Нормальный закон распределения

Функция распределения F(x) и соответствующая ей плотность распределения f(x) представляют собой некоторую математическую модель свойств исследуемой случайной величины (отклика), значения которой регистрируются в ходе эксперимента.

Поэтому одной из **основных задач** статистической обработки опытных данных является **нахождение таких функций распределения**, которые, с одной стороны, достаточно хорошо описывали бы наблюдаемые значения случайной величины, а с другой – были бы удобны для дальнейшего статистического анализа.

При этом **вид функции** распределения предпочтительно выбирать на основе **представлений о физической природе** рассматриваемого явления, т.к. в этом случае исключаются возможные погрешности при распространении найденных закономерностей за пределы изучаемого в эксперименте интервала варьирования (изменения) случайной величины (отклика).

Из всех изученных к настоящему времени случайных величин при обработке экспериментальных данных исследователи чаще всего оперируют со случайными величинами, которые имеют так называемое нормальное (Гауссово) распределение (рис.3).

Не вдаваясь в подробные математические выкладки, отметим, что, согласно **центральной предельной теореме математической статистики**,

«при определенных условиях распределение нормированной суммы n независимых случайных величин, распределенных по произвольному закону, стремится к нормальному, когда n стремится к бесконечности».

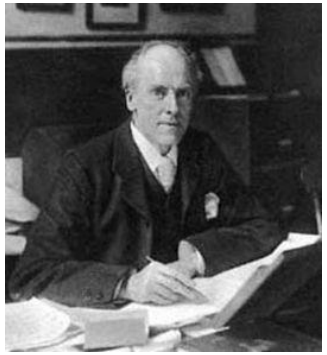
Необходимые условия, при которых эта теорема оказывается справедливой, состоят в том, что различные случайные величины должны иметь **конечные дисперсии** и дисперсия любой случайной величины не должна быть слишком большой по сравнению с дисперсиями других.

При обработке экспериментальных данных эта теорема имеет очень большое значение, поскольку отклик становится случайной величиной в результате влияния неконтролируемых факторов, **число которых скорее всего стремится к бесконечности.**

Кроме того, если при проведении опытов все наиболее существенные факторы контролируются, то воздействие на отклик каждого из **неконтролируемых факторов не должно быть слишком большим по сравнению с остальными неконтролируемыми факторами**.

Другими словами, та дисперсия (рассеивание) отклика, которую вызывает какой-либо из неконтролируемых факторов, не должна сильно отличаться от дисперсий, связанных с влиянием остальных неконтролируемых факторов. В противном случае фактор, дисперсия от которого существенно отличается от других, обязательно **должен быть переведен в разряд контролируемых**.





Следовательно, если при планировании эксперимента учтены все наиболее существенные факторы и затем, при проведении опытов, они контролируются, то при обработке экспериментальных данных можно предполагать, что отклик не должен противоречить нормальному распределению.

При обработке результатов наблюдений исследователи прежде всего предполагают именно нормальное распределение отклика.

Большинство других распределений, которые используются в математической статистике (Стьюдента, Фишера, Пирсона, Кохрена, а также распределения, по которым составлены различные критериальные таблицы), получены на основе нормального распределения.

Нельзя, однако, абсолютизировать значение нормального распределения.

Не все случайные величины распределены по нормальному закону. Тем не менее на практике, если явление подвержено действию многих случайных факторов, их суммарное воздействие вполне оправданно можно описать с помощью нормального закона.

Как уже было отмечено, для случайной величины, которая не противоречит нормальному закону, функция распределения (12) и соответствующая ей плотность распределения

$$F(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{x} e^{-\frac{(x-M_x)^2}{2\sigma_x^2}} dx, \qquad (11)$$

определяются двумя параметрами: M_x – математическим ожиданием и σ_x^2 – дисперсией.

Отметим некоторые свойства нормального закона распределения.

1. Кривая плотности распределения симметрична относительно значения M_x, называемого иногда центром распределения.

2. При больших значениях σ_x^2 кривая f(x) более пологая, т.е. σ_x^2 является мерой величины рассеивания значения случайной величины около значений M_x. При уменьшении параметра σ_x^2 кривая нормального распределения сжимается вдоль оси OX и вытягивается вдоль f(x).

