Нормальное распределение

Числовые случайные величины могут быть либо дискретными, либо непрерывными (подробнее см. Типы данных). Дискретные случайные величины (т.е. величины, возникающие в результате подсчета событий) были рассмотрены ранее (см. Биноминальное распределение, Гипергеометрическое распределение, Распределение Пуассона). В этой и нескольких следующих заметках мы изучим непрерывные случайные величины, которые возникают в результате измерений. Непрерывная случайная величина может принимать любое значение, принадлежащее числовой оси или интервалу.[1] Примером такой случайной величины может служить вес какой-нибудь коробки, время загрузки Web-страницы, расходы на рекламу, доходы от продаж, время обслуживания клиента и время между двумя приходами клиентов в банк.

Математическое выражение, описывающее распределение значений непрерывной случайной величины, называется плотностью непрерывного распределения вероятностей (рис. 1). На панели А представлена плотность нормального распределения. Эта функция является симметричной и колоколообразной. Следовательно, большинство значений такой случайной величины концентрируется вокруг математического ожидания, которое совпадает с медианой. Несмотря на то что нормально распределенная случайная величина может принимать любые числовые значения, вероятность очень больших положительных или отрицательных значений крайне мала. На панели Б изображена плотность равномерного распределения. Значения случайной величины, равномерно распределенной на интервале от *a* до *b*, равновероятны. Иногда это распределение называют прямоугольным. Оно является симметричным, и, следовательно, его математическое ожидание равно медиане. На панели В показана плотность экспоненциального распределения. Это распределение имеет ярко выраженную положительную асимметрию, и, следовательно, его математическое ожидание больше медианы. Экспоненциально распределенные случайные величины изменяются от нуля до плюс бесконечности, однако очень большие значения крайне мало вероятны.

Рис. 1. Три непрерывных распределения

В этой заметке описывается одно из наиболее важных распределений в статистике — нормальное распределение, которое также называют гауссовым. То, что это одно из наиболее часто используемых распределений, косвенно подтверждается, в частности тем, что ранее я уже писал на эту тему ☺, см. Нормальное распределение. Построение графика в Excel. Концепция шести сигм. Страница с этой заметкой является третьей по посещаемости из более чем 370 на сайте.

Плотность нормального распределения изображена на рис. 1а. Можно вычислить вероятность того, что нормально распределенная случайная величина лежит в заданном интервале. Однако вероятность того, что она принимает наперед заданное значение, равна нулю. Это отличает непрерывные случайные величины (измеряемые) от дискретных (подсчитываемых). Например, время измеряется, а не подсчитывается. Следовательно, можно вычислить вероятность того, что Web-страница будет загружаться от 7 до 10 с. Сужая заданный интервал, можно вычислить вероятность того, что она будет загружаться от 8 до 9 с. Кроме того, можно вычислить вероятность того, что она будет загружаться от 8,99 до 9,01 с. Однако вероятность того, что Web-страница будет загружаться ровно 8 с, равна нулю.

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 346–363

Определение вероятностей или вычисление математического ожидания и стандартного отклонения непрерывной случайной величины требуют знания интегрального исчисления и не будут здесь рассматриваться. Однако использование функций Excel позволяет легко преодолеть эти трудности.

Важность нормального распределения в статистике обусловлена тремя причинами:

1. Оно описывает (точно или приблизительно) распределение многих непрерывных случайных величин.
2. С помощью нормального распределения можно аппроксимировать разнообразные дискретные распределения.
3. Нормальное распределение лежит в основе классической теории статистических выводов, поскольку оно тесно связано с центральной предельной теоремой.

Нормальное распределение:

- Имеет колоколообразную (а значит, симметричную) форму.
- Его математическое ожидание, медиана и мода совпадают друг с другом.
- Половина нормально распределенных значений лежит в интервале, длина которого равна 4/3 стандартного отклонения. Это значит, что межквартильный размах находится в интервале от 2/3 стандартного отклонения левее среднего значения до 2/3 стандартного отклонения правее среднего значения.
- Значения нормально распределенной случайной величины лежат на всей числовой оси ($-\infty < X < +\infty$).

На практике многие случайные величины являются лишь приближенно нормальными. Иначе говоря, их свойства лишь аппроксимируют теоретические свойства нормального распределения, перечисленные выше. Рассмотрим в качестве примера таблицу на рис. 2.

