3.4. СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ВЫБОРОЧНЫХ ЗНАЧЕНИЙ ПРОГНОЗНЫХ МОДЕЛЕЙ

До сих пор мы рассматривали способы построения прогнозных моделей стационарных процессов, не учитывая одной весьма важной особенности. Все используемые статистические данные являются выборочными значениями. Они дают некоторое представление о сути процесса, также как и построенные модели. Но это представление является приближённым, поскольку полное представление о стационарном процессе нам могут дать только характеристики его генеральной совокупности, а не выборки из неё. Таким образом, те точечные оценки, которые получались во всех предыдущих построениях, являются лишь приближением к истине, но не истиной. Но что делать, если мы не в состоянии собрать и обработать данные всей генеральной совокупности, а вынуждены работать лишь с выборкой из неё размера N? Выход здесь только один – оценить, насколько полученные точечные значения статистических характеристик прогнозируемого процесса дают представление о генеральной совокупности в целом. Логика здесь очевидна – чем большее число выборочных значений мы подвергаем статистической обработке, тем более точно их обобщающие характеристики представляют свойства исследуемого процесса. Следовательно, можно оценить вероятность того, насколько точны статистические характеристики и важнейшим показателем этого выступает размер выборки N.

Из всего многообразия возможных проявлений случайности, чаще всего в экономической практике она проявляется в форме нормального закона распределении вероятностей. Именно это положение и лежит в основе рассматриваемой процедуры.

Прежде всего, необходимо вспомнить, что функция распределения вероятностей случайной величины $F(x)$ определяет вероятность того, что случайная величина X при испытании примет значение, меньшее произвольно изменяемого действительного числа x ($-\infty<x<+\infty$):

$$F(x) = P(X < x). \qquad (3.4.1)$$

Эта функция положительна и меньше единицы. Для непрерывно возрастающего x график этой функции является возрастающим. На рис.3.4 построен график функции нормального распределения вероятностей. Если взять на оси x этого графика две произвольные точки x_0 и $x_0+\Delta x$, то их ординатами соответственно будут $F(x_0)$ и $F(x_0+\Delta x)$. Для приращения функции распределения вероятностей на этом участке имеем:

$$F(x_0 + \Delta x) - F(x_0) = P(x_0 < X < x_0 + \Delta x). \qquad (3.4.2)$$

Первая производная функции распределения вероятностей получила название плотности вероятности и имеет вид:

$$\varphi(x) = \lim_{\Delta x \to 0} \frac{F(x_0 + \Delta x) - F(x_0)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x_0 < X < x_0 + \Delta x)}{\Delta x}, \qquad (3.4.3)$$

из чего следует, что плотность вероятности представляет собой предел отношения вероятности того, что случайная величина X примет значение, лежащее в границах от x_0 до $x_0+\Delta x$, к величине интервала Δx, когда этот интервал стремится к нулю.