3. Максимум ординаты кривой плотности распределения определяется выражением

$$f_{\max} = \frac{1}{\sqrt{2\pi\sigma_x^2}},$$

что при $\sigma_x^2 = 1$ соответствует значению примерно 0,4.

4. Для нормального распределения математическое ожидание, мода и медиана совпадают:

$$M_x = M_0 = M_e. \quad (23)$$

В ряде случаев рассматривается не сама случайная величина X, а ее отклонение от математического ожидания:

$$Y = X - M. \quad (24)$$

Такая случайная величина Y называется центрированной.

Отношение случайной величины X к ее среднему квадратичному отклонению

$$V = \frac{X}{\sigma_x} \quad (25)$$

называется нормированной случайной величиной.

Таким образом, центрированная случайная величина – разность между данной случайной величиной и ее математическим ожиданием, а нормированная случайная величина – отношение данной случайной величины к ее среднему квадратичному отклонению.

Очевидно, что математическое ожидание центрированной случайной величины равно нулю, $M_y = 0$, а дисперсия нормированной случайной величины равна единице, $\sigma_V^{\,2} = 1$.

Приведенная случайная величина – центрированная и нормированная случайная величина

$$Z = \frac{X - M_x}{\sigma_x} \quad (26)$$

Математическое ожидание и дисперсия приведенной случайной величины Z равны соответственно нулю, $M_z = 0$, и единице, $\sigma_z^2 = 1$.

Нормальное распределение с параметрами $M_z = 0$ и $\sigma_z^2 = 1$ называется **стандартным (нормированным**)

Для приведенной случайной величины нормальное стандартное распределение принимает вид

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz = \Phi(z), \quad (27)$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \phi(z). \quad (28)$$