Толщина (дюймы)	Относительная частота
< 0,0180	48/10 000=0,0048
0,0180 < 0,0182	122/10 000=0,0122
0,0182 < 0,0184	325/10 000=0,0325
0,0184 < 0,0186	695/10 000=0,0695
0,0186 < 0,0188	1 198/10 000=0,1198
0,0188 < 0,0190	1 664/10 000=0,1164
0,0190 < 0,0192	1 896/10 000=0,1896
0,0192 < 0,0194	1 664/10 000=0,1664
0,0194 < 0,0196	1 198/10 000=0,1198
0,0196 < 0,0198	695/10 000=0,0695
0,0198 < 0,0200	325/10 000=0,0325
≥0,0202	48/10 000=0,0048
Всего:	1,0000

Рис. 2. Толщина 10 000 медных дисков

Здесь перечислены результаты измерения толщины 10 000 медных дисков. Толщина представляет собой непрерывную случайную величину, распределение которой аппроксимируется нормальным. Основная масса значений этой величины лежит в интервале от 0,0190 до 0,0192 дюймов и распределена симметрично относительно этого интервала, формируя колоколообразную кривую. Как следует из таблицы, разбиение числовой прямой на интервалы образует группы взаимоисключающих и исчерпывающих событий, сумма вероятностей которых равна единице. Таким образом, распределение вероятностей можно интерпретировать как распределение относительных частот (подробнее см. Представление числовых данных в виде таблиц и диаграмм и последний раздел заметки Определение среднего значения, вариации и формы распределения. Описательные статистики), соответствующих средним точкам интервалов.

На рис. 3 изображена гистограмма относительных частот и полигон распределения толщины 10 000 медных дисков. Как видим, первые три условия нормального распределения выполняются, а четвертое — нет. Толщина диска не может быть отрицательной или равной нулю. Из таблицы (рис. 2) следует, что из 10 000 медных дисков только 48 толще 0,0202 дюйма и такое же количество дисков тоньше 0,0180 дюйма. Таким образом, вероятность случайно выбрать слишком толстый или слишком тонкий диск равна 0,0048+0,0048=0,0096, т.е. меньше 1 из 100.

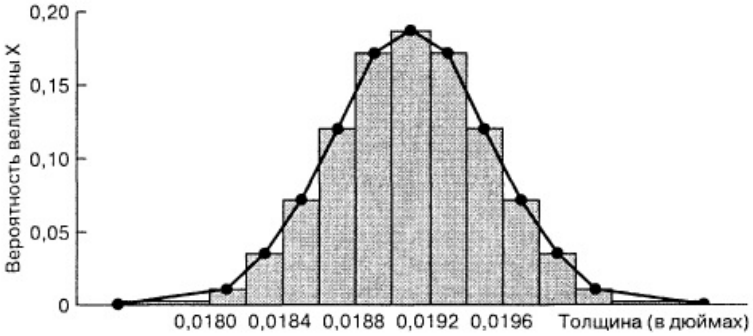

Рис. 3. Гистограмма относительных частот и полигон распределения ширины 10 000 медных дисков

Плотность нормального распределения:

$$(1)\quad f(X) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

где e — основание натурального логарифма, константа равная 2,71828, μ — математическое ожидание генеральной совокупности, σ — стандартное отклонение генеральной совокупности, X — произвольное значение непрерывной случайной величины, $-\infty < X < +\infty$.

Поскольку величины e и π являются математическими константами, плотность нормального распределения зависит только от двух параметров — математического ожидания μ и стандартного отклонения σ (рис. 4). Разным комбинациям этих параметров соответствуют разные плотности нормального распределения. Распределения А и Б имеют одинаковое математическое ожидание μ, но разные стандартные отклонения. С другой стороны, распределения А и В имеют одинаковое стандартное отклонение σ, но разные математические ожидания. Кроме того, распределения Б и В имеют разные математические ожидания и стандартные отклонения.

