получили вкладчики банков или кредитных союзов за тот же период времени. Если упорядочить значения доходности, то легко заметить, что восемь фондов имеют доходность выше, а семь — ниже среднего значения. Среднее

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 178–209

арифметическое играет роль точки равновесия, так что фонды с низкими доходами уравновешивают фонды с высокими доходами. В вычислении среднего задействованы все элементы выборки. *Ни одна из других оценок среднего значения распределения не обладает этим свойством.*

Когда следует вычислять среднее арифметическое. Поскольку среднее арифметическое зависит от всех элементов выборки, наличие экстремальных значений значительно влияет на результат. В таких ситуациях среднее арифметическое может исказить смысл числовых данных. Следовательно, описывая набор данных, содержащий экстремальные значения, необходимо указывать медиану либо среднее арифметическое и медиану. Например, если удалить из выборки доходность фонда RS Emerging Growth, выборочное среднее доходности 14 фондов уменьшится почти на 1% и составит 5,19%.

Медиана

Медиана представляет собой срединное значение упорядоченного массива чисел. Если массив не содержит повторяющихся чисел, то половина его элементов окажется меньше, а половина — больше медианы. Если выборка содержит экстремальные значения, для оценки среднего значения лучше использовать не среднее арифметическое, а медиану. Чтобы вычислить медиану выборки, ее сначала необходимо упорядочить.

$$\text{Медиана} = \frac{n+1}{2}\text{-й элемент упорядоченного массива}$$

Эта формула неоднозначна. Ее результат зависит от четности или нечетности числа *n*:

- Если выборка содержит нечетное количество элементов, медиана равна *(n+1)/2*-му элементу.
- Если выборка содержит четное количество элементов, медиана лежит между двумя средними элементами выборки и равна среднему арифметическому, вычисленному по этим двум элементам.

Чтобы вычислить медиану выборки, содержащей данные о доходности 15 взаимных фондов с очень высокий уровнем риска, сначала необходимо упорядочить исходные данные (рис. 2). Тогда медиана будет напротив номера среднего элемента выборки; в нашем примере №8. В Excel есть специальная функция =МЕДИАНА(), которая работает и с неупорядоченными массивами тоже.

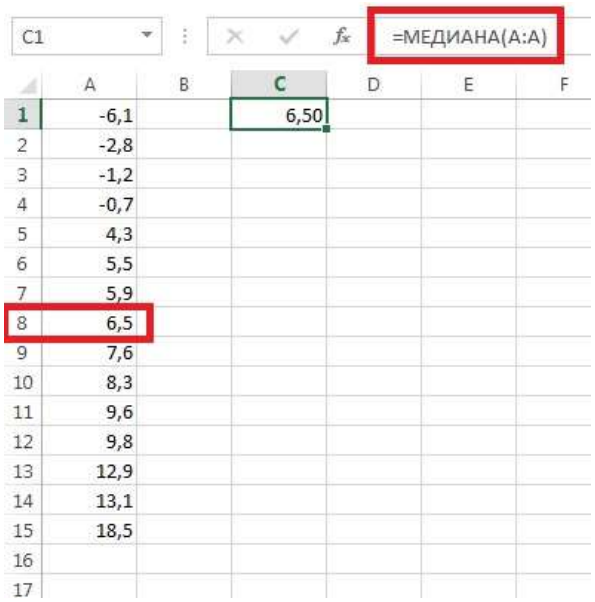


Рис. 2. Медиана 15 фондов

Таким образом, медиана равна 6,5. Это означает, что доходность одной половины фондов с очень высоким уровнем риска не превышает 6,5, а доходность второй половины — превышает ее. Обратите внимание на то, что медиана, равная 6,5, ненамного больше среднего значения, равного 6,08.

Если удалить из выборки доходность фонда RS Emerging Growth, то медиана оставшихся 14 фондов уменьшится до 6,2%, то есть не так значительно, как среднее арифметическое (рис. 3).