Графики этих функций показаны на рис. 3 в, г, причем

$$\Phi(-z) = 1 - \Phi(z), \quad (29)$$

$$\varphi(-z) = \varphi(z). \quad (30)$$

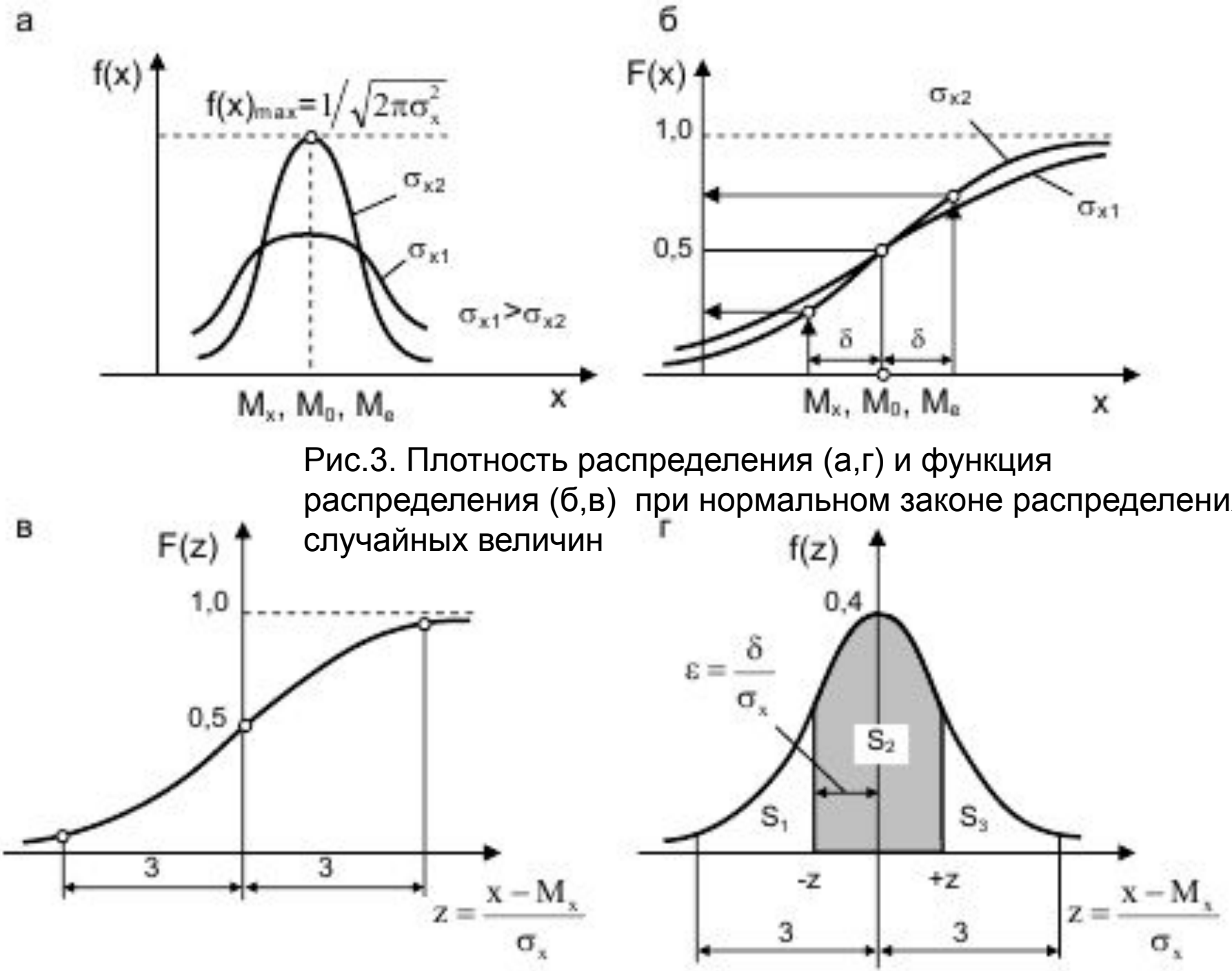
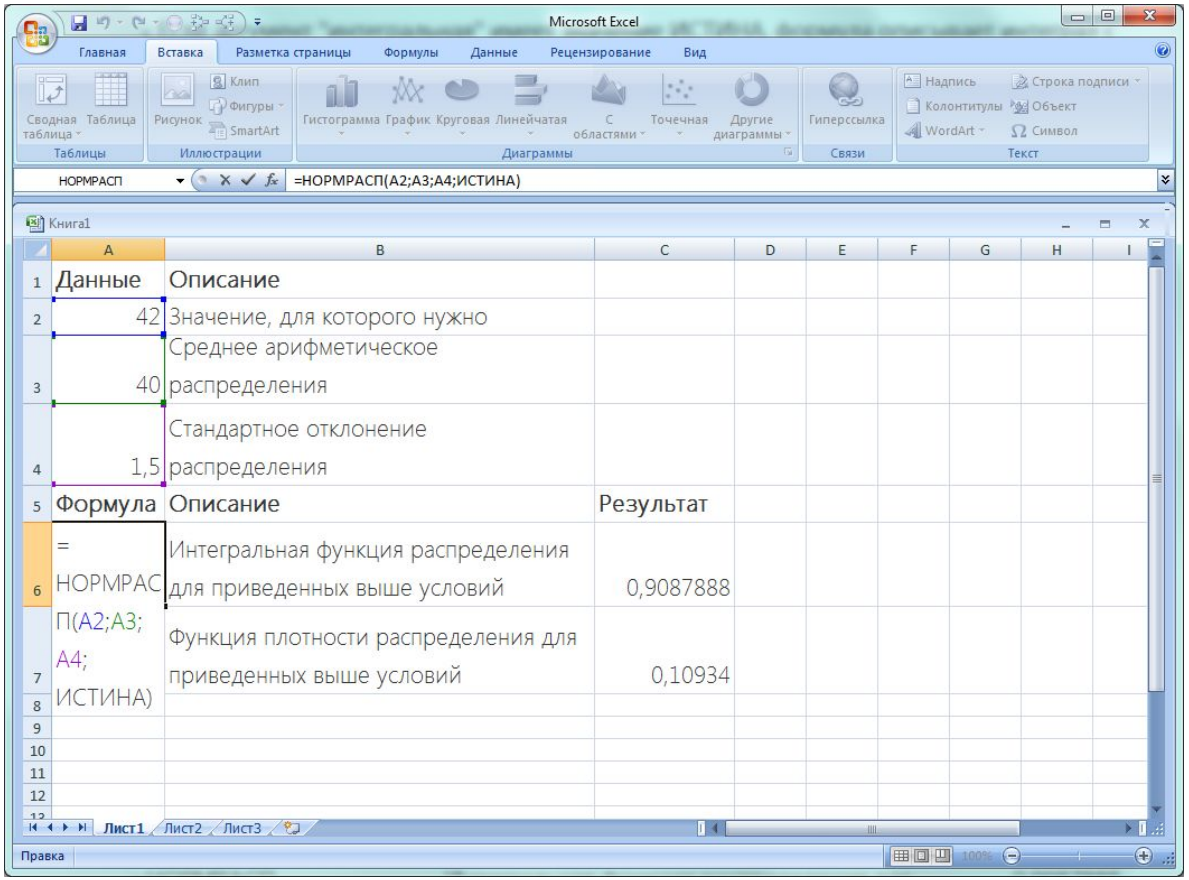