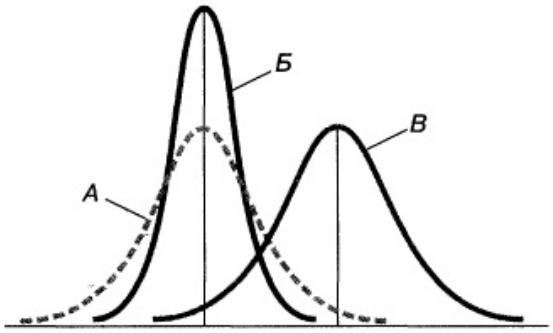

Рис. 4. Три нормальных распределения, соответствующие разным комбинациям параметров μ и σ

К сожалению, вычислить математическое выражение, заданное формулой (1), довольно сложно. Чтобы упростить задачу, значения плотности нормального распределения, как правило, табулируют. Поскольку количество возможных комбинаций параметров μ и σ бесконечно, для вычислений понадобилось бы бесконечное количество таблиц. Однако, если нормировать данные, все распределения можно свести к одной таблице. Используя формулу преобразования, любую нормально распределенную случайную величину X можно преобразовать в нормированную нормально распределенную случайную величину Z.

Величина Z равна разности между величиной X и математическим ожиданием генеральной совокупности p, деленной на стандартное отклонение σ:

$$(2)\quad Z = (X - \mu)/\sigma$$

Математическое ожидание стандартизованного нормального распределения равно нулю, а стандартное отклонение — единице. Плотность стандартизованного нормального распределения можно получить, подставив формулу (2) в формулу (1):

$$(3) \quad f(Z) = \frac{1}{\sqrt{2\pi}} \; e^{-\frac{1}{2}Z^2}$$

Таким образом, любое множество нормально распределенных величин можно преобразовать в стандартизованную форму. Проиллюстрируем процедуру нормирования. Например, время загрузки Web-страницы распределено нормально, причем его математическое ожидание равно p = 7 с, а стандартное отклонение σ = 2 с. Как показывает рис. 5, каждому значению переменной X соответствует нормированное значение Z, полученное с помощью формулы преобразования (2). Следовательно, время загрузки, равное 9 с, на одну стандартную единицу превышает математическое ожидание: Z = (9 – 7) / 2 = +1, а время загрузки равное 1 с на три стандартные единицы (стандартных отклонения) меньше математического ожидания: Z = (1 – 7) / 2 = –3.

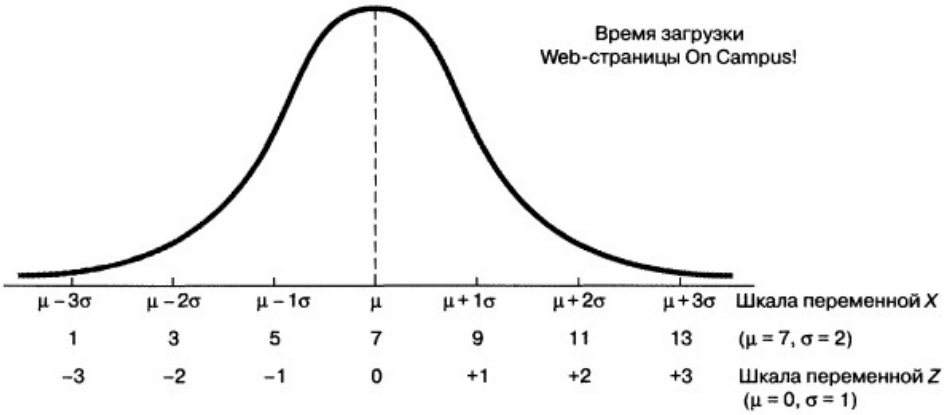

Рис. 5. Преобразование шкал для загрузки Web-сайта; μ = 7, σ = 2

Таким образом, стандартное отклонение становится единицей измерения. Иначе говоря, время загрузки, равное 9 с, на 2 с (т.е. на одно стандартное отклонение) превышает математическое ожидание, а время, равное 1 с, на 6 с (т.е. на три стандартных отклонения) меньше математического ожидания. Допустим теперь, что среднее время загрузки другого Web-сайта равно 4 с, а стандартное отклонение 1 с (рис. 6).