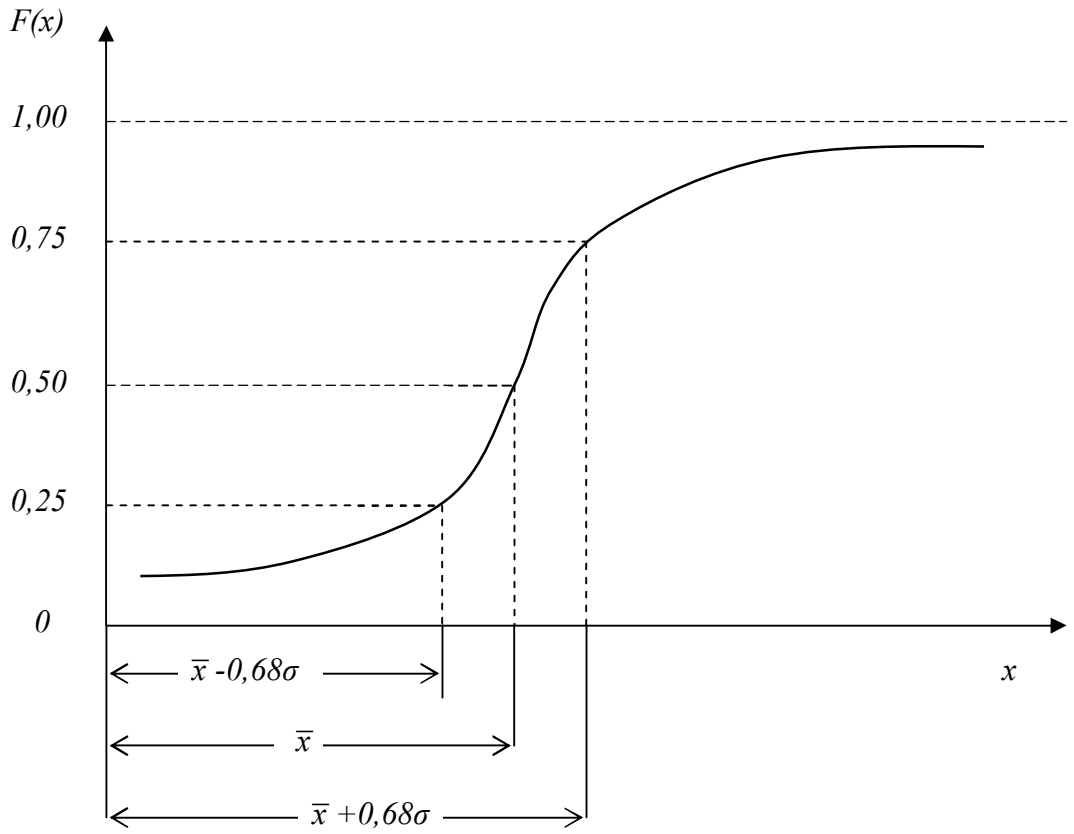

Рис. 3.4. График функции нормального распределения

Функция распределения вероятностей является первообразной функцией по отношению к функции плотности вероятности, поэтому вероятность (3.4.2) того, что случайная величина X примет значение, лежащее в границах от x_0 до $x_0+\Delta x$ может быть найдена так:

$$P(x_0 < X < x_0 + \Delta x) = F(x_0 + \Delta x) - F(x_0) = \int_{x_0}^{x_0+\Delta x} \varphi(x)dx\,. \qquad (3.4.4)$$

На графике плотности вероятности полученная вероятность (3.4.4) будет представлять собой площадь криволинейной трапеции с основанием от x_0 до $x_0+\Delta x$, ограниченную сверху кривой плотности вероятности (рис. 3.5).

Иногда функцию распределения вероятностей $F(x)$ называют «интегральной функцией распределения», а функцию плотности вероятности $\varphi(x)$ называют «дифференциальной кривой распределения», исходя из их математического смысла.