Рис.3. Плотность распределения (а,г) и функция распределения (б,в) при нормальном законе распределения случайных величин

Значения нормированной функции (27) нормального распределения (функции Лапласа) и значения плотности нормированного нормального распределения (28) табулированы и приведены в различных учебниках и справочниках по математической статистике

В списке статистических функций электронных таблиц Microsoft Excel им соответствуют НОРМРАСП(x; 0; 1; ИСТИНА) или НОРМСТРАСП(z) – для (27) и НОРМРАСП(x; 0; 1; ЛОЖЬ) – для (28).



Геометрически функция Лапласа представляет площадь под кривой f(z) в интервале от $-\infty$ до некоторой конкретной величины z.

Заметим, что иногда вместо функции Φ(z) табулируется функция $\Phi_0(z)$:

$$\Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{z^2}{2}} dz,$$

равная площади под графиком стандартного нормального распределения от 0 до z (см. рис. 3, г).

В силу симметрии

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{z^2}{2}} dz = 1/2.$$

Поэтому между функциями и существует простая зависимость

$$\Phi(z) = S + \Phi 0(z).$$

Функция $\Phi_0(z)$ нечетна:

$$\Phi_0(-z) = -\Phi_0(z).$$

В самом деле,

$$\Phi_0(-z) = \Phi(-z) - S = 1 - \Phi(z) - S = S - (1/2 + \Phi_0(-z)) = -\Phi_0(z).$$

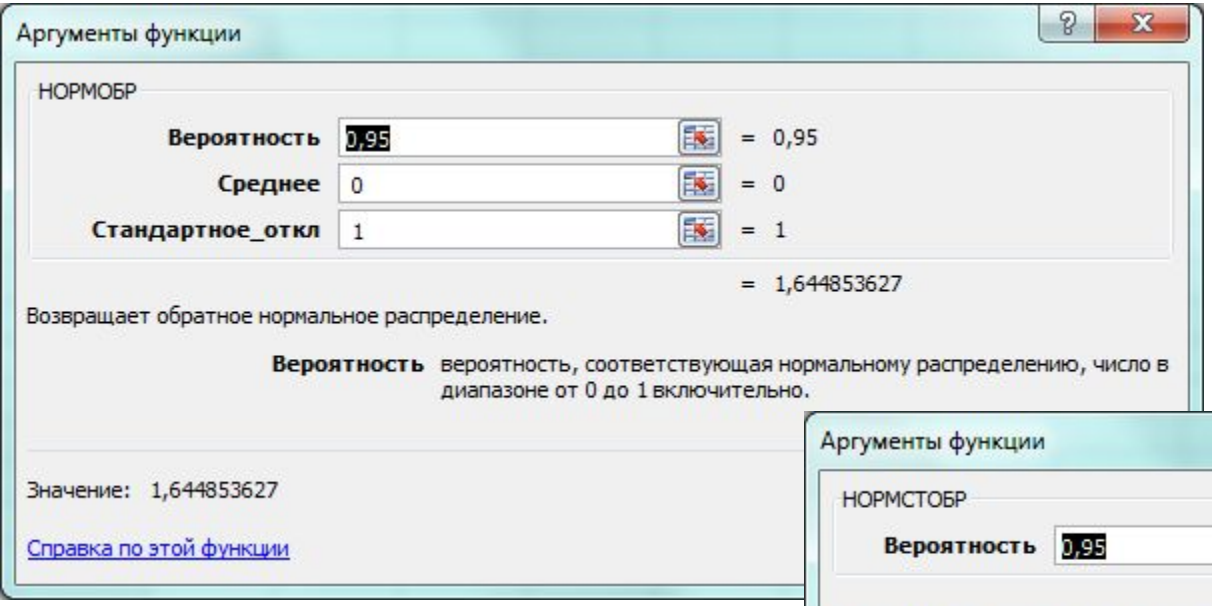
В соответствии с (19) квантиль z_p порядка p, нормированного нормального закона распределения - это такое значение приведенной случайной величины Z, для которого функция распределения (27) принимает значение P:

