Глава 2

Математика вероятностей

Наука начинается, когда сущности, о которых мы говорим, становятся величинами, значения которых можно измерить. Числа – это фундамент любой теории, они же являются её конечной целью.

В физике и химии процесс измерения не может быть сделан сколь угодно точным. В биологии и социальных науках ситуация ещё более сложная. При проведении однотипных наблюдений получаются различные результаты, какими бы точными "приборами" и методиками мы не пользовались. Социальная, экономическая среда постоянно изменяется, являясь в известной мере субъективной и подверженной многочисленным внешним и внутренним случайным факторам. Поэтому необходимо уметь нейтрализовать неизбежную неточность эмпирических исследований.

"Вероятность" часто оказывается синонимом к слову "будущее". Значение цены финансового инструмента через год, завтра, или даже через одну минуту существует только в виде спектра возможностей, каждая из которых *может* реализоваться. Методы математики вероятностей лежат в основе большинства современных финансовых теорий, которые учитывают случайный, стохастический характер поведения исследуемых объектов. Вероятностные закономерности, определяющие динамику цен на финансовых рынках, лежат в основе теории диверсификации, деривативов и методов прогнозирования.

В этой главе мы рассмотрим базовый стохастический "язык", который будет использоваться в дальнейшем. Основным объектом является случайная переменная, распределение её вероятностей, интегральные характеристики и корреляционные связи с другими величинами.

2.1 Среднее

Рассмотрим некоторую величину, различные наблюдения (измерения) которой дают нам набор чисел $x_1, x_2, ...$ Это могут быть значения ежедневных цен финансового инструмента или доходности некоторой торговой системы. Например, ежедневные логарифмические *изменения в процентах* американского фондового индекса S&P500 в течение недели до и после октябрьского краха 1987 года имели значения:

$$x = \quad \text{-0.54, 1.64, -3.00, -2.37, -5.30, } \mathbf{-22.9}\text{, 5.20, 8.71, -4.00, -0.01}$$

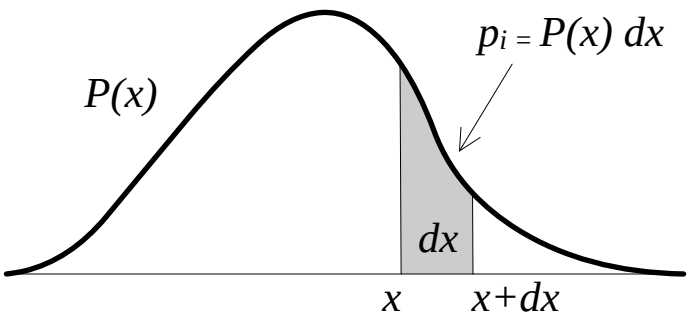
Числа $x_1, x_2, .., x_n$ можно рассматривать, как возможные реализации случайной величины x. На первом этапе исследования предполагается, что все x_i *независимы* друг от друга. Иными словами, мы не интересуемся порядком их поступления и можем эту последовательность произвольным образом перемешать.

• Пусть значение x_i встречается n_i раз, а общее количество чисел равно n. Мы называем *арифметическим средним*, или просто *средним* x величину:

$$\bar{x} = \langle x \rangle = \frac{1}{n} \sum_i x_i n_i = \sum_i x_i p_i = \int_{-\infty}^{\infty} x \; P(x) dx, \tag{2.1}$$

где $p_i = n_i/n$ - относительные *частоты* (или *вероятности*, стр.450) появления того или иного значения x_i. Если все x_i различны, то среднее равно их сумме, делённой на n. Чем вероятнее значение x_i, тем больший вклад оно даёт в среднее.

Большинство финансовых величин – это вещественные числа, поэтому удобно перейти к непрерывному пределу, в результате которого вместо суммы возникает интеграл. На рисунке заштрихованная область дает частоту того, что значение x будет оказываться в интервале от x до $x + dx$. Таким образом, $P(x)$ – это *плотность распределения вероятностей*:



Для получения вероятности p_i необходимо умножить функцию $P(x)$ на интервал dx. Интеграл (стр.494) – это сокращенное обозначение процедуры вычисления площади под кривой, равной сумме узких столбиков шириной dx.

• Вероятность того, что x принимает какое-либо значение, равна единице и соответствует полной площади под кривой плотности распределения:

$$\sum_i p_i = \int_{-\infty}^{\infty} P(x)dx = 1.$$

Это равенство называют *условием нормировки.*

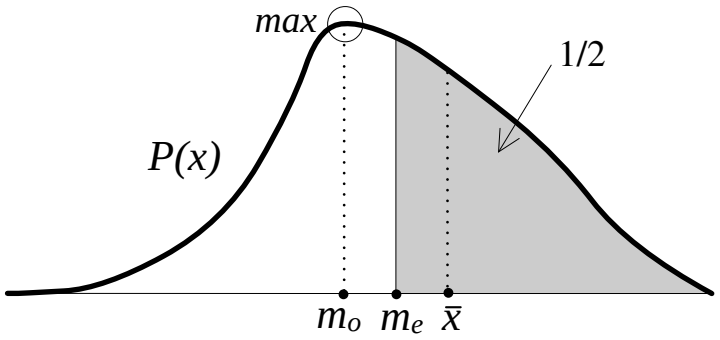
• *Медианным средним* случайной величины x является такое значение m_e, для которого число наблюдений, меньших медианного ($x < m_e$) и больших ($x > m_e$), одинаково. Поэтому слева и справа от m_e площади под кривой $P(x)$ совпадают и равны 1/2. Для вычисления медианного среднего в случае конечного набора n чисел x_1,.., x_n их необходимо отсортировать в порядке возрастания. Тогда медиана для нечётных n будет равна центральному числу, а для чётных – середине между двумя центральными числами. Для 10-ти изменений S&P500: $\bar{x} = -2.26\%$, а $m_e = -1.46\% = (-2.37 - 0.54)/2$.

$$-22.9, \quad -5.3, \quad -4.00, \quad \underline{-3.00, \quad \overline{-2.37, \quad -0.54}, \quad -0.01}, \quad 1.64, \quad 5.20, \quad 8.71$$

В данных, имеющих большие выбросы, медианное среднее зачастую более адекватно характеризует величину x, чем обычное арифметическое среднее. В русской версии Excel среднее вычисляется при помощи вызова функции СРЗНАЧ(), а для медианного среднего предназначена функция МЕДИАНА().

Возможна промежуточная между медианой и средним мера, когда отбрасывают четверть самых больших и самых маленьких значений, а по оставшимся вычисляют простое арифметическое среднее -1.48 = (-3.00-2.37-0.54-0.01)/4.