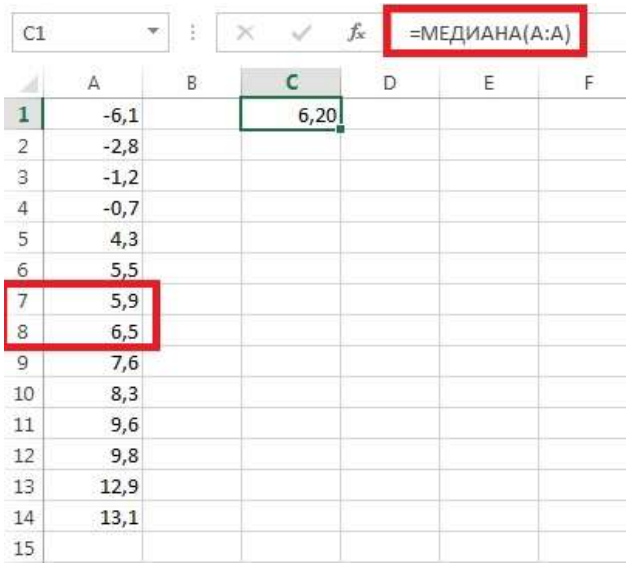


Рис. 3. Медиана 14 фондов

Мода

Термин был впервые введен Пирсоном в 1894 г. Мода — это число, которое чаще других встречается в выборке (наиболее модное). Мода хорошо описывает, например, типичную реакцию водителей на сигнал светофора о прекращении движения. Классический пример использования моды — выбор размера выпускаемой партии обуви или цвета обоев. Если распределение имеет несколько мод, то говорят, что оно мультимодально или многомодально (имеет два или более «пика»). Мультимодальность распределения дает важную информацию о природе исследуемой переменной. Например, в социологических опросах, если переменная представляет собой предпочтение или отношение к чему-то, то мультимодальность может означать, что существуют несколько определенно различных мнений. Мультимодальность также служит индикатором того, что выборка не является однородной и наблюдения, возможно, порождены двумя или более «наложенными» распределениями. В отличие от среднего арифметического, выбросы на моду не влияют. Для непрерывно распределенных случайных величин, например, для показателей среднегодовой доходности взаимных фондов, мода иногда вообще не существует (или не имеет смысла). Поскольку эти показатели могут принимать самые разные значения, повторяющиеся величины встречаются крайне редко.

Квартили

Квартили — это показатели, которые чаще всего используются для оценки распределения данных при описании свойств больших числовых выборок. В то время как медиана разделяет упорядоченный массив пополам (50% элементов массива меньше медианы и 50% — больше), квартили разбивают упорядоченный набор данных на четыре части. Величины Q_1, медиана и Q_3 являются 25-м, 50-м и 75-м перцентилем соответственно. Первый квартиль Q_1 — это число, разделяющее выборку на две части: 25% элементов меньше, а 75% — больше первого квартиля.

$Q_1 = \frac{n+1}{4}$-й элемент упорядоченного массива

Третий квартиль Q_3 — это число, разделяющее выборку также на две части: 75% элементов меньше, а 25% — больше третьего квартиля.

$Q_3 = \frac{3(n+1)}{4}$-й элемент упорядоченного массива

Для расчета квартилей в версиях Excel до 2007 г. использовалась функция =КВАРТИЛЬ(массив;часть). Начиная с версии Excel2010 применяются две функции:[2]

- =КВАРТИЛЬ.ВКЛ(массив;часть)
- =КВАРТИЛЬ.ИСКЛ(массив;часть)

[2] Функция КВАРТИЛЬ оставлена для совмещения с более ранними версиями Excel

Эти две функции дают немного различные значения (рис. 4). Например, при вычислении квартилей выборки, содержащей данные о среднегодовой доходности 15 взаимных фондов с очень высоким уровнем риска Q_1 = 1,8 или –0,7 для КВАРТИЛЬ.ВКЛ и КВАРТИЛЬ.ИСКЛ, соответственно. Кстати функция КВАРТИЛЬ, использовавшаяся ранее соответствует современной функции КВАРТИЛЬ.ВКЛ. Для расчета квартилей в Excel с помощью вышеприведенных формул массив данных можно не упорядочивать.

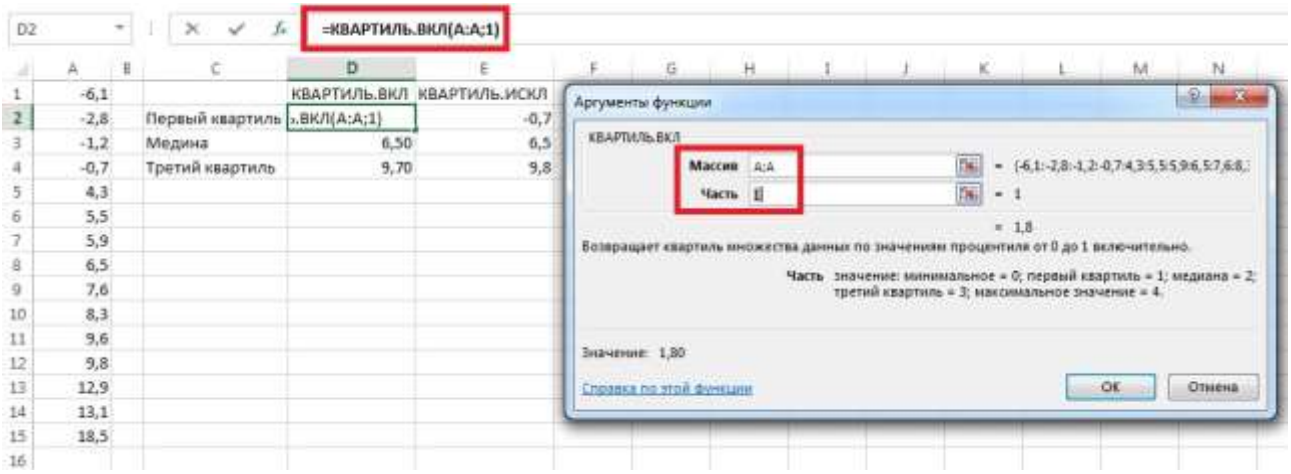


Рис. 4. Вычисление квартилей в Excel

Подчеркнем еще раз. Excel умеет рассчитывать квартили для одномерного *дискретного ряда*, содержащего значения случайной величины. Расчет квартилей для распределения на основе частот приведен ниже в разделе *Вычисление описательных статистик для распределения на основе частот*.