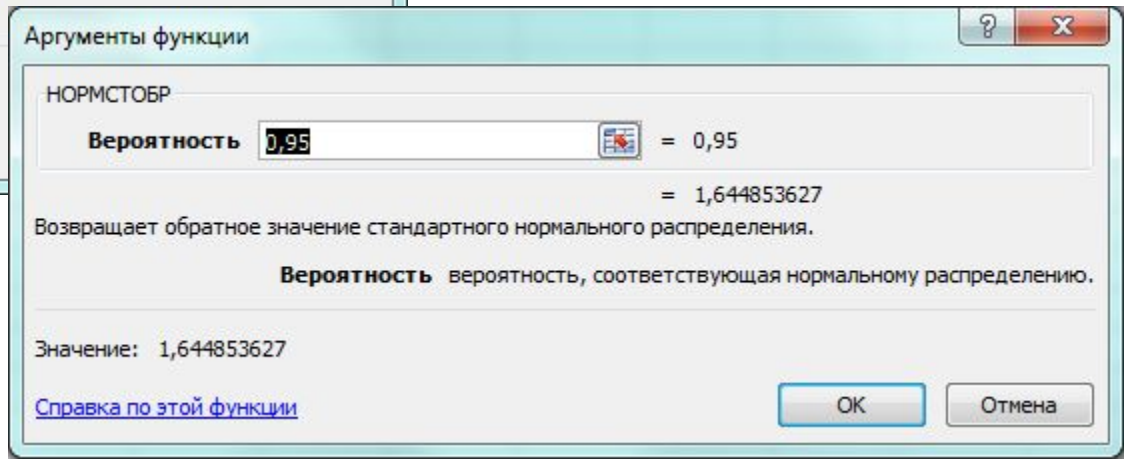
$$\Phi(z_p) = P. \qquad (31)$$

При определении квантили z_p необходимо решать задачу, обратную задаче определения значений функции Лапласа, т.е. по известному значению P этой функции (27) находить соответствующее ему значение аргумента z_p.

Для этого можно либо воспользоваться табулированными значениями функции Лапласа (например, поскольку Ф(1,64) = 0,94950, а Ф(1,65) = 0,9505, то z0,95 ≈ 1,645), либо воспользоваться таблицами для функции, обратной функции Лапласа, т.е. табулированными значениями квантилей нормированного нормального закона распределения.

Определение квантили z_p в электронных таблицах Microsoft Excel сводится к вычислению статистической функции НОРМОБР(P; 0; 1) или НОРМСТОБР(P) (например, НОРМОБР(0,95; 0; 1) = НОРМСТОБР(0,95) = 1,644853).