• Наиболее вероятное значение m_o называют *модой*. В точке моды плотность вероятности достигает своего максимального значения $P(m_o) = max$. Большинство распределений, встречающихся на практике, имеют не более одного максимума, и, соответственно, мода обычно хорошо определена:



В общем случае для не симметричного вокруг моды распределения $P(x)$ арифметическое среднее, медиана и мода имеют *различные* значения. Каждая из этих величин так или иначе характеризует случайные числа, их наиболее "*типичные*" значения.

• Плотность вероятности позволяет вычислить среднее любой функции $R(x)$:

$$\overline{R(x)} = \langle R(x) \rangle = \sum_i R(x_i) \cdot p_i = \int_{-\infty}^{\infty} R(x) \cdot P(x)\, dx. \qquad (2.2)$$

Например, если $R(x)$ равна доходу, получаемому при том или ином значении случайного параметра x, то формула (2.2) даёт нам величину среднего дохода как взвешенную сумму всех возможных доходов (стр. 450).

Мы *не будем различать* при обозначении среднего черту сверху или фигурные скобки, используя для коротких выражений черту, а для длинных – скобки. В математической литературе для среднего $\langle R \rangle$ используют также следующие обозначения: $\mathbf{E}(R)$ или $\mathbf{M}(R)$. Среднее является *функционалом*, так как, подставляя в интеграл или сумму ту или иную *функцию* $R(x)$, мы получаем (ставим ей в соответствие) *число*.

При вычислении среднего от функции случайной переменной x за знак усреднения можно выносить константу и разбивать среднее на слагаемые:

$$\langle \alpha f(x) \rangle = \alpha \langle f(x) \rangle, \qquad \langle f(x) + g(x) \rangle = \langle f(x) \rangle + \langle g(x) \rangle.$$

Но это и всё! Нелинейные функции, в общем случае, не могут быть вынесены из-под знака среднего (суммы или интеграла): $\langle x^2 \rangle \neq \langle x \rangle^2$, $\langle \sqrt{x} \rangle \neq \sqrt{\langle x \rangle}$.

• Это свойство среднего приводит к различным необычным эффектам, часто проявляющимся при проведении финансовых вычислений. Рассмотрим забавный *парадокс Сигеля* (Siegel, 1972). Предположим, что сегодня курс евро по отношению к доллару равен единице $x = €/\$ = 1$, и известно, что равновероятно он может вырасти до 2 или опуститься до $1/2$. В какой валюте целесообразно держать свои средства? Среднее значение будущего курса больше единицы:

$$\langle x \rangle = \sum_{i=1}^{2} x_i \cdot p_i = 2 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{5}{4} > 1,$$

поэтому евро выглядит предпочтительнее. Однако, с другой стороны, курс доллара к евро $1/x = \$/€$ также равновероятно примет значения $1/2$ и 2, поэтому его среднее значение тоже оказывается больше единицы. Так какая же из валют более доходна?

Причина возникновения парадокса состоит в том, что, в общем случае:

$$1/\langle x \rangle \neq \langle 1/x \rangle.$$

Ситуация осложняется, если сегодня необходимо определить "справедливое" значение форвардного курса с поставкой в будущем. Если его значение равно среднему будущих курсов, то форварды не удовлетворяют естественному соотношению $\langle €/\$ \rangle \cdot \langle \$/€ \rangle = 1$.

• Ещё одна необычная ситуация возникает при рассмотрении среднего от логарифма цены. Так как

$$\langle \ln x \rangle \;\neq\; \ln \langle x \rangle \,,$$

то стоимость некоторого финансового актива x может в среднем увеличиваться, в то время как логарифмическая доходность от его обладания быть отрицательной. Например, если начальная цена x_0 равновероятно в два раза увеличивается $2x_0$ или в 4 раза уменьшается $x_0/4$, несложно видеть (≪ стр.550), что средняя логарифмическая доходность отрицательна, хотя средняя стоимость актива растёт:

$$\left\langle \ln \frac{x}{x_0} \right\rangle = -\ln\sqrt{2} < 0, \qquad\qquad \langle x \rangle = \frac{9x_0}{8} > x_0.$$

Стоит ли инвестировать в такой актив?

• Отметим ещё *петербургский парадокс* (Бернули, 1713). Пусть будущие цены $x_k = 2^k \;=\; 2,\; 4,\; 8, ..$ финансового инструмента реализуются с вероятностями $p_k = 1/2^k = 1/2, 1/4, 1/8, ..$ Какая его текущая цена является справедливой и для покупателя и для продавца? Легко видеть, что $\langle x \rangle$ равно бесконечности. Однако вряд ли кто-либо готов заплатить один миллион за этот инструмент, так как с вероятностью $p_1 + .. + p_{19}$ =0.999998 он получит убыток ($x_{20} = 2^{20} \sim 10^6$).

• Происхождение подобных парадоксов связано с некритичным "абсолютизированием" среднего значения как *единственной* меры, характеризующей возможную реализацию случайного события. В Стохастическом Мире будущее многовариантно и определяется функцией распределения вероятности $P(x)$, а не одним числом $\langle x \rangle$. Например, две *различные* плотности вероятности будущей цены могут иметь *одинаковые* средние, однако это не означает, что они эквивалентны для инвестора. Напротив, для них могут быть различны медиана, мода или вероятность получить доход. Следовательно, существуют определённые критерии для выбора между этими инвестиционными возможностями.

Не менее важным является *неприятие риска* и желание чаще получать положительные эмоции. В результате сильно несимметричная плотность вероятности, дающая большие доходы с малой вероятностью, для большинства окажется менее предпочтительной, чем более симметричная, даже при одинаковых средних значениях. Имеет смысл сравнить собственное восприятие двух торговых стратегий, одна из которых равновероятно даёт доход 0 или 2, а вторая, тоже имея среднее равное 1, в девяти случаях из десяти приносит убыток -1, и только с вероятностью 0.1 доход 19.

Оставляя дальнейшие размышления философского характера Читателю, заметим, что лучше строить такие торговые стратегии, которые имеют симметричную плотность распределения, принося *заметную* прибыль независимо от способов вычисления среднего и других вероятностных критериев ☺.

2.2 Волатильность

Не смотря на то, что распределение вероятностей $P(x)$ полностью определяет свойства случайной переменной x, на практике оказывается удобнее пользоваться не всей функцией $P(x)$, а лишь несколькими интегральными параметрами, которые её достаточно полно характеризуют.