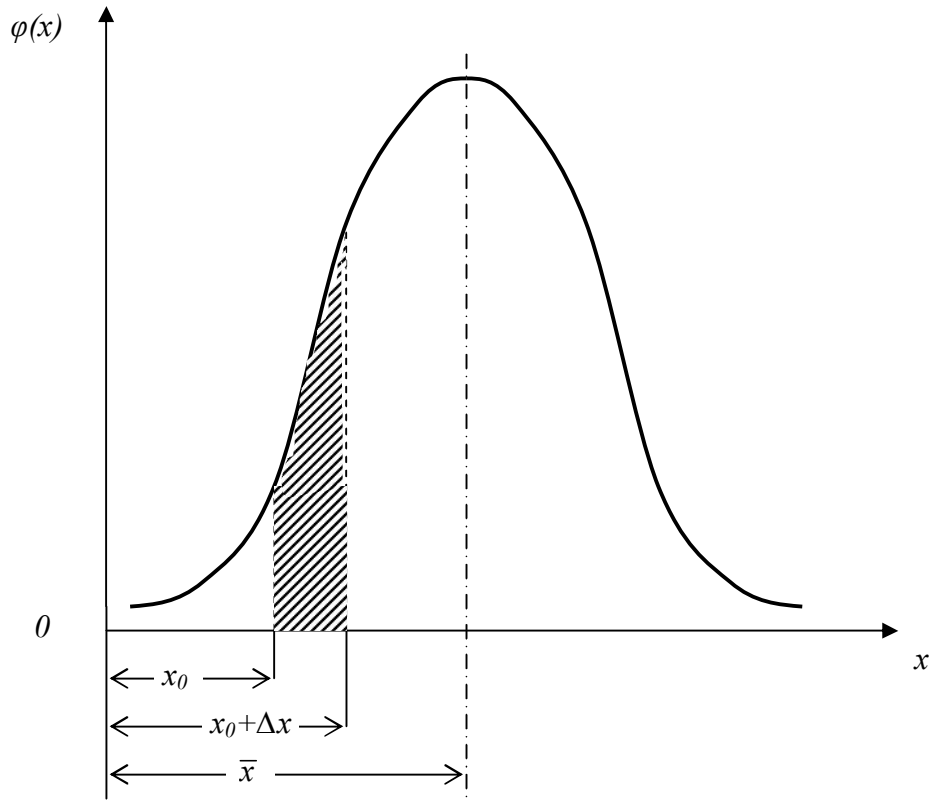

Рис. 3.5 График плотности вероятности нормального распределения

В математике среди элементарных функций известна функция Гаусса, которая имеет вид:

$$Y = e^{-x^2}. \qquad (3.4.5)$$

Эта функция, как легко заметить, симметрична относительно нулевого значения *x*, это во-первых, во-вторых, она всегда положительна, а в-третьих, она принимает своё максимальное значение, равное единице в том случае, когда *x=0*. По своему виду эта функция как нельзя лучше подходит для описания графика плотности вероятности нормального распределения (рис. 3.5), что и дало возможность Гауссу предложить функцию, аппроксимирующую нормальный закон распределения вероятностей, и носящую его имя:

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-M(x)}{\sigma})^2}. \qquad (3.4.6)$$

Характер функции, как легко заметить, определяется двумя характеристиками – дисперсией σ^2 и математическим ожиданием *M(x)*. Увеличение математического ожидания *M(x)* приводит к сдвигу кривой вправо вдоль оси *0x*, а её уменьшение – к сдвигу влево. С возрастанием дисперсии максимальная ордината нормальной кривой убывает, а сама кривая становится более пологой.

Так как, варьируя эти параметры можно получить любое семейство кривых, то следует взять за основу функцию плотности вероятности при каких-то фиксированных стандартных значениях, а затем – подставлять в неё эти две характеристики. Именно так и поступил в своё время Лаплас. Для этого он принял *M(x)=0*, то есть ситуацию, когда график рис. 3.5 симметричен относительно нулевого значения на оси *x*, и равенство единице дисперсии распределения. Тогда получается нормированная кривая плотности распределения:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} . \qquad (3.4.7)$$

К этому виду можно привести любой ряд *x*, для чего следует от каждого значения ряда отнять его математическое ожидание *M(x)*, а затем полученные значения разделить на среднее квадратичное отклонение σ:

$$z = \frac{x - M(x)}{\sigma} . \qquad (3.4.8)$$

Такой ряд будет называться нормированным.

Больший интерес, чем функция (3.4.7), представляет её первообразная, которая характеризует вероятность того, что *X* лежит в интервале от нуля до некоторого значения *Z*:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^Z e^{-\frac{1}{2}z^2} dz . \qquad (3.4.9)$$

Эта функция получила название нормированной функции Лапласа. Подставляя в неё различные значения *x,* можно получить разные значения вероятностей. Эта работа и была в своё время выполнена, а расчётные значения вероятности сведены в соответствующие таблицы, которые можно встретить в любом учебнике по теории вероятностей и математической статистике. Очевидно, что для условия *x=0*, то есть для ситуации, когда выборочное значение *X* точно соответствует математическому ожиданию процесса, вероятность будет равна нулю. А вот уже вероятность того, что *X* лежит в интервале от нуля до значения *Z=0,01* не равна нулю. Подставляя это значение в функцию (3.4.9), или просто заглянув в соответствующую строку таблицы, получим вероятность, равную 0.0040. А вот вероятность того, что для нормированной величины её значение окажется в интервале от нуля до значения *Z=5,00* равна 0,4999997, то есть такая вероятность очень высока.