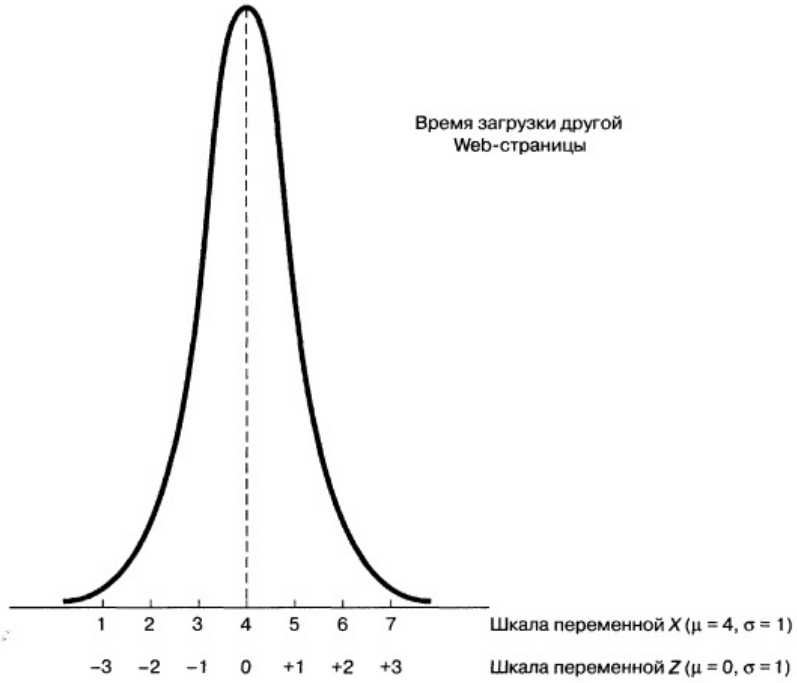

Рис. 6. Преобразование шкал для загрузки Web-сайта; μ = 4, σ = 1

Сравнивая рисунки 5 и 6, легко обнаружить, что время загрузки, равное 5 с, на одно стандартное отклонение больше среднего времени загрузки: Z = (5 – 4) / 1 = +1, а время загрузки, равное 1 с, на три стандартных отклонения меньше математического ожидания: Z = (1 – 4) / 1 = –3. На рис. 5 и 6 показаны полигоны относительных частот, соответствующие времени загрузки двух Web-сайтов. Поскольку результаты измерений образуют полную генеральную совокупность, сумма вероятностей, т.е. площадь фигуры, лежащей под кривой, должна быть равной единице.

Предположим, нам необходимо определить вероятность того, что время загрузки Web-сайта (μ = 7, σ = 2, рис. 5) меньше 9 с. Поскольку это время на одно стандартное отклонение превышает математическое ожидание, следует найти вероятность того, что время загрузки не превышает величины, равной математическому ожиданию плюс одно стандартное отклонение. В Excel2013 для работы с нормально распределенными случайными величинами используется довольно много функций. Для решения нашей задачи идеально подходит =НОРМ.СТ.РАСП(z;интегральная) (рис. 7). В Excel до версии 2007 используется функция =НОРМСТРАСП(z). В ней только один параметр, так как второй параметр (интегральная) по умолчанию равен ИСТИНА.