Наиболее важным (после среднего) параметром является среднеквадратичная ошибка σ, или, на финансовом жаргоне, *волатильность* (volatility) величины x:

$$\sigma^2 = D = \left\langle (x - \bar{x})^2 \right\rangle = \sum_i (x_i - \bar{x})^2 \, p_i = \int_{-\infty}^{\infty} (x - \bar{x})^2 \, P(x) dx, \qquad (2.3)$$

Сигма (волатильность) – это корень из суммы квадратов отклонения каждого x_i от среднего $\bar{x}$. Квадрат сигмы называют *дисперсией* D или *моментом 2-го порядка*. Волатильность характеризует степень "размазанности", ширины распределения, т.е. показывает, насколько сильно x_i могут на практике отклоняться от среднего $\bar{x}$. В Excel волатильность данных вычисляется при помощи вызова функции СТАНДОТКЛОН().

• Дисперсию можно вычислить ещё одним способом. Для этого раскроем квадрат в определении (2.3):

$$\left\langle (x - \bar{x})^2 \right\rangle = \left\langle x^2 - 2\bar{x} \cdot x + \bar{x}^2 \right\rangle = \left\langle x^2 \right\rangle - \left\langle 2\bar{x} \cdot x \right\rangle + \bar{x}^2.$$

Множитель $2\bar{x}$ является числом, и его всегда можно вынести за сумму (интеграл) усреднения, а так как обозначения для среднего $\langle x \rangle$ и $\bar{x}$ эквивалентны: $\left\langle 2\bar{x} \cdot x \right\rangle = 2\bar{x} \cdot \langle x \rangle = 2 \langle x \rangle^2$, то мы приходим к двум способам вычисления дисперсии:

$$\boxed{\sigma^2 = D = \left\langle (x - \bar{x})^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2} \, . \qquad (2.4)$$

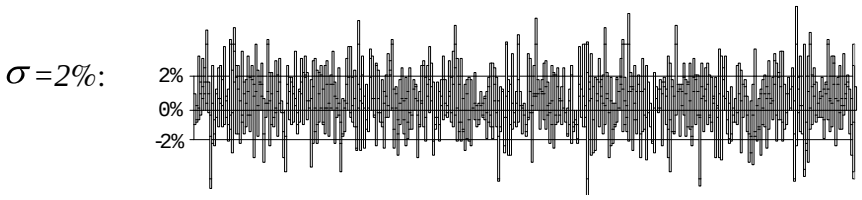
Дисперсия равна или среднему квадрату разности величины и её среднего, или разности среднего квадрата и квадрата среднего случайной величины ☺.

Уместно ещё раз обратить внимание на простой, но важный факт. Среднее квадрата и квадрат среднего – это две большие разницы:

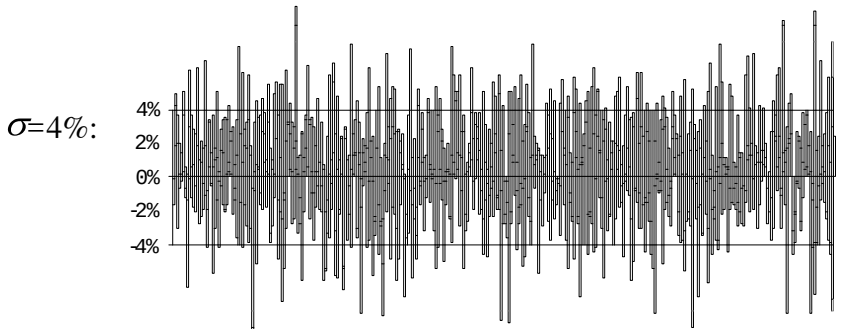
$$\left\langle x^2 \right\rangle = \sum_i x_i^2 p_i, \qquad \langle x \rangle^2 = \left(\sum_i x_i p_i \right)^2 .$$

Т.е. в первом случае мы сначала возводим в квадрат, а затем суммируем, а во втором – наоборот: $a^2 + b^2 \neq (a + b)^2$. Мы опять сталкиваемся с особенностями вычисления средних для нелинейных функций.

• Чтобы визуально представить себе финансовый смысл волатильности, рассмотрим ежедневные изменения в процентах цены некоторого финансового инструмента (акции, индекса) с небольшим средним и высокой волатильностью (именно это обычно и наблюдается на практике). Если волатильность $\sigma = 2\%$, то мы будем наблюдать следующие значения ежедневных доходностей:



Для более волатильной бумаги с $\sigma = 4\%$ диаграмма будет выглядеть значительно более "пушистой" или, в зависимости от мировосприятия, – "колючей":



Колючесть ценовой динамики часто отождествляют с рискованностью ценной бумаги. Мечта инвестора – это высокие средние доходности при низких волатильностях, другими словами, максимальность *отношения Шарпа*: $\bar{x}/\sigma$.

Если распределение вероятности симметрично и его максимум совпадает со средним, то волатильность иногда называют *стандартной ошибкой*, так как она характеризует наиболее типичный диапазон реализации случайной величины вокруг её среднего:

$$x = \bar{x} \pm \sigma = [\bar{x} - \sigma \ ... \ \bar{x} + \sigma].$$

Для сильно несимметричных распределений диапазон $\bar{x} \pm \sigma$ уже не столь "характерен". Например, для приведенного в предыдущем разделе примера с $x_i = \{-1, 19\}$ и $p_i = \{0.9, 0.1\}$ несложно получить $\langle x \rangle = 1$, а $\langle x^2 \rangle = 37$. Поэтому волатильность $\sigma = 6$. Однако диапазон 1 ± 6 мало что отражает для понимания возможных значений x_i.

Для симметричного случая с $x_i = \{0, 2\}$ и $p_i = \{0.5, 0.5\}$ имеем $\langle x \rangle = 1$, $\sigma = 1$ и диапазон $x = 1 \pm 1$ вполне соответствует действительности.

Заметим, что в первом случае волатильность в шесть раз выше, чем во втором. Именно это было интегральной причиной неприятия более высокого риска, ассоциируемого с волатильностью.

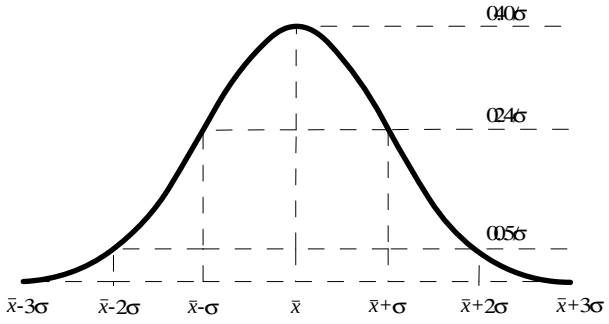
2.3 Нормальное распределение

• На практике часто возникает плотность вероятности, близкая к *нормальному распределению* или *распределению Гаусса*:

$$\boxed{P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right]}. \tag{2.5}$$

Оно характерно для ситуации, когда на переменную x оказывает воздействие множество несвязанных друг с другом внешних факторов, и она флуктуирует вокруг среднего значения. Однако термин "нормальное" не должен вводить в заблуждение. Эмпирическое распределение вероятностей иногда может существенно отличаться от нормального.