Таким образом, можно увидеть, что имеется возможность оценить то, с какой вероятностью выборочное значение попадёт в заданный интервал от нуля до *Z*.

Можно, конечно, на практике каждый имеющийся ряд *x* отцентрировать относительно его математического ожидания *M(x)*, а затем полученные значения разделить на среднее квадратичное отклонение σ. Но значительно удобнее, воспользовавшись функцией Лапласа, привести формулу (3.4.4) к такому виду, чтобы при нормальном распределении можно было сразу определить вероятность того, что случайная величина *X* примет значение,

лежащее в границах от x_0 до $x_0+\Delta x$. Для этого воспользуемся очевидным равенством:

$$P(x_0 < X < x_0 + \Delta x) = \int_{x_0}^{x_0+\Delta x} \varphi(x)dx = \int_{0}^{x_0+\Delta x} \varphi(x)dx - \int_{0}^{x_0} \varphi(x)dx . \qquad (3.4.10)$$

Для того чтобы применить функцию Лапласа, определим из (3.4.8) исходную переменную x: $x=\sigma z+M(x)$, а $dx=\sigma dz$. Теперь можно найти новые пределы интегрирования: если $x=x_0+\Delta x$, то в соответствии с (3.4.8) $z=(x_0+\Delta x-M(x))/\sigma$, а если $x=x_0$, то $z=(x_0 -M(x))/\sigma$. Теперь, подставляя в (3.4.10) функцию Лапласа с этими пределами интегрирования, получим:

$$P(x_0 < X < x_0 + \Delta x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x_0+\Delta x-M(x)} e^{-\frac{z^2}{2}} dz - \frac{1}{\sqrt{2\pi}} \int_{0}^{x_0-M(x)} e^{-\frac{z^2}{2}} dz = \qquad (3.4.11)$$

$$\Phi(\frac{x_0 + \Delta x - M(x)}{\sigma}) - \Phi(\frac{x_0 - M(x)}{\sigma})$$

Пусть теперь необходимо решить задачу нахождения вероятности того, что выполняется неравенство: $|X-M(x)|<\delta$. Это неравенство равносильно двойному неравенству: $M(x)- \delta <X< M(x)+\delta$, что, как легко заметить, даёт формулировку в терминах задачи (3.4.11), поэтому решение этой задачи легко найти с помощью (3.4.11):

$$P(|X - M(x)| < \delta) = \Phi(\frac{(M(x)+\delta) - M(x)}{\sigma}) - \Phi(\frac{(M(x)-\delta) - M(x)}{\sigma}) = \qquad (3.4.12)$$

$$\Phi(\frac{\delta}{\sigma}) - \Phi(-\frac{\delta}{\sigma}) = 2\Phi(\delta/\sigma)$$

Так как δ – некоторое наперёд заданное число, его можно задавать различными способами. В частности, можно задать как некоторую линейную функцию от среднеквадратичного отклонения, например, так: $\delta=\sigma t$. Откуда $t=\delta/\sigma$. Пусть, например, $t=3$. Тогда вероятность того, что отклонение случайной величины X от его математического ожидания $M(x)$ по абсолютной величине будет меньше $\delta=3\sigma$ равна:

$$P(|X - M(x)| < 3\sigma) = 2\Phi(t) = 2\Phi(3) = 0,9973 ,$$

что известно в математической статистике под правилом трёх сигм.

Неравенство $|X-M(x)|< \sigma t$ равносильно не только двустороннему неравенству : $M(x)- \sigma t <X< M(x)+ \sigma t$, но и другому двустороннему неравенству, а именно:

$$X - \sigma t < M(x) < X+ \sigma t. \qquad (3.4.13)$$

Здесь следует обратить внимание вот на что. В последнем двойном неравенстве неизвестно математическое ожидание и среднеквадратичное отклонение. Для того, чтобы продолжить дальше, необходимо вспомнить некоторые свойства дисперсии и, соответственно, среднеквадратичного отклонения.