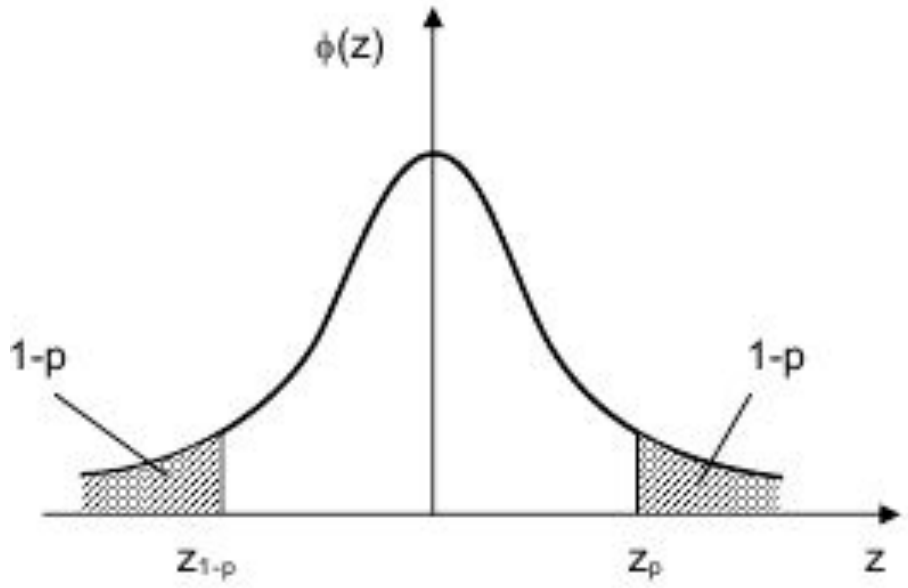


Для квантили стандартного нормального распределения справедливо следующее равенство:

$$z_{1-p} = -z_p. \qquad (27)$$

Рассмотрим график плотности стандартного нормального распределения (рис.4). Площадь под графиком левее квантили z_p по определению равна p. Значит, площадь правее этой точки равна 1 – p. Такая же площадь расположена левее точки z_{1-p} Итак, площади левее z1 –p и правее z_p равны. Поскольку график симметричен относительно оси ординат, из этого следует, что эти точки расположены на одинаковом расстоянии от нуля.

Рис.4. Квантиль стандартного нормального распределения



Зная квантиль z_p порядка p нормированного нормального закона распределения ($M_z = 0$ и $\sigma_z^2 = 1$), всегда можно найти квантиль x_p соответствующего порядка p для нормального распределения с произвольными параметрами σ_z^2

Поскольку

$$F(x_p) = P(X \le x_p) = P\left(\frac{X - M_x}{\sigma_x} \le \frac{x_p - M_x}{\sigma_x} \right) =$$

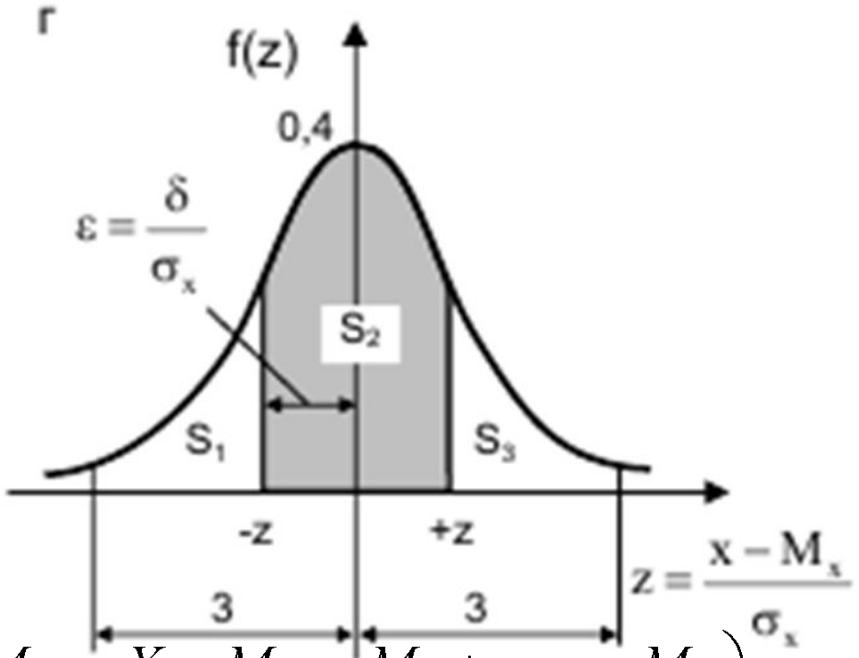
$$P\left(Z \le \frac{x_p - M_x}{\sigma_x} \right) = \Phi\left(\frac{x_p - M_x}{\sigma_x} \right) = P = \Phi(z_p),$$

то

$$\frac{x_p - M_x}{\sigma_x} = z_p$$

и, следовательно, $x_p = M_x + z_p \sigma_x$. (62a)

В ряде случаев важно знать вероятность того, что случайная величина X, подчиняющаяся нормальному закону распределения, не будет отличаться от своего математического ожидания M_x больше чем на величину $\pm\delta = \varepsilon \cdot \sigma_x$ (см.рис.3,г).