Функция (2.5) определяется двумя параметрами – $\bar{x}$ и σ. Для нормального распределения мода, медианное и арифметическое средние совпадают по величине и равны параметру $\bar{x}$. Параметр σ представляет собой волатильность случайной величины x (≺ стр.550).



Нормальное распределение симметрично относительно прямой $x = \bar{x}$ и изменяет свою кривизну в точках перегиба $x = \bar{x} \pm \sigma$. В максимуме плотность вероятности достигает значения $P(\bar{x}) = 1/\sigma\sqrt{2\pi} \approx 0.40/\sigma$. Чем больше σ, тем "колокол" распределения шире и ниже. Поэтому при больших σ более вероятны существенные отклонения величины x от своего среднего $\bar{x}$.

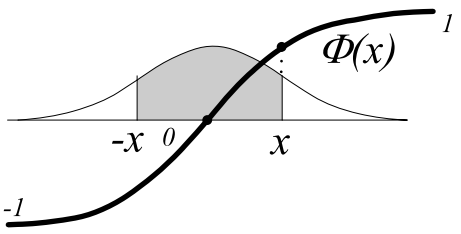
• Для получения вероятности того, что случайная величина окажется в интервале $x = [x_1..x_2]$, необходимо вычислить интеграл от x_1 до x_2: В случае распределения Гаусса после замены $z = (x-\bar{x})/\sigma$ эта вероятность может быть записана в следующем виде:

$$p(x_1 \leqslant x \leqslant x_2) = \int_{x_1}^{x_2} P(x)\, dx \;=\; \int_{(x_1-\bar{x})/\sigma}^{(x_2-\bar{x})/\sigma} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz. \tag{2.6}$$

В результате подынтегральная функция не зависит от параметров $\bar{x}$, σ, и они "перебираются" в пределы интегрирования.

• Рассмотрим случайную величину, распределённую по Гауссу с параметрами $\bar{x} = 0$, $\sigma = 1$. Вероятность того, что мы будем наблюдать её в интервале $[-x...x]$, определяется *функцией Лапласа*:

$$\Phi(x) = \int_{-x}^{x} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz.$$



При $x = 0$ площадь под кривой равна нулю, поэтому $\Phi(0) = 0$. Часто используемые значения $\Phi(1) = 0.683$, $\Phi(2) = 0.955$, $\Phi(3) = 0.997$. При больших x функция Лапласа стремится к единице (условие нормировки). К тому же она нечётная (антисимметричная) функция $\Phi(-x) = -\Phi(x)$.

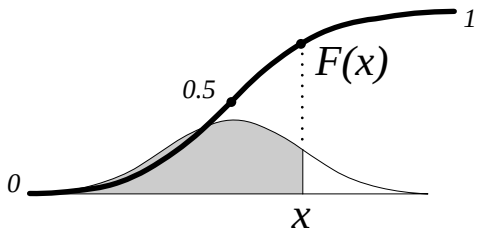
Вероятность отклонения случайной величины x с *произвольными* параметрами $\bar{x}$, σ от среднего не более, чем на Δx в соответствии с (2.6) равна:

$$p(\bar{x} - \Delta x \leqslant x \leqslant \bar{x} + \Delta x) = \Phi\left(\frac{\Delta x}{\sigma}\right).$$

В частности, вероятность отклонения на величину, не превышающую параметр σ, равна $\Phi(1\sigma/\sigma) = \Phi(1) = 0.68$, или реализуется в 68% случаев. Аналогично, отклонение не более, чем на две сигмы $\Phi(2\sigma/\sigma) = \Phi(2) = 0.95$, происходит в 95% случаев. Экспонента убывает очень быстро, поэтому популярно *правило трех сигм*, которое с возможной вероятностью ошибки 0.003=1-0.997 предсказывает, что выбранный наугад x_i будет лежать в диапазоне $\bar{x} \pm 3\sigma$.

• *Интегральным распределением* мы называем вероятность того, что случайная величина не превышает некоторого значения x. Для нормального распределения с параметрами $\bar{x} = 0$ и $\sigma = 1$ имеем:

$$F(x) = \int_{-\infty}^{x} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz.$$



Эта функция, как и $\Phi(x)$, не выражается через более элементарные. В Excel $F(x)$ можно вычислить при помощи вызова НОРМСТРАСП(x). Понятно, что $F(\infty) = 1$ и $F(0) = 1/2$ (≪ стр.550). Из геометрических соображений несложно получить следующие соотношения (≪ стр.550):

$$F(-x) = 1 - F(x), \qquad \Phi(x) = 2F(x) - 1.$$

Заметим также, что производная интегрального распределения (стр.496) даёт функцию плотности распределения $F'(x) = P(x)$.

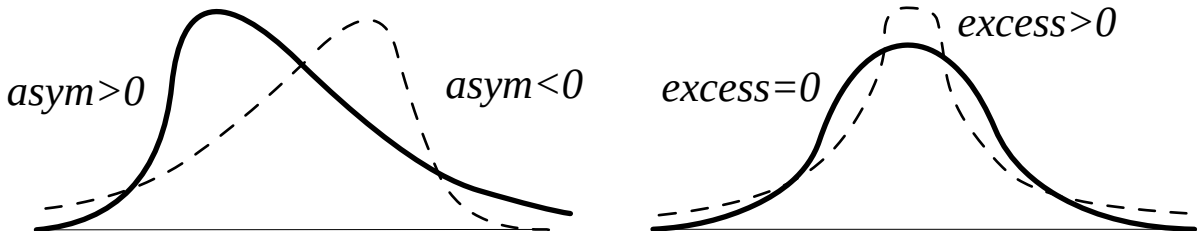
2.4 Отклонение от нормальности

Для характеристики распределения вероятности кроме волатильности полезными оказываются также и моменты отклонения от среднего более высоких степеней, чем вторая. Они называются асимметрией и эксцессом:

$$\boxed{asym = \frac{\left\langle (x-\bar{x})^3 \right\rangle}{\sigma^3}}, \qquad \boxed{excess = \frac{\left\langle (x-\bar{x})^4 \right\rangle}{\sigma^4} - 3}. \tag{2.7}$$

После деления моментов на σ в соответствующей степени они, в отличие от среднего и волатильности, становятся безразмерными. Для вычисления в Excel асимметрии нужно вызвать функцию СКОС(), а для эксцесса – ЭКСЦЕСС().