Среднее геометрическое

В отличие от среднего арифметического среднее геометрическое позволяет оценить степень изменения переменной с течением времени. Среднее геометрическое — это корень *n*-й степени из произведения *n* величин (в Excel используется функция =СРГЕОМ):

$\overline{X}_G = (X_1 * X_2 * \ldots * X_n)^{1/n}$

Похожий параметр – среднее геометрическое значение нормы прибыли – определяется формулой:

$\overline{R}_G = [(1 + R_1) * (1 + R_2) * \ldots * (1 + R_n)]^{1/n} - 1,$

где R_i – норма прибыли за *i*-й период времени

Например, предположим, что объем вложенных средств в исходный момент времени равен 100 000 долл. К концу первого года он падает до уровня 50 000 долл., а к концу второго года восстанавливается до исходной отметки 100 000 долл. Норма прибыли этой инвестиции за двухлетний период равна 0, поскольку первоначальный и финальный объем средств равны между собой. Однако среднее арифметическое годовых норм прибыли равно $\overline{X}$ = (–0,5 + 1) / 2 = 0,25 или 25%, поскольку норма прибыли в первый год R_1 = (50 000 – 100 000) / 100 000 = –0,5, а во второй R_2 = (100 000 – 50 000) / 50 000 = 1. В то же время, среднее геометрическое значение нормы прибыли за два года равно: $\overline{R}_G = [(1–0,5) * (1+1)]^{1/2} – 1 = [0,5*2,0]^{½} – 1 = 1 – 1 = 0$. Таким образом, среднее геометрическое точнее отражает изменение (точнее, отсутствие изменений) объема инвестиций за двухлетний период, чем среднее арифметическое.

Интересные факты. Во-первых, среднее геометрическое всегда будет меньше среднего арифметического тех же чисел. За исключением случая, когда все взятые числа равны друг другу. Во-вторых, рассмотрев свойства прямоугольного треугольника, можно понять, почему среднее называется геометрическим. Высота прямоугольного треугольника, опущенная на гипотенузу, есть среднее пропорциональное между проекциями катетов на гипотенузу, а каждый катет есть среднее пропорциональное между гипотенузой и его проекцией на гипотенузу (рис. 5). Это даёт геометрический способ построения среднего геометрического двух (длин) отрезков: нужно построить окружность на сумме этих двух отрезков как на диаметре, тогда высота, восставленная из точки их соединения до пересечения с окружностью, даст искомую величину: $BH = \sqrt{AH * HC} = \sqrt{ab}$.

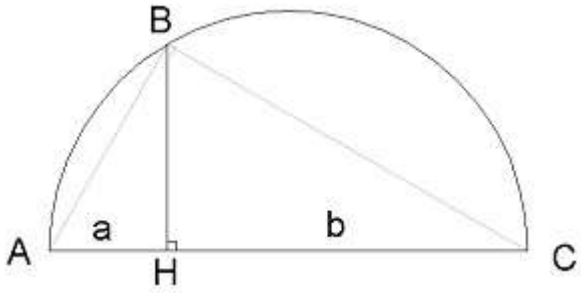


Рис. 5. Геометрическая природа среднего геометрического (рисунок из Википедии)

* * *

Второе важное свойство числовых данных — их **вариация**, характеризующая степень дисперсии данных. Две разные выборки могут отличаться как средними значениями, так и вариациями. Однако, как показано на рис. 6 и 7, две выборки могут иметь одинаковые вариации, но разные средние значения, либо одинаковые средние значения и совершенно разные вариации. Данные, которым соответствует полигон В на рис. 7, изменяются намного меньше, чем данные, по которым построен полигон А.

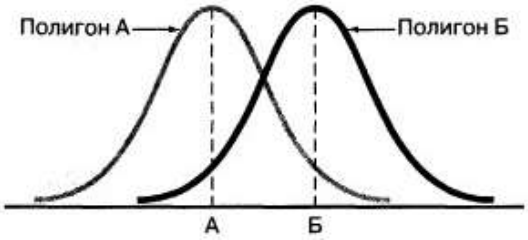


Рис. 6. Два симметричных распределения колоколообразной формы с одинаковым разбросом и разными средними значениями

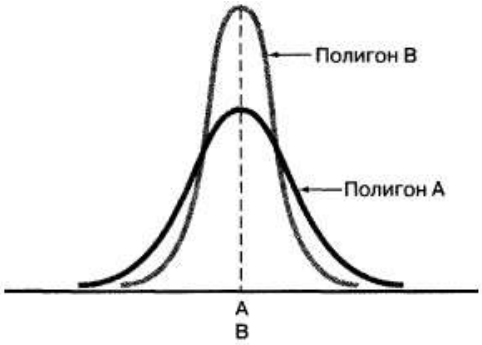


Рис. 7. Два симметричных распределения колоколообразной формы с одинаковыми средними значениями и разным разбросом

Существует пять оценок вариации данных:

- размах,
- межквартильный размах,
- дисперсия,
- стандартное отклонение,
- коэффициент вариации.

Размахом называется разность между наибольшим и наименьшим элементами выборки:

Размах = $X_{Max} – X_{Min}$

Размах выборки, содержащей данные о среднегодовой доходности 15 взаимных фондов с очень высоким уровнем риска, можно вычислить, используя упорядоченный массив (см. рис. 4): Размах = 18,5 – (–6,1) = 24,6. Это значит, что разница между наибольшей и наименьшей среднегодовой доходностью фондов с очень высоким уровнем риска равна 24,6% .

Размах позволяет измерить общий разброс данных. Хотя размах выборки является весьма простой оценкой общего разброса данных, его слабость заключается в том, что он никак не учитывает, как именно распределены данные между минимальным и максимальным элементами. Этот эффект хорошо прослеживается на рис. 8, который иллюстрирует выборки, имеющие одинаковый размах. Шкала В демонстрирует, что если выборка содержит хотя бы одно экстремальное значение, размах выборки оказывается весьма неточной оценкой разброса данных.