Дисперсия, как известно, представляет собой меру отклонений случайной величины от его математического ожидания. Применительно к дискретному случаю (а именно его мы рассматриваем в нашей дисциплине), дисперсия может быть записана так:

$$\sigma^2 = M(X - M(x))^2 . \qquad (3.4.14)$$

Для того чтобы вычислить дисперсию, необходимо иметь все значения случайной переменной *X* и точно знать величину её математического ожидания *M(x)*. На практике эти значения не известны, а известны лишь некоторые выборочные значения *x* и их средняя арифметическая $\bar{x}$. Покажем, как, используя эти значения, можно определить пределы нахождения математического ожидания случайной величины (3.4.13). Вспомним, что дисперсия двух независимых случайных величин равна сумме дисперсий этих величин. Поэтому дисперсию (3.4.14) можно представить так:

$$\sigma^2 = M(X - \bar{x} + \bar{x} - M(x))^2 = M(X - \bar{x})^2 + M(\bar{x} - M(x))^2 = s^2 + \sigma(\bar{x})^2 . \qquad (3.4.15)$$

где s^2 – дисперсия случайных величин относительно их средней арифметической,

$\sigma(\bar{x})^2$ - дисперсия средней арифметической относительно математического ожидания.

Дисперсия случайных величин относительно их средней арифметической вычисляется легко, а что представляет собой $\sigma(\bar{x})^2$ - дисперсия средней арифметической относительно математического ожидания? Это квадрат отклонения средней арифметической от математического ожидания случайной величины. Для нормального распределения вероятностей средняя арифметическая является лучшей оценкой математического ожидания, причём, чем большее число наблюдений включается в расчёт средней арифметической, тем ближе находится средняя арифметическая к математическому ожиданию. Из этого со всей очевидностью следует, что с ростом числа наблюдений *n* дисперсия средней арифметической относительно математического ожидания стремится к нулю. Чтобы понять, как это происходит, вычислим дисперсию средней арифметической *n* одинаково распределённых взаимно независимых случайных величин:

$$\sigma(\bar{x})^2 = M(\bar{x} - M(x))^2 = M(\frac{\sum_{i=1}^{n} X_i - nM(x)}{n})^2 = \frac{1}{n} M(X - M(x))^2 = \frac{1}{n}\sigma^2 . \qquad (3.4.16)$$

Таким образом, дисперсия средней арифметической относительно математического ожидания в *n* раз меньше дисперсии всей генеральной совокупности.

Из равенств (3.4.15) и (3.4.16) получим:

$$\sigma^2 = s^2 + \sigma(\bar{x})^2 = s^2 + \frac{1}{n}\sigma^2 \rightarrow \sigma^2 = \frac{n}{n-1} s^2 . \qquad (3.4.17)$$

Итак, если у нас есть дисперсия случайных величин относительно их средней арифметической s^2, то, с помощью формулы (3.4.17) можно оценить общую дисперсию.

Теперь у нас есть все основания для решения следующей задачи: оценить возможный интервал значений математического ожидания случайной величины по известному значению числа наблюдений *n* и среднему арифметическому $\bar{x}$. Воспользовавшись (3.4.12), получим, заменив X на $\bar{x}$ и общую дисперсию σ^2 на дисперсию средней арифметической относительно математического ожидания $\sigma(\bar{x})^2$:

$$P(|\bar{x} - M(x)| < \delta) = \Phi(\delta / \sigma(\bar{x})). \quad (3.4.18)$$

Так как в соответствии с (3.4.16) $\sigma(\bar{x})^2 = \sigma^2 / n$, а ранее мы рассмотрели замену $\delta = \sigma t$, которая для случая средней арифметической примет вид: $\delta = t\sigma/\sqrt{n}$, то (3.4.18) можно записать в другой форме, а именно:

$$P(|\bar{x} - M(x)| < t\sigma/\sqrt{n}) = 2\Phi(t(\sigma/\sqrt{n})/(\sigma/\sqrt{n})) = 2\Phi(t). \quad (3.4.19)$$