Асимметрия характеризует степень и направление скошенности распределения относительно среднего. Если асимметрия равна нулю, распределение, вероятнее всего, симметрично относительно точки $x = \bar{x}$. Если она положительна, то у распределения более толстый и длинный "хвост" справа, если отрицательна, то слева от $\bar{x}$. Для асимметричного распределения среднее, мода и медиана (стр.35), вообще говоря, различны.



Эксцесс показывает, насколько медленно убывает вероятность распределения при больших отклонениях от среднего. В случае распределения Гаусса эксцесс равен нулю, для этого из безразмерного момента специально вычитают тройку (≺ стр.551). Если распределение на бесконечности убывает медленнее, чем распределение Гаусса, то эксцесс будет положительным, иначе – отрицательным.

Положительность эксцесса означает наличие "толстых хвостов" в графике распределения вероятности. А это, в свою очередь, приводит к тому, что чаще, по сравнению с "нормальным" случаем (гауссово распределение), могут происходить редкие события. Например, обвалы на фондовом рынке!

Так же, как и в случае с волатильностью, можно вычислять асимметрию и эксцесс при помощи различных формул. Так, для асимметрии, возводя в третью степень под знаком усреднения, получаем:

$$asym \cdot \sigma^3 = \left\langle x^3 \right\rangle - 3\left\langle x^2 \right\rangle \left\langle x \right\rangle + 2\left\langle x \right\rangle^3 .$$

Поэтому значение асимметрии может быть получено при помощи трёх средних.

• При вычислении асимметрии и эксцесса необходимо помнить, что они *очень чувствительны* к редким событиям или ошибочным выбросам даже при большом количестве наблюдений x_i.

Фондовый индекс $S\&P500$ является прекрасным "полигоном" для проведения стохастических исследований. Индекс представляет собой усреднение цен 500 крупнейших американских компаний и отражает общее "настроение" фондового рынка. На finance.yahoo.com доступны исторические данные с 1950 года. В качестве случайных чисел будем рассматривать ежедневные относительные *логарифмические изменения* индекса в процентах: $x_t = \ln(p_t/p_{t-1})$.

Если за 57 лет (1950-2007) вычислить асимметрию и эксцесс величин x_t, то получатся следующие результаты (первая строка):

n	min	m_e	max	$\bar{x}$	σ	$asym$	$excess$	$x > \bar{x}$	$> 1\sigma$	$> 2\sigma$	$> 3\sigma$
14591	-22.9	0.045	8.7	0.031	0.897	-1.30	34.3	51.1	23	4.8	1.2
14589	-9.0	0.044	5.6	0.031	0.876	-0.33	5.9	51.1	23	5.2	1.4

В 1987 году произошел обвал рынка. В понедельник, 19-го октября, индекс к закрытию упал на -22.9%. После этого во вторник и среду произошло стремительное восстановление на 5.2% и 8.7% соответственно. Если эти три исключительных дня заменить на одно суммарное изменение на -9.0%, то значения асимметрии и эксцесса радикально изменятся (вторая строка). Обратим внимание, что число торговых дней n, взятых для вычисления статистических параметров, равнялось очень большому для практической статистики числу $n = 14591$.

• Более устойчивым к выбросам признаком отклонения от нормальности может служить доля чисел, отклонившихся от среднего более, чем на σ, 2σ и 3σ. В таблице последние три колонки соответствуют проценту дней, для которых $|x - \bar{x}| > n\sigma$. В случае нормального распределения эти числа должны быть равны (32, 4.6, 0.27)%. Видно, что "редкие события" ($> 3\sigma$) на фондовом рынке происходят в 5 раз чаще, чем в нормальном случае. Изменения же, меньшие сигмы, происходят на треть реже.

Доля примеров, которые больше среднего, не всегда характеризует знак асимметричности распределения. Например, для индекса S&P500 крахи происходят достаточно редко, но они большие по величине, поэтому его асимметрия отрицательна. Тем не менее, число дневных изменений, больших среднего (колонка $x > \bar{x}$), было равно 51%, т.е. вероятность роста рынка немного больше, чем его падения.

Для проведения статистических расчетов обычно требуется предварительная подготовка и верификация (проверка) данных. Например, исключение явных выбросов. Следует, однако, учитывать, что процедура эта достаточно произвольна, и злоупотреблять ею не стоит.

2.5 Эмпирические распределения

• Рассмотрим сначала случай небольшого количества значений непрерывной случайной величины. Например, за 20 лет (1987-2006 г.) ежеквартальные изменения (в процентах) реального *Валового Внутреннего Продукта* США (GDP в ценах 2000 г.) дают нам $n = 80$ чисел:

n	min	m_e	max	$\bar{x}$	σ_x	$asym$	$excess$	$x > \bar{x}$	$> 1\sigma$	$> 2\sigma$	$> 3\sigma$
80	-0.80	0.73	1.8	0.75	0.50	-0.34	0.58	48.8	26	6.3	1.3

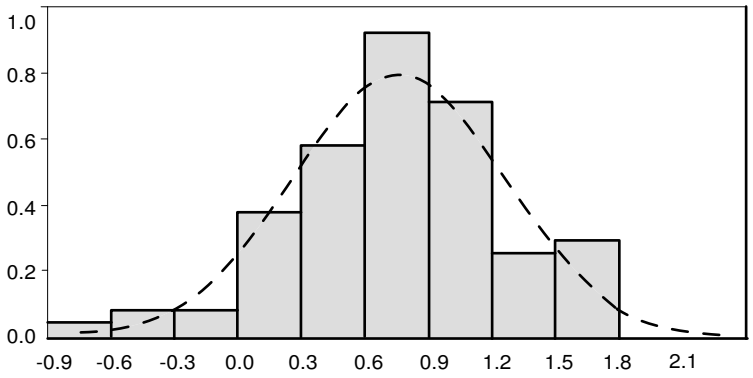
Разобьём некоторый диапазон (обычно $[x_{min}..x_{max}]$) на интервалы $[x_i..x_{i+1}]$ с равной шириной Δx и подсчитаем количество попаданий эмпирических данных в каждый из интервалов. Выбор ширины интервалов достаточно произволен, и скорее преследует цель достижения эстетического баланса между "гладкостью" распределения и большим числом точек на кривой. Часто рекомендуют выбирать число интервалов, равное $1 + 1.4 \ln n$ (*формула Стерджерса*).