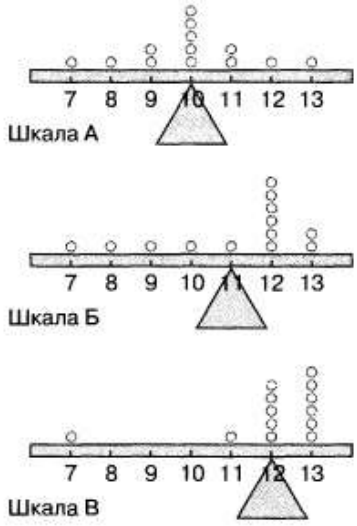


Рис. 8. Сравнение трех выборок, имеющих одинаковый размах; треугольник символизирует опору весов, и его расположение соответствует среднему значению выборки

Межквартильный размах или средний размах — это разность между третьим и первым квартилями выборки:

Межквартильный размах = $Q_3 - Q_1$

Эта величина позволяет оценить разброс 50% элементов и не учитывать влияние экстремальных элементов. Межквартильный размах выборки, содержащей данные о среднегодовой доходности 15 взаимных фондов с очень высоким уровнем риска, можно вычислить, используя данные на рис. 4 (например, для функции КВАРТИЛЬ.ИСКЛ): Межквартильный размах = 9,8 – (–0,7) = 10,5. Интервал, ограниченный числами 9,8 и –0,7, часто называют средней половиной.

Следует отметить, что величины Q_1 и Q_3, а значит, и межквартильный размах, не зависят от наличия выбросов, поскольку при их вычислении не учитывается ни одна величина, которая была бы меньше Q_1 или больше Q_3. Суммарные количественные характеристики, такие как медиана, первый и третий квартили, а также межквартильный размах, на которые не влияют выбросы, называются устойчивыми показателями.

Дисперсия и стандартное отклонение

Хотя размах и межквартильный размах позволяют оценить общий и средний разброс выборки соответственно, ни одна из этих оценок не учитывает, как именно распределены данные. *Дисперсия и стандартное отклонение* лишены этого недостатка. Эти показатели позволяют оценить степень колебания данных вокруг среднего значения. *Выборочная дисперсия* является приближением среднего арифметического, вычисленного на основе квадратов разностей между каждым элементом выборки и выборочным средним. Для выборки X_1, X_2, … X_n выборочная дисперсия (обозначаемая символом S^2 задается следующей формулой:

$$S^2 = \frac{\left(X_1 - \bar{X}\right)^2 + \left(X_2 - \bar{X}\right)^2 + \ldots + \left(X_n - \bar{X}\right)^2}{n-1}$$

В общем случае выборочная дисперсия — это сумма квадратов разностей между элементами выборки и выборочным средним, деленная на величину, равную объему выборки минус один:

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

где $\bar{X}$ — арифметическое среднее, n — объем выборки, X_i — i-й элемент выборки X. В Excel до версии 2007 для расчета выборочной дисперсии использовалась функция =ДИСП(), с версии 2010 используется функция =ДИСП.В().

Наиболее практичной и широко распространенной оценкой разброса данных является *стандартное выборочное отклонение*. Этот показатель обозначается символом S и равен квадратному корню из выборочной дисперсии:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

В Excel до версии 2007 для расчета стандартного выборочного отклонения использовалась функция =СТАНДОТКЛОН(), с версии 2010 используется функция =СТАНДОТКЛОН.В(). Для расчета этих функций массив данных может быть неупорядоченным.

Ни выборочная дисперсия, ни стандартное выборочное отклонение не могут быть отрицательными. Единственная ситуация, в которой показатели S^2 и S могут быть нулевыми, — если все элементы выборки равны между собой. В этом совершенно невероятном случае размах и межквартильный размах также равны нулю.

Числовые данные по своей природе изменчивы. Любая переменная может принимать множество разных значений. Например, разные взаимные фонды имеют разные показатели доходности и убытков. Вследствие изменчивости числовых данных очень важно изучать не только оценки среднего значения, которые по своей природе являются суммарными, но и оценки дисперсии, характеризующие разброс данных.

Дисперсия и стандартное отклонение позволяют оценить разброс данных вокруг среднего значения, иначе говоря, определить, сколько элементов выборки меньше среднего, а сколько — больше. Дисперсия обладает некоторыми ценными математическими свойствами. Однако ее величина представляет собой квадрат единицы измерения — квадратный процент, квадратный доллар, квадратный дюйм и т.п. Следовательно, естественной оценкой дисперсии является стандартное отклонение, которое выражается в обычных единицах измерений — процентах дохода, долларах или дюймах.

Стандартное отклонение позволяет оценить величину колебаний элементов выборки вокруг среднего значения. Практически во всех ситуациях основное количество наблюдаемых величин лежит в интервале плюс-минус одно стандартное отклонение от среднего значения. Следовательно, зная среднее арифметическое элементов выборки и стандартное выборочное отклонение, можно определить интервал, которому принадлежит основная масса данных.

Стандартное отклонение доходности 15 взаимных фондов с очень высоким уровнем риска равно 6,6 (рис. 9). Это значит, что доходность основной массы фондов отличается от среднего значения не более чем на 6,6% (т.е. колеблется в интервале от $\bar{X} - S$ = 6,2 – 6,6 = –0,4 до $\bar{X} + S$ = 12,8). Фактически в этом интервале лежит пятилетняя среднегодовая доходность 53,3% (8 из 15) фондов.