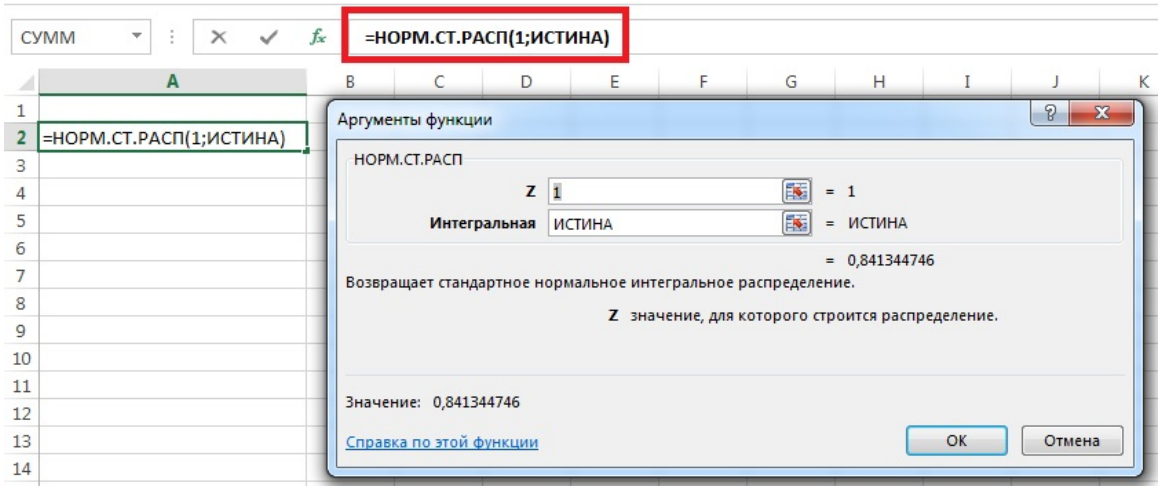

Рис. 7. Расчет вероятности того, что время загрузки Web-сайта (μ = 7, σ = 2, рис. 5) меньше 9 с

Параметр *z* – это координата *X* на нормированной оси (рис. 8). Мы же с помощью функции =НОРМ.СТ.РАСП() определили вероятность того, что случайная величина будет левее *X.*

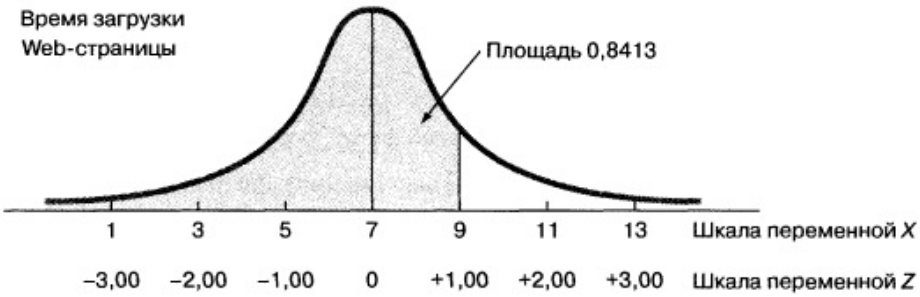

Рис. 8. Площадь фигуры, ограниченной интегральной кривой стандартизованного нормального распределения

С другой стороны, для распределения с параметрами μ = 4, σ = 1 (рис. 6) время загрузки, равное 5 с, на одно стандартное отклонение превышает математическое ожидание, т.е. 4 с. Следовательно, вероятность того, что Web-страница загрузится быстрее, чем за 5 с, также равна 0,8413. На рис. 9 показано, как два отличающихся распределения преобразуются в одно и тоже стандартизованное распределение.

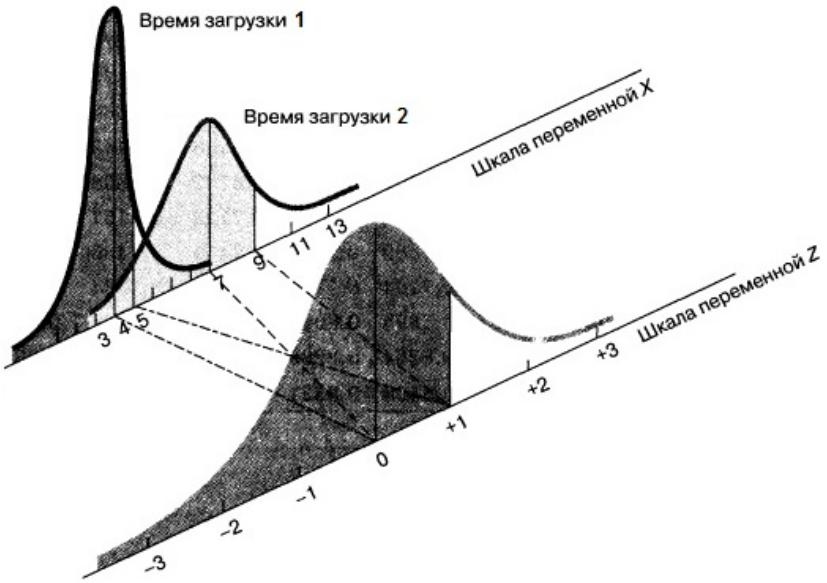

Рис. 9. Преобразование шкал для площадей фигур, ограниченных интегральными кривыми двух нормальных распределений

Рассмотрим несколько примеров.

Пример 1. Определите вероятность того, что время загрузки (μ = 7, σ = 2) превысит 9 с.

Решение. Вероятность того, что время загрузки не превысит 9 с, равна 0,8413, следовательно, искомая вероятность равна 1 – 0,8413 = 0,1587.