В случае с GDP зададим $\Delta x = 0.3$ и расположим интервалы так, чтобы середина центрального пришлась на $\bar{x} = 0.75$ (см. файл 02_stat.xls):

левая граница x_1	-0.9	-0.6	-0.3	0.0	0.3	0.6	0.9	1.2	1.5
правая граница x_2	-0.6	-0.3	0.0	0.3	0.6	0.9	1.2	1.5	1.8
количество n_i	1	2	2	9	14	22	17	6	7
плотность $P(x)$	0.04	0.08	0.08	0.38	0.58	0.92	0.71	0.25	0.29

Для того, чтобы частоты можно было сравнивать с теоретическими функциями распределения, необходимо n_i разделить на общее количество чисел $n = 80$ (вероятность) и на ширину интервала (плотность): $P(x) = (n_i/n)/\Delta x$.

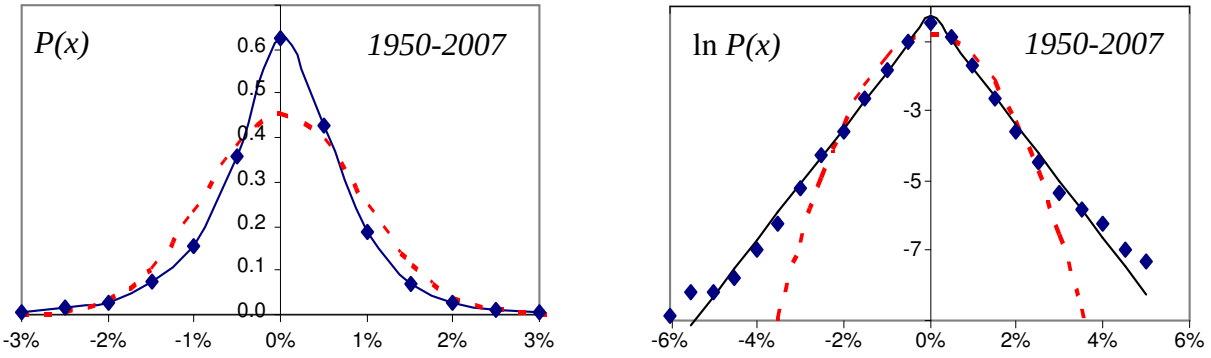
Так как интервалов немного, естественно представить плотность вероятности в виде *гистограммы* (столбиков):



Пунктирная линия – это нормальное распределение с параметрами $\bar{x} = 0.75$ и $\sigma = 0.50$. Видно, что формула Гаусса достаточно хорошо описывает эмпирическое распределение частот, однако небольшие асимметрия и эксцесс всё же сказываются.

Любопытно, что нулевой или отрицательный эксцесс – достаточно редкое явление для распределения *изменений* различных экономических и финансовых величин.

• Построим теперь распределение вероятности для 14589 ежедневных логарифмических изменений $x_t = \ln p_t/p_{t-1}$ фондового индекса S&P500 (левая диаграмма). Чтобы увидеть поведение "хвостов" распределения, необходимо перейти к $\ln P(x)$. Логарифмическая функция растёт очень медленно, поэтому в логарифмическом масштабе удобно сравнивать сильно различающиеся значения. В частности, распределение Гаусса превращается в параболу (пунктир), а эмпирическое распределение имеет следующий вид (правая диаграмма):



Сплошные линии на правом рисунке соответствуют *распределению Лапласа*:

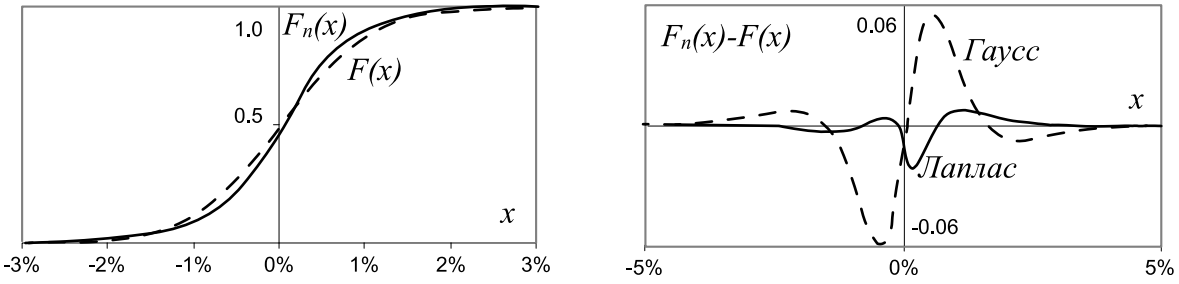
$$P(x) = \frac{\lambda}{2}\, e^{-\lambda|x-\bar{x}|},$$

имеющему волатильность $\sigma = \sqrt{2}/\lambda$ и эксцесс $excess = 3$ (⋖ стр.551). Видно, что распределение Лапласа существенно лучше, чем распределение Гаусса.

Существует множество методик подбора аналитической функции $P(x)$ и проверки её адекватности. Один из критериев, предложенный Колмогоровым – это минимизация максимального отклонения D эмпирического *интегрального распределения* $F_n(x)$ от теоретического $F(x)$:

$$D = \max_x |F_n(x) - F(x)|, \qquad \text{где} \quad F(x) = \int_{-\infty}^{x} P(z)dz.$$

На рисунке слева даны эмпирическое интегральное распределение (сплошная линия) и гауссово (пунктирная). На правом рисунке – отклонения эмпирического распределения от гауссового (пунктир) и лапласового (сплошная линия):



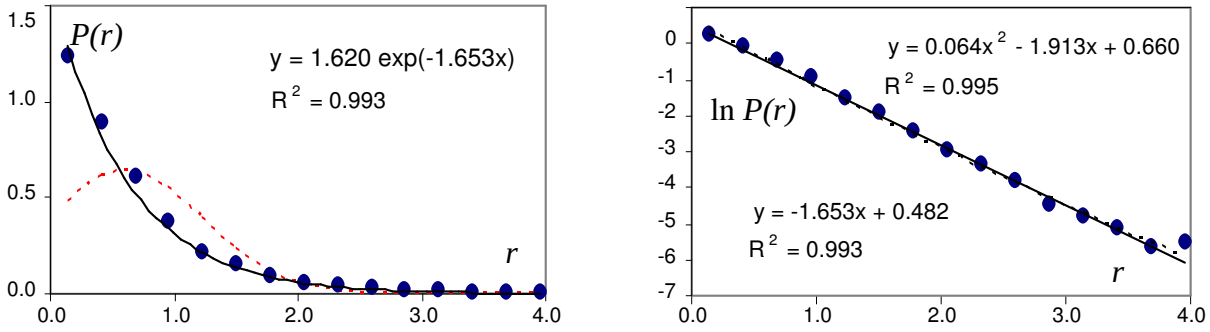
Видно, что наибольшее отклонение эмпирического интегрального распределения происходит в окрестности $\bar{x} \pm \sigma$.