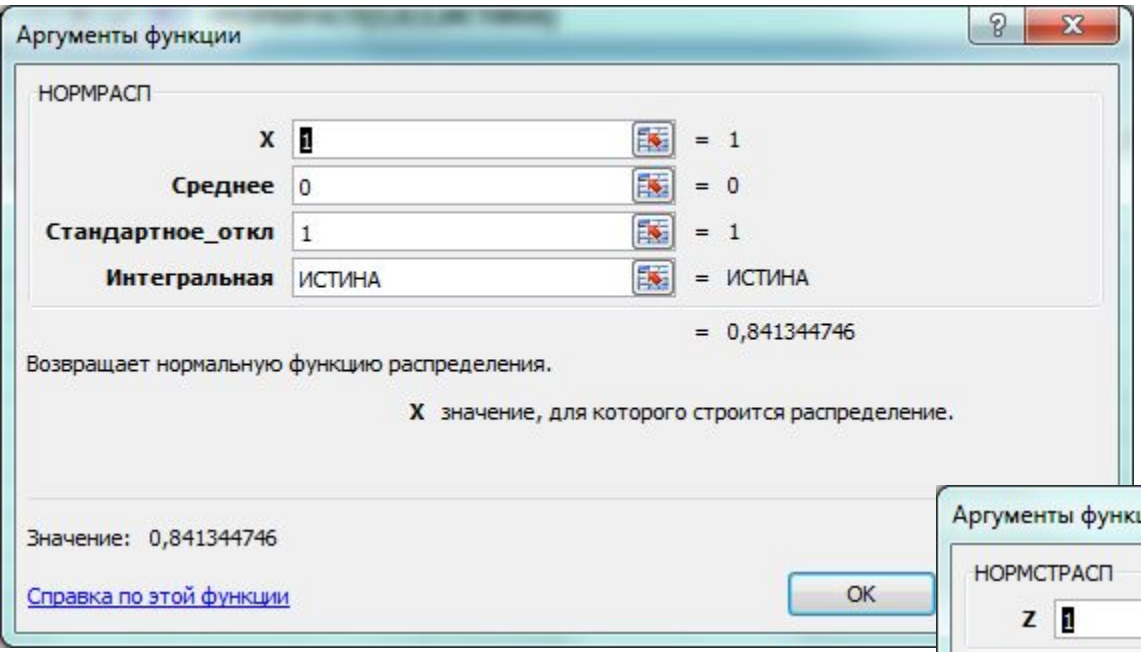


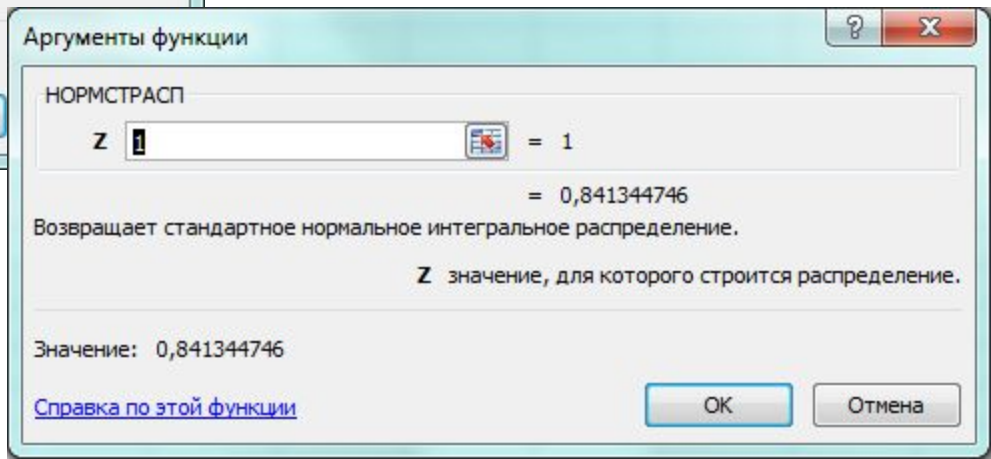
$$P(M_x - \varepsilon < X \le M_x + \varepsilon) = P\left(\frac{M_x - \varepsilon\sigma_x - M_x}{\sigma_x} < \frac{X - M_x}{\sigma_x} \le \frac{M_x + \varepsilon\sigma_x - M_x}{\sigma_x}\right) =$$