Пример 2. Определите вероятность того, что время загрузки (μ = 7, σ = 2) лежит в интервале 7–9 с.

Решение. P(7 < X < 9) = P(X < 9) – P(X < 7). Можно было бы, как и выше, сначала привести нормальное распределение к стандартному виду, а потом воспользоваться функцией =НОРМ.СТ.РАСП (рис. 10).

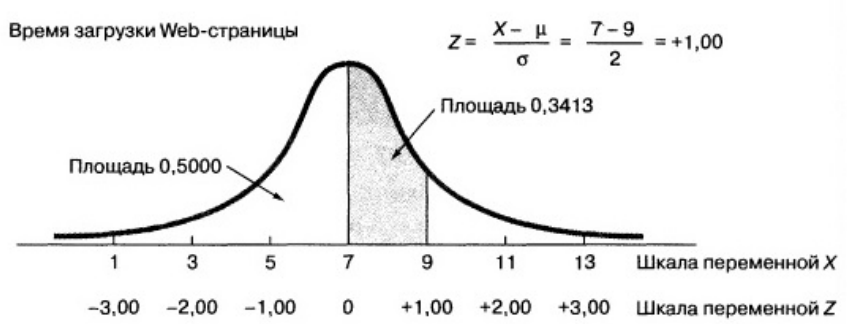

Рис. 10. Определение вероятности P(7 < X < 9)

Однако в Excel есть функция и для нестандартизированного нормального распределения (рис. 11).

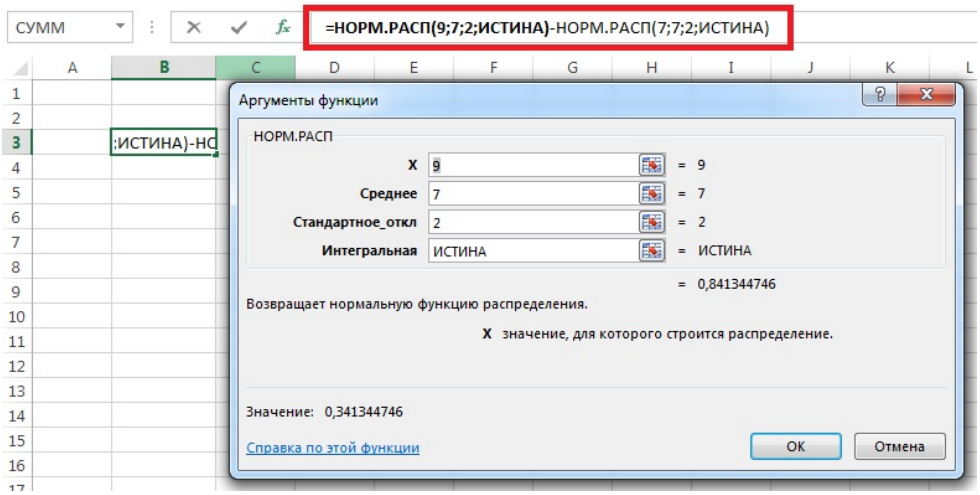

Рис. 11. Определение вероятности того, что время загрузки (μ = 7, σ = 2) лежит в интервале от 7 до 9 с

Обратите внимание, что, поскольку математическое ожидание и медиана нормального распределения совпадают между собой, вероятность того, что загрузка продлится меньше 7 с, равна 0,5, то есть, =НОРМ.РАСП(7;7;2;ИСТИНА) = 0,5.

Пример 3. Определите вероятность того, что время загрузки (μ = 7, σ = 2) лежит в интервале 5–9 с.

Решение. P(5 < X < 9) = P(X < 9) – P(X < 5) =НОРМ.РАСП(9;7;2;ИСТИНА)-НОРМ.РАСП(5;7;2;ИСТИНА) = 0,6826 (рис. 12).