• Рассмотрим также распределение плотности вероятности *модуля* изменения $r_t = |\ln p_t/p_{t-1}|$ индекса S&P500 за период 1950-2006 (всегда $r_t \geqslant 0$):

n	*min*	m_e	*max*	$\bar{r}$	σ	*asym*	*excess*	$\rangle 1\sigma$	$\rangle 2\sigma$	$\rangle 3\sigma$
14338	0.0	0.448	9.0	0.619	0.617	2.71	15.1	13	4.3	1.7

Интегральные параметры свидетельствуют о существенном отклонении от нормального распределения в данном случае. Обращают на себя внимание также близкие значения для среднего и волатильности.

Эмпирическое распределение *не имеет максимума* в окрестности среднего значения и монотонно спадает при увеличении r_t. Пунктирная линия соответствует нормальному распределению $(\bar{r},\ \sigma)$, а сплошная – обычной экспоненте (левая диаграмма):



Справа представлен логарифм этого распределения. Видно, что эмпирические данные в логарифмическом масштабе достаточно хорошо ложатся на прямую линию. Пунктирная линия соответствует трёхпараметрической параболе и практически не улучшает аппроксимации.

Таким образом, абсолютные изменения фондового индекса за период 1950-2006 в первом приближении можно описать *показательным (экспоненциальным) распределением*:

$$\boxed{P(r) = \lambda\, e^{-\lambda r}}\,.$$

Аналогичное распределение справедливо для модулей изменений цен самых разнообразных финансовых инструментов на различных временных интервалах. Для показательного распределения среднее значение и волатильность совпадают $\langle r\rangle = \sigma_r = 1/\lambda$ (⋖ стр.552). Например, курс евро к доллару для модулей часовых изменений имеет среднее 0.082 и волатильность 0.092, и неплохо, в логарифмическом масштабе, ложится на прямую.

Для показательного распределения справедливо важное соотношение:

$$P(x_1 + x_2) = P(x_1) \cdot P(x_2).$$

В качестве упражнения стоит вычислить значения асимметрии $asym = 2$ и эксцесса $excess = 6$ показательного распределения (⋖ стр.552) и, сравнив их с эмпирическими, сделать вывод о характере распределения при $r \gg \sigma$.

• Обобщением показательного распределения является *гамма-распределение* с двумя параметрами λ и α ($x \geqslant 0$):

$$\boxed{P(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\lambda x}}. \tag{2.8}$$

Это распределение имеет нулевую плотность вероятности при $x < 0$, т.е. описывает всегда не отрицательную случайную величину. Его интегральные параметры: $\bar{x} = \alpha/\lambda$, $\sigma = \sqrt{\alpha}/\lambda$, $asym = 2/\sqrt{\alpha}$ и $excess = 6/\alpha$ (≪ стр.552).

Можно также обобщить и распределения Лапласа, для устранения из него "не гладкого" модуля: $P(x) \sim \exp\bigl[-\lambda \cdot \sqrt{(x-\bar{x})^2 + \mu^2}\,\bigr]$.

• Распределения изменений цен различных финансовых инструментов обладают любопытным свойством наличия "толстых хвостов" (fat) по сравнению с распределением Гаусса. При больших отклонениях от среднего распределение иногда аппроксимируют асимптотическим *степенным законом*:

$$P(x) \sim A|x|^{-\mu}, \qquad x \to \pm\infty.$$

Степенная функция убывает медленнее, чем экспоненциальная или гауссова. Заметим, что при больших μ экспоненциальная и степенная асимптотики могут совпадать, так как $(1 + ax/\mu)^{-\mu} \to \exp(-ax)$, при $\mu \to \infty$.

Для обнаружения степенной зависимости необходимо строить график плотности распределения в логарифмическом масштабе по обоим осям, так как в этом случае получится прямая: $\ln P(x) \sim -\mu \cdot \ln x$. К сожалению надёжность восстановления асимптотического поведения плотности распределения невысока из-за малого числа примеров при $|x - \bar{x}|/\sigma_x \gg 1$.

К степенному поведению с $\mu = 2$ приводит *распределение Коши*:

$$\boxed{P(x) = \frac{a/\pi}{a^2 + x^2}}. \tag{2.9}$$

Эта плотность имеет нулевое среднее, характерную ширину a, однако *не обладает* конечной волатильностью σ и моментами более высоких порядков.

Существует большое количество работ, в которых изучаются функциональные зависимости эмпирических распределений изменений цен на финансовые инструменты. Эта деятельность, безусловно, представляет значительный академический интерес. Однако в силу существенной нестационарности рынков и, следовательно, малого числа данных для восстановления зависимости в "новейшей истории", она не имеет большого практического значения. Тем не менее, общепринятым результатом этих исследований является тот факт, что распределение цен существенно ненормально, с толстыми хвостами, приводящими к заметному эксцессу, и, следовательно, к высокой вероятности "редких событий", обычно проявляющихся на рынке в форме финансовых крахов.

2.6 Логнормальное распределение

Рассмотрим ситуацию, когда логарифмические изменения цены x имеют нормальное распределение. Возникает вопрос – какова при этом плотность распределения для самой цены. Таким образом, мы имеем две случайные величины r и x, связанные соотношением $x = e^r x_0$ или $r = \ln(x/x_0)$, где $x_0 = const$ – начальная цена. Величина r имеет нормальное распределение с параметрами $\bar{r}$, σ, требуется найти распределение для *всегда положительной* цены x.

Вычислим сначала среднее значение цены (⋖ стр. 553), считая, что доходности описываются функцией Гаусса $P(r)$:

$$\langle x \rangle = x_0 \langle e^r \rangle = x_0 \int\limits_{-\infty}^{\infty} e^r P(r) dr = x_0 \exp\left(\bar{r} + \frac{\sigma^2}{2}\right), \tag{2.10}$$

Эффективным или мартингальным рынком называется модель, в которой средняя цена остаётся неизменной: $\langle x \rangle = x_0$. В этом случае:

$$\boxed{\bar{r} = -\frac{\sigma^2}{2}}. \tag{2.11}$$

Это довольно забавный результат, означающий, что для постоянства средней цены необходимо, чтобы логарифмическое изменение было в среднем *отрицательным*. Мы видели, что $\langle \ln x \rangle$, вообще говоря, не равно $\ln \langle x \rangle$. Поэтому среднее отношение $\langle \ln(x/x_0) \rangle$ может быть равно 0, но при этом $\langle x \rangle \neq x_0$.