$$P(-\varepsilon < Z \le +\varepsilon) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\varepsilon} e^{-\frac{z^2}{2}} dz - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\varepsilon} e^{-\frac{z^2}{2}} dz = \Phi(\varepsilon) - \Phi(-\varepsilon),$$

$$= \Phi(\varepsilon) - (1 - \Phi(\varepsilon)) = 2\Phi(\varepsilon) - 1.$$

Так, при $\delta = \sigma_x$ ($\varepsilon = 1$) получаем, что $P(M_x + \sigma_x < X \le M_x - \sigma_x) = 2\Phi(1) - 1$, а поскольку по таблицам $\Phi(1) = 0{,}84135$ (или в Microsoft Excel НОРМРАСП(1;0;1; ИСТИНА) = НОРМСТРАСП(1) = 0,84135), то для случайной величины с нормальным законом распределения вероятность того, что
она примет такое значение, которое не будет отличаться от ее математического ожидания более чем на одно среднее квадратическое отклонение, равна $2 \cdot 0{,}84135 - 1 = 0{,}68$. Иными словами, при нормальном распределении примерно 2/3 всех значений случайной величины (отклика) лежит в интервале $M_x \pm \sigma_x$.





Аналогично можно подсчитать, что интервалу $M_x \pm 1{,}96\sigma_x \approx M_x \pm 2\sigma_x$ соответствует вероятность 0,95 (Ф(1,96) = 0,975002), а интервалу $M_x \pm 3\sigma_x$ - 0,997 (Ф(3) = 0,99865)

.

Отметим дополнительно, что 90% значений случайной величины лежат в диапазоне $M_x \pm 1{,}64\sigma_x$ (Ф(1,64) = 0,949497).

Следовательно, отличие какого-либо из значений случайной величины с нормальным законом распределения от ее математического ожидания не превосходит утроенного среднего квадратичного отклонения с вероятностью 0,997.

Это свойство в математической статистике носит название «правило трех сигм».

Чем больше величина интервала $M_x \pm \delta$, тем с большей вероятностью случайная величина X попадает в этот интервал.

Рассмотрим небольшой пример.

Пример 2. Предположим, что математическое ожидание содержания серы в угле равно M_S=0,6%, а среднеквадратичное отклонение σ_S=0,15%.

В этом случае мы можем быть уверены в том, что величина фактически измеренного значения процентного содержания серы в угле будет находиться в интервалах:

$0,6 \pm 1,00 \cdot 0,15 = 0,6 \pm 0,15$ с вероятностью 0,68;
$0,6 \pm 1,64 \cdot 0,15 = 0,6 \pm 0,25$ с вероятностью 0,90;
$0,6 \pm 1,96 \cdot 0,15 = 0,6 \pm 0,29$ с вероятностью 0,95;
$0,6 \pm 3,00 \cdot 0,15 = 0,6 \pm 0,45$ с вероятностью 0,997,

т.е. из 1000 проб только 3 пробы по содержанию серы в угле будут выходить из диапазона от 0,15 до 1,05%.

Заметим, однако: при рассмотрении примера 2 мы предполагали, что процентное содержание серы в угле не противоречит нормальному закону распределения, а также то, что нам изначально были известны математическое ожидание M_x и среднеквадратичное отклонение σ_x этой случайной величины, т.е. было выполнено большое (в пределе бесконечное) число измерений.

Как же работать со случайными величинами в реальных условиях проведения эксперимента, когда число измерений весьма ограничено?

К рассмотрению методологии решения подобных задач мы и перейдем в следующих лекциях.