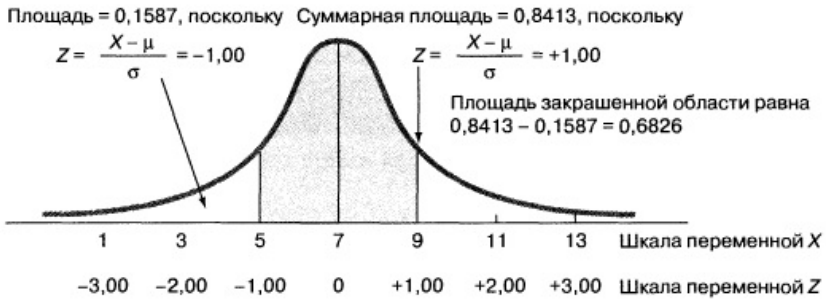

Рис. 12. Определение вероятности P(5 < X < 9)

Полученный результат довольно важен. Для любого нормального распределения вероятность того, что случайно выбранное число лежит в окрестности математического ожидания на расстоянии, не превышающем одно стандартное отклонение, равно 0,6826. В окрестности математического ожидания на расстоянии, не превышающем двух стандартных отклонений, лежит чуть более 95% нормально распределенных величин (рис. 13).

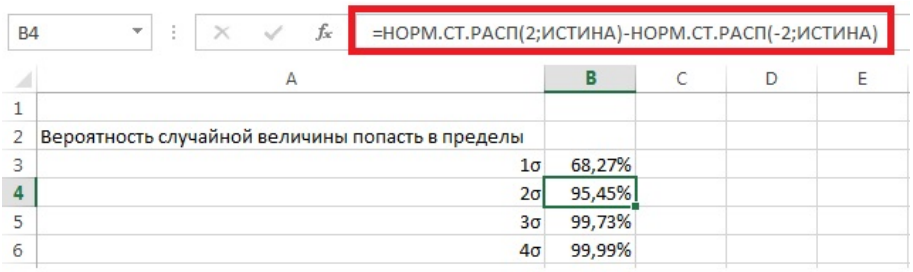

Рис. 13. Вероятность случайной величины попасть в пределы σ-окрестности

В окрестности математического ожидания на расстоянии, не превышающем трех стандартных отклонений, расположено 99,7% всех нормально распределенных величин. Следовательно, 99,73% результатов измерений времени загрузки Web-страницы лежат в интервале от 1 до 13 с. Таким образом, весьма маловероятно (0,0027, или 27 шансов из 10 000), что время загрузки Web-страницы будет меньше 1 с или больше 13 с. Вот почему на практике считают, что интервал длиной 6σ, центром которого является математическое ожидание, содержит практически все значения нормально распределенной случайной величины.

В примерах 1–3 мы вычислили вероятности, связанные с разными значениями измеренной величины. Примеры 4 и 5 посвящены обратной задаче: как определить значение переменной, соответствующей заданной вероятности?

Пример 4. Найдите значение переменной X, соответствующей интегральной вероятности, равной 0,1. Сколько секунд длится загрузка Web-страницы в 10% случаев?

Решение. Поскольку предполагается, что в 10% случаев Web-страница загружается не более чем за X с, площадь фигуры, ограниченной гауссовой кривой и осью абсцисс, равна 0,1 (рис. 14). Для обратной задачи в Excel до версии 2007 существуют две функции =НОРМСТОБР() – возвращает обратное значение стандартного нормального распределения, и =НОРМОБР() – возвращает обратное нормальное распределение (не стандартизированное). В версии Excel, начиная с 2010, им соответствуют функции: =НОРМ.СТ.ОБР() и =НОРМ.ОБР(). В нашем примере =НОРМ.ОБР(0,1;7;2) = 4,4 с (рис. 15).