Так как неравенство $|\bar{x} - M(x)| < t\sigma/\sqrt{n}$ равносильно не только двустороннему неравенству : $M(x) - t\sigma/\sqrt{n} < \bar{x} < M(x) + t\sigma/\sqrt{n}$, но и другому двустороннему неравенству, а именно: $\bar{x} - t\sigma/\sqrt{n} < M(x) < \bar{x} + t\sigma/\sqrt{n}$, получим вероятность того, что математическое ожидание случайной нормально распределённой величины лежит в пределах, определяемых средней арифметической и дисперсией:

$$P(\bar{x} - t\sigma/\sqrt{n} < M(x) < \bar{x} + t\sigma/\sqrt{n}) = 2\Phi(t) = \alpha. \quad (3.4.20)$$

Как следует из (3.4.17) при достаточно большом числе наблюдений n выборочная дисперсия s^2 практически равна общей дисперсии σ^2, поэтому можно утверждать, что с заданной доверительной вероятностью α математическое ожидание случайной величины лежит в пределах:

$$\bar{x} - ts/\sqrt{n} < M(x) < \bar{x} + ts/\sqrt{n}. \quad (3.4.21)$$

Но при малых выборках ($n<30$), с которыми в основном и приходится иметь дело в прогнозировании социально-экономических процессов, выборочная дисперсия отличается от общей (3.4.17), поэтому такая замена не совсем корректна. Более того, введение поправочного коэффициента (3.4.17), что кажется вполне логичным, не меняет ситуацию, потому что при малых выборках выборочное значение дисперсии ведёт себя иначе, чем это следовало бы из нормального закона распределения. В результате и нормированный ряд (3.4.8), в котором вместо дисперсии подставляется её выборочное значение, а вместо случайной переменной – средняя арифметическая:

$$\frac{\bar{x} - M(x)}{s}\sqrt{n} = t, \quad (3.4.22)$$

не будет распределён нормально и к нему функция Лапласа неприемлема.

Английский статистик В.Госсет предложил описывать распределение величины (3.4.22) близким по форме к нормальному, которое получило название «распределение Стьюдента», поскольку именно под этим псевдонимом В.Госсет и опубликовал соответствующие материалы. Это распределение также симметрично, как и функция Лапласа, и имеет такую же форму, но её максимум несколько меньше, а с увеличением величины t функция более пологая, чем функция Лапласа нормального распределения. С увеличением числа наблюдений n распределение случайной переменной (3.4.22) стремится к нормальному и при $n>30$ практически совпадает с ним.

Плотность распределения величины t, определяемой как (3.4.22), определяется только одним параметром – количеством наблюдений $n=k+1$:

$$s_k = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{k\pi}}(1 + \frac{t^2}{k})^{-\frac{k+1}{2}}, \quad (3.4.23)$$

где *Г(y)* – Гамма функция Эйлера в точке *y*,

k = n-1 – величина, получившая название «число степеней свободы».

С помощью этой плотности распределения случайной величины *t* можно рассчитать вероятность *1-q* того, что истинное значение нормированной переменной *t* лежит в пределах от минус $t_{q,k}$ до плюс $t_{q,k}$:

$$P(-t_{q,k} < t < t_{q,k}) = 1 - q \,. \qquad (3.4.24)$$

Эта вероятность получила название «доверительной вероятности».

Подставляя в это равенство значения *t,* взятые из (3.4.22), получим, что для выборочных значений средней арифметической и дисперсии с доверительной вероятностью *1-q* математическое ожидание случайной величины лежит в пределах:

$$\overline{x} - t_{q,k}\frac{s}{\sqrt{n}} < M(x) < \overline{x} + t_{q,k}\frac{s}{\sqrt{n}} \,. \qquad (3.4.25)$$

Здесь множитель $t_{q,k}$ рассчитывается по функции (3.4.23) или берётся из таблиц *t*-статистики Стьюдента. В таблицах эта величина выбирается, исходя из числа степеней свободы *k=n-1* и доверительной вероятности *1-q*. Иногда в таблицах приводится только значение *q*, что нисколько не затрудняет поиск в ней $t_{q,k}$.