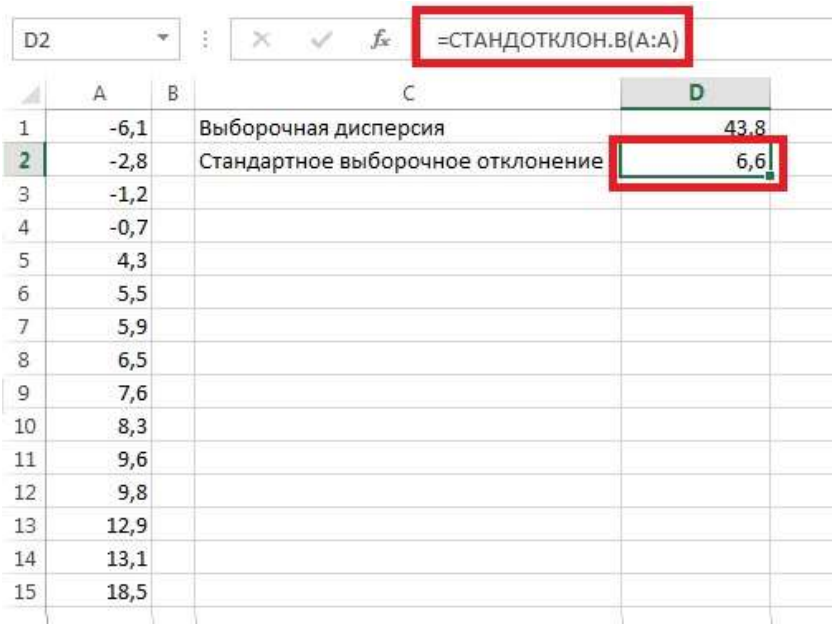


Рис. 9. Стандартное выборочное отклонение

Обратите внимание на то, что в процессе суммирования квадратов разностей элементы выборки, лежащие дальше от среднего значения, приобретают больший вес, чем элементы, лежащие ближе. Это свойство является основной причиной того, что для оценки среднего значения распределения чаще всего используется среднее арифметическое значение.

Коэффициент вариации

В отличие от предыдущих оценок разброса, коэффициент вариации является относительной оценкой. Он всегда измеряется в процентах, а не в единицах измерения исходных данных. Коэффициент вариации, обозначаемый символами CV, измеряет рассеивание данных относительно среднего значения. Коэффициент вариации равен стандартному отклонению, деленному на среднее арифметическое и умноженному на 100%:

$$CV = \frac{S}{\overline{X}} * 100\%$$

где S — стандартное выборочное отклонение, $\overline{X}$ — выборочное среднее.

Коэффициент вариации позволяет сравнить две выборки, элементы которых выражаются в разных единицах измерения. Например, управляющий службы доставки корреспонденции намеревается обновить парк грузовиков. При погрузке пакетов следует учитывать два вида ограничений: вес (в фунтах) и объем (в кубических футах) каждого пакета. Предположим, что в выборке, содержащей 200 пакетов, средний вес равен 26,0 фунтов, стандартное отклонение веса 3,9 фунтов, средний объем пакета 8,8 кубических футов, а стандартное отклонение объема 2,2 кубических фута. Как сравнить разброс веса и объема пакетов?

Поскольку единицы измерения веса и объема отличаются друг от друга, управляющий должен сравнить относительный разброс этих величин. Коэффициент вариации веса равен CV_W = 3,9 / 26,0 * 100% = 15%, а коэффициент вариации объема CV_V = 2,2 / 8,8 * 100% = 25%. Таким образом, относительный разброс объема пакетов намного больше относительного разброса их веса.

Форма распределения

Третье важное свойство выборки — форма ее распределения. Это распределение может быть симметричным или асимметричным. Чтобы описать форму распределения, необходимо вычислить его среднее значение и медиану. Если эти два показателя совпадают, переменная считается симметрично распределенной. Если среднее значение переменной больше медианы, ее распределение имеет положительную асимметрию (рис. 10). Если медиана больше среднего значения, распределение переменной имеет отрицательную асимметрию. Положительная асимметрия возникает, когда среднее значение увеличивается до необычайно высоких значений. Отрицательная асимметрия возникает, когда среднее значение уменьшается до необычайно малых