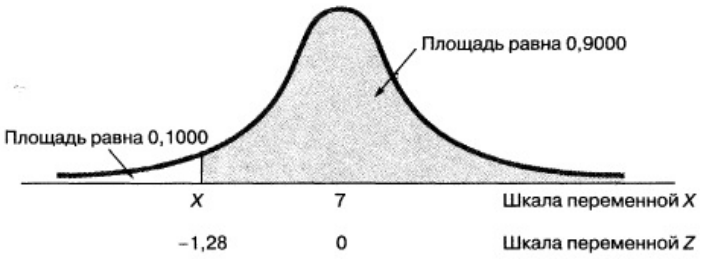

Рис. 14. Интегральная вероятность, равная 0,1

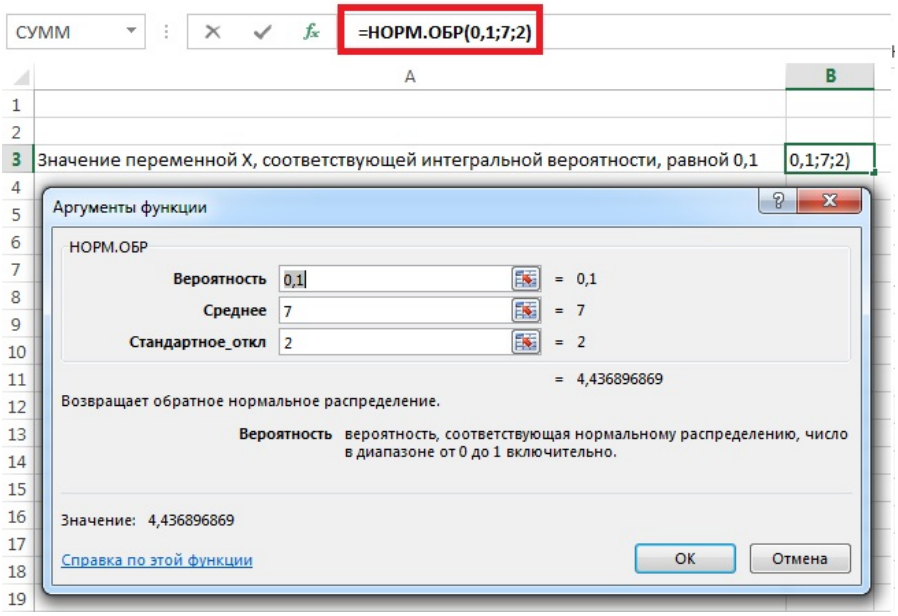

Рис. 15. Использование функции =НОРМ.ОБР()

В общем случае формула для определения величины *X* может быть выведена на основе формулы (2) $Z = (X - \mu)/\sigma$:

(4) $X = \mu + Z\sigma$

Пример 5. Для стандартного нормального распределения определите нижнюю и верхнюю границы интервала с центром в математическом ожидании, который содержит 90% значений случайной величины.

Решение. Нижняя граница Z соответствует такой интегральной вероятности p(Z), которая меньше (1 – 90%) / 2, то есть меньше 5%. Верхняя граница Z соответствует такой интегральной вероятности p(Z), которая больше (1 – 90%) / 2 + 90%, то есть больше 95%.

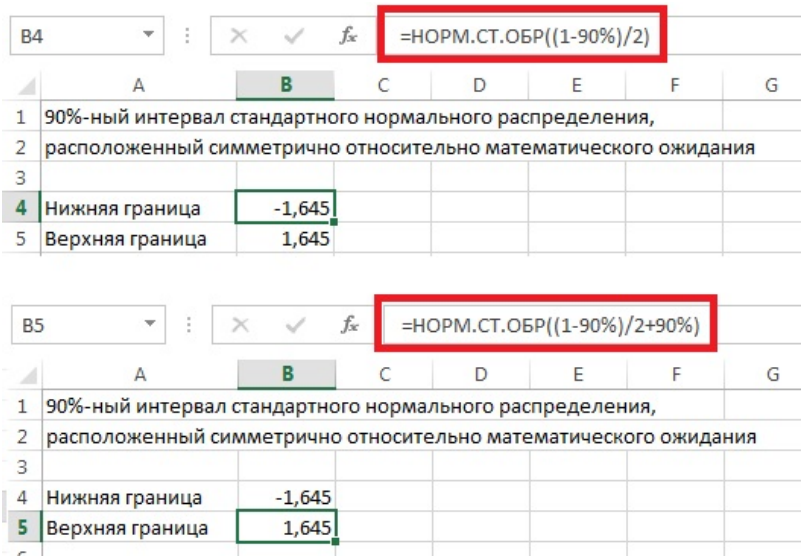

Рис. 16. Определение величин Z, соответствующих значениям ±45%

Таким образом, с вероятностью 90% случайная величина попадает в окрестность ±1,65σ математического ожидания. 90%-ные интервалы находят широкое применение в оценочных суждениях; см., например, Дуглас Хаббард. Как измерить всё, что угодно. Оценка стоимости нематериального в бизнесе.

Предыдущая заметка Распределение Пуассона

Следующая заметка

К оглавлению Статистика для менеджеров с использованием Microsoft Excel