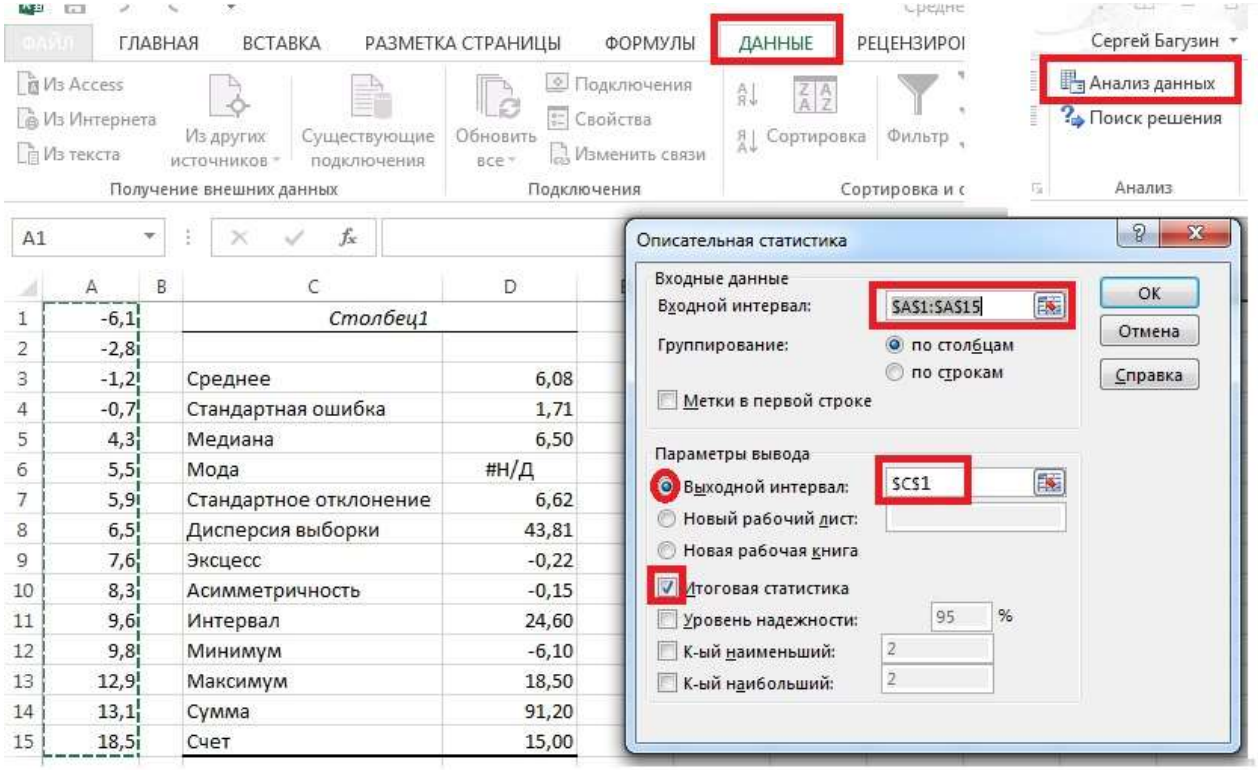
значений. Переменная является симметрично распределенной, если она не принимает никаких экстремальных значений ни в одном из направлений, так что большие и малые значения переменной уравновешивают друг друга.

Рис. 10. Три вида распределений

Данные, изображенные на шкале А, имеют отрицательную асимметрию. На этом рисунке виден длинный хвост и перекос влево, вызванные наличием необычно малых значений. Эти крайне малые величины смещают среднее значение влево, и оно становится меньше медианы. Данные, изображенные на шкале Б, распределены симметрично. Левая и правая половины распределения являются своими зеркальными отражениями. Большие и малые величины уравновешивают друг друга, а среднее значение и медиана равны между собой. Данные, изображенные на шкале В, имеют положительную асимметрию. На этом рисунке виден длинный хвост и перекос вправо, вызванные наличием необычайно высоких значений. Эти слишком большие величины смещают среднее значение вправо, и оно становится больше медианы.

В Excel описательные статистики можно получить с помощью надстройки *Пакет анализа*. Пройдите по меню *Данные* → *Анализ данных*, в открывшемся окне выберите строку *Описательная статистика* и кликните *Ok*. В окне *Описательная статистика* обязательно укажите *Входной интервал* (рис. 11). Если вы хотите увидеть описательные статистики на том же листе, что и исходные данные, выберите переключатель *Выходной интервал* и укажите ячейку, куда следует поместить левый верхний угол выводимых статистик (в нашем примере C1). Если вы хотите вывести данные на новый лист или в новую книгу, достаточно просто выбрать соответствующий переключатель. Поставьте галочку напротив *Итоговая статистика*. По желанию также можно выбрать *Уровень сложности, k-ый наименьший и k-й наибольший*.

Если на вкладе *Данные* в области *Анализ* у вас не отображается пиктограмма *Анализ данных*, нужно предварительно установить надстройку *Пакет анализа* (см., например, Представление числовых данных в виде таблиц и диаграмм).



	A	B	C	D
1	-6,1		*Столбец1*	
2	-2,8			
3	-1,2		Среднее	6,08
4	-0,7		Стандартная ошибка	1,71
5	4,3		Медиана	6,50
6	5,5		Мода	#Н/Д
7	5,9		Стандартное отклонение	6,62
8	6,5		Дисперсия выборки	43,81
9	7,6		Эксцесс	-0,22
10	8,3		Асимметричность	-0,15
11	9,6		Интервал	24,60
12	9,8		Минимум	-6,10
13	12,9		Максимум	18,50
14	13,1		Сумма	91,20
15	18,5		Счет	15,00

Рис. 11. Описательные статистики пятилетней среднегодовой доходности фондов с очень высоким уровнями риска, вычисленные с помощью надстройки *Анализ данных* программы Excel

Excel вычисляет целый ряд статистик, рассмотренных выше: среднее, медиану, моду, стандартное отклонение, дисперсию, размах (*интервал*), минимум, максимум и объем выборки (*счет*). Кроме того, Excel вычисляет некоторые новые для нас статистики: стандартную ошибку, эксцесс и асимметричность. *Стандартная ошибка* равна стандартному отклонению, деленному на квадратный корень объема выборки. *Асимметричность* характеризует отклонение от симметричности распределения и является функцией, зависящей от куба разностей между элементами выборки и средним значением. Эксцесс представляет собой меру относительной концентрации данных вокруг среднего значения по сравнению с хвостами распределения и зависит от разностей между элементами выборки и средним значением, возведенных в четвертую степень.

Вычисление описательных статистик для генеральной совокупности

Среднее значение, разброс и форма распределения, рассмотренные выше, представляют собой характеристики, определяемые по выборке. Однако, если набор данных содержит числовые измерения всей генеральной совокупности, можно вычислить ее параметры. К числу таких параметров относятся математическое ожидание, дисперсия и стандартное отклонение генеральной совокупности.

Математическое ожидание равно сумме всех значений генеральной совокупности, деленной на объем генеральной совокупности:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

где μ — математическое ожидание, X_i — i-е наблюдение переменной X, N — объем генеральной совокупности. В Excel для вычисления математического ожидания используется та же функция, что и для среднего арифметического: =СРЗНАЧ().

Дисперсия генеральной совокупности равна сумме квадратов разностей между элементами генеральной совокупности и мат. ожиданием, деленной на объем генеральной совокупности:

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

где σ^2 – дисперсия генеральной совокупности. В Excel до версии 2007 для вычисления дисперсии генеральной совокупности используется функция =ДИСПР(), начиная с версии 2010 =ДИСП.Г().

Стандартное отклонение генеральной совокупности равно квадратному корню, извлеченному из дисперсии генеральной совокупности:

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

В Excel до версии 2007 для вычисления стандартного отклонения генеральной совокупности используется функция =СТАНДОТКЛОНП(), начиная с версии 2010 =СТАНДОТКЛОН.Г(). Обратите внимание на то, что формулы для дисперсии и стандартного отклонения генеральной совокупности отличаются от формул для вычисления выборочной дисперсии и стандартного отклонения. При вычислении выборочных статистик S^2 и S знаменатель дроби равен $n - 1$, а при вычислении параметров σ^2 и σ — объему генеральной совокупности N.

Эмпирическое правило

В большинстве ситуаций крупная доля наблюдений концентрируется вокруг медианы, образуя кластер. В наборах данных, имеющих положительную асимметрию, этот кластер расположен левее (т.е. ниже) математического ожидания, а в наборах, имеющих отрицательную асимметрию, этот кластер расположен правее (т.е. выше) математического ожидания. У симметричных данных математическое ожидание и медиана совпадают, а наблюдения концентрируются вокруг математического ожидания, формируя колоколообразное распределение. Если распределение не имеет ярко выраженной асимметрии, а данные концентрируются вокруг некоего центра тяжести, для оценки изменчивости можно применять эмпирическое правило, которое гласит: если данные имеют колоколообразное распределение, то приблизительно 68% наблюдений отстоят от математического ожидания не более чем на одно стандартное отклонение, приблизительно 95% наблюдений отстоят от математического ожидания не более чем на два стандартных отклонения и 99,7% наблюдений отстоят от математического ожидания не более чем на три стандартных отклонения.

Таким образом, стандартное отклонение, представляющее собой оценку среднего колебания вокруг математического ожидания, помогает понять, как распределены наблюдения, и идентифицировать выбросы. Из эмпирического правила следует, что для колоколообразных распределений лишь одно значение из двадцати отличается от математического ожидания больше, чем на два стандартных отклонения. Следовательно, значения, лежащие за пределами интервала $\mu \pm 2\sigma$, можно считать выбросами. Кроме того, только три из 1000 наблюдений отличаются от математического ожидания больше чем на три стандартных отклонения. Таким образом, значения, лежащие за пределами интервала $\mu \pm 3\sigma$ практически всегда являются выбросами. Для распределений, имеющих сильную асимметрию или не имеющих колоколообразной формы, можно применять эмпирическое правило Бьенамэ-Чебышева.

Более ста лет назад математики Бьенамэ и Чебышев независимо друг от друга открыли полезное свойство стандартного отклонения. Они обнаружили, что для любого набора данных, независимо от формы распределения, процент наблюдений, лежащих на расстоянии не превышающем k стандартных отклонений от математического ожидания, не меньше $(1 - 1/k^2)*100\%$.

Например, если $k = 2$, правило Бьенамэ-Чебышева гласит, что как минимум $(1 - (1/2)^2) \times 100\% = 75\%$ наблюдений должно лежать в интервале $\mu \pm 2\sigma$. Это правило справедливо для любого k, превышающего единицу. Правило Бьенамэ-Чебышева носит весьма общий характер и справедливо для распределений любого вида. Оно указывает минимальное количество наблюдений, расстояние от которых до математического ожидания не превышает заданной величины. Однако, если распределение имеет колоколообразную форму, эмпирическое правило более точно оценивает концентрацию данных вокруг математического ожидания.

Вычисление описательных статистик для распределения на основе частот. Если исходные данные недоступны, единственным источником информации становится распределение частот. В таких ситуациях можно вычислить приближенные значения количественных показателей распределения, таких как среднее арифметическое, стандартное отклонение, квартили.

Если выборочные данные представлены в виде распределения частот, приближенное значение среднего арифметического можно вычислить, предполагая, что все значения внутри каждого класса сосредоточены в средней точке класса:

$$\bar{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

где $\bar{X}$ — выборочное среднее, n — количество наблюдений, или объем выборки, c — количество классов в распределении частот, m_j — средняя точка j-го класса, f_j — частота, соответствующая j-му классу.

Для вычисления стандартного отклонения по распределению частот также предполагается, что все значения внутри каждого класса сосредоточены в средней точке класса.

$$S = \sqrt{\frac{\sum_{j=1}^{c} (m_j - \bar{X})^2 f_j}{n-1}}$$

Чтобы понять, как определяются квартили ряда на основе частот, рассмотрим расчет нижнего квартиля на основе данных за 2013 г. о распределении населения России по величине среднедушевых денежных доходов (рис. 12).

В процентах	2010	2011	2012	2013	Накопленная сумма процентов за 2013
до 5000	9,4	7,3	5,9	6,0	6,0
5000,1 – 7000,0	9,4	8,1	7,0	7,4	13,4
7000,1 – 10 000,0	14,6	13,4	12,1	13,0	26,4
10 000,1 – 14 000,0	16,6	16,2	15,4	16,4	42,8
14000,1 – 19 000,0	15,2	15,6	15,5	16,2	59,0
19 000,1 – 27 000,0	14,7	15,9	16,6	16,7	75,7
27 000,1 – 45 000,0	13,3	15,1	17,0	15,9	91,6
свыше 45 000	6,8	8,4	10,5	8,4	100,0
Всего, %	100	100	100	100	
Данные 2013 года относятся к первому кварталу и являются предварительными					

Рис. 12. Доля населения России со среднедушевыми денежными доходами в среднем за месяц, рублей

Для расчета первого квартиля ряда на основе частот можно воспользоваться формулой:[3]

$$Q_1 = x_{Q_1} + i \cdot \frac{\frac{1}{4}\sum f - S_{Q_1-1}}{f_{Q_1}}$$

где Q_1 – величина первого квартиля, x_{Q1} – нижняя граница интервала, содержащего первый квартиль (интервал определяется по накопленной частоте, первой превышающей 25%); i – величина интервала; Σf – сумма частот всей выборки; наверное, всегда равна 100%; S_{Q1-1} – накопленная частота интервала, предшествующего интервалу, содержащему нижний квартиль; f_{Q1} – частота интервала, содержащего нижний квартиль. Формула для третьего квартиля отличается тем, что во всех местах вместо Q_1 нужно использовать Q_3, а вместо ¼ подставить ¾.

В нашем примере (рис. 12) нижний квартиль находится в интервале 7000,1 – 10 000, накопленная частота которого равна 26,4%. Нижняя граница этого интервала – 7000 руб., величина интервала – 3000 руб., накопленная частота интервала, предшествующего интервалу, содержащему нижний

[3] Используются материалы Структурные характеристики вариационного ряда распределения

квартиль – 13,4%, частота интервала, содержащего нижний квартиль – 13,0%. Таким образом: Q_1 = 7000 + 3000 * (¼ * 100 – 13,4) / 13 = 9677 руб.

Ловушки, связанные с описательными статистиками

В этой заметке мы рассмотрели, как описать набор данных с помощью различных статистик, оценивающих его среднее значение, разброс и вид распределения. Следующим этапом является анализ и интерпретация данных. До сих пор мы изучали объективные свойства данных, а теперь переходим к их субъективной трактовке. Исследователя подстерегают две ошибки: неверно выбранный предмет анализа и неправильная интерпретация результатов.

Анализ доходности 15 взаимных фондов с очень высоким уровнем риска является вполне беспристрастным. Он привел к совершенно объективным выводам: все взаимные фонды имеют разную доходность, разброс доходности фондов колеблется от –6,1 до 18,5, а средняя доходность равна 6,08. Объективность анализа данных обеспечивается правильным выбором суммарных количественных показателей распределения. Было рассмотрено несколько способов оценки среднего значения и разброса данных, указаны их преимущества и недостатки. Как же выбрать правильную статистику, обеспечивающую объективный и беспристрастный анализ? Если распределение данных имеет небольшую асимметрию, следует ли выбирать медиану, а не среднее арифметическое? Какой показатель более точно характеризует разброс данных: стандартное отклонение или размах? Следует ли указывать на положительную асимметрию распределения?

С другой стороны, интерпретация данных является субъективным процессом. Разные люди приходят к разным выводам, истолковывая одни и те же результаты. У каждого своя точка зрения. Кто-то считает суммарные показатели среднегодовой доходности 15 фондов с очень высоким уровнем риска хорошими и вполне доволен полученным доходом. Другим может показаться, что эти фонды имеют слишком низкую доходность. Таким образом, субъективность следует компенсировать честностью, нейтральностью и ясностью выводов.

Этические проблемы

Анализ данных неразрывно связан с этическими вопросами. Следует критически относиться к информации, распространяемой газетами, радио, телевидением и Интерентом. Со временем вы научитесь скептически относиться не только к результатам, но и к целям, предмету и объективности исследований. Лучше всего об этом сказал известный британский политик Бенджамин Дизраэли: «Существуют три вида лжи: ложь, наглая ложь и статистика».

Как было отмечено в заметке Искусство графического представления данных этические проблемы возникают при выборе результатов, которые следует привести в отчете. Следует публиковать как положительные, так и отрицательные результаты. Кроме того, делая доклад или письменный отчет, результаты необходимо излагать честно, нейтрально и объективно. Следует различать неудачную и нечестную презентации. Для этого необходимо определить, каковы были намерения докладчика. Иногда важную информацию докладчик пропускает по невежеству, а иногда — умышленно (например, если он применяет среднее арифметическое для оценки среднего значения явно асимметричных данных, чтобы получить желаемый результат). Нечестно также замалчивать результаты, которые не соответствуют точке зрения исследователя.

Предыдущая заметка Искусство графического представления данных

Следующая заметка

К оглавлению Статистика для менеджеров с использованием Microsoft Excel