Волатильность цены вычисляется аналогичным образом, и, естественно, *не совпадает* с волатильностью логарифмической доходности:

$$\sigma_x = \bar{x} \cdot \sqrt{\exp(\sigma^2) - 1} \approx \bar{x}\ \sigma. \tag{2.12}$$

Чем больше волатильность логарифмической доходности σ, тем большая будет волатильность цены σ_x. Обычно $\sigma \sim 1\% = 0.01$, поэтому приближенное равенство будет выполняться с хорошей точностью.

• Для определения плотности распределения *цены* необходимо вычислить среднее значение произвольной функции x:

$$\langle F(x) \rangle = \int\limits_{-\infty}^{\infty} F\big(x_0 e^r\big)\ e^{-(r-\bar{r})/2\sigma^2} \frac{dr}{\sigma\sqrt{2\pi}} = \int\limits_{0}^{\infty} F(x)\ P_L(x)\ dx.$$

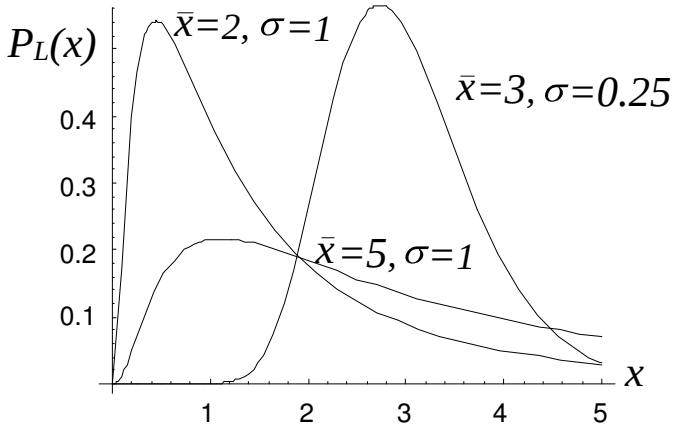
Первый интеграл является вычислением такого среднего при помощи нормального распределения для случайной величины r. Если мы сможем при помощи замен преобразовать этот интеграл во второй, то функция $P_L(x)$ по определению и будет плотностью вероятности положительной случайной величины x.

Проведём в интеграле по r замену $r \to x = x_0 e^r$, $dr = dx/x$. В результате получим стандартное выражение для среднего положительной случайной величины x:

$$\boxed{P_L(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x/\bar{x}) + \sigma^2/2)^2}{2\sigma^2}\right]},$$

где $P_L(x)$ называют *логнормальной плотностью распределения*, а среднее цены $\bar{x} = x_0 \exp(\bar{r} + \sigma^2/2)$. Так как цена x всегда остаётся положительной, то плотность распределения её вероятности равна нулю при $x < 0$.

Таким образом, хотя относительные изменения описываются симметричным нормальным законом Гаусса, цена будет иметь несимметричный логнормальный вид:



Логнормальное распределение имеет не нулевые асимметрию и эксцесс:

$$asym = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1} \approx 3\sigma, \qquad excess = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6 \approx 16\sigma^2.$$

Мода (максимум) логнормального распределения соответствует точке:

$$m_0 = \bar{x}e^{-3\sigma^2/2} \approx \bar{x}, \qquad P_L(m_0) = \frac{e^{\sigma^2}}{\bar{x}\sigma\sqrt{2\pi}} \approx \frac{1}{\sigma_x\sqrt{2\pi}}.$$

При маленькой волатильности изменений цены σ логнормальное и нормальное распределения имеют высокий узкий максимум в окрестности $x = m_0 \approx \bar{x}$ и практически совпадают.

Применённый выше приём построения неизвестной плотности распределения случайной величины, которая является функцией известной величины, достаточно общий. В качестве упражнения предлагается найти плотность распределения курса: $x = €/\$$, считая, что оно, в силу симметрии, такое же, как и у обратного кросс-курса $y = 1/x = \$/€$ (≺ стр. 553).

2.7 Статистические распределения *

Большинство распределений, применяющихся в статистике, выводятся из нормального распределения. При этом рассматривают случайную величину u, являющуюся некоторой комбинацией $u = f(x_1, .., x_n)$ *независимых* случайных чисел x_i. Обратим внимание, что индексом мы сейчас нумеруем не результаты измерений, а различные случайные числа. Плотность вероятности независимых величин по определению (стр.453) равна произведению вероятностей каждой из них:

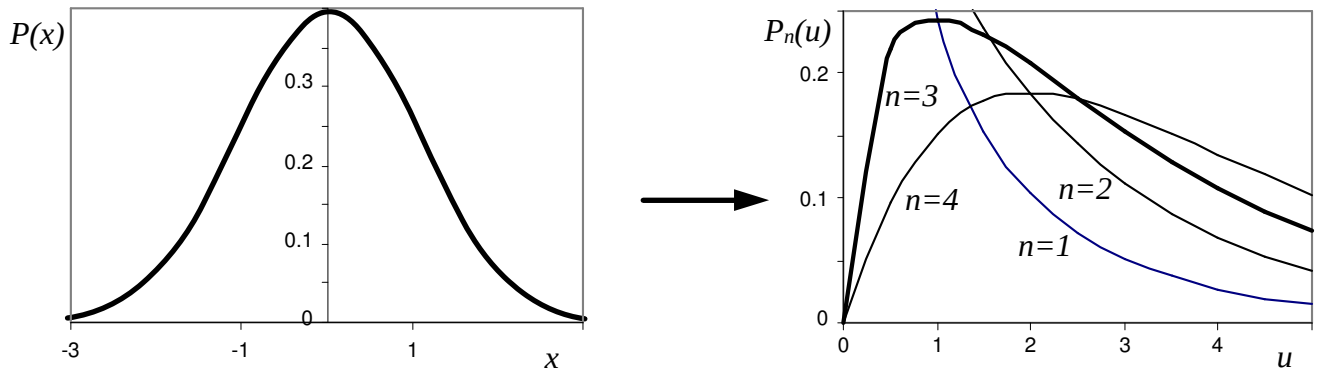
$$P(x_1, .., x_n) = P(x_1) \cdot .. \cdot P(x_n).$$

Среднее некоторой произвольной функции $\langle F(u) \rangle$ можно вычислить двумя способами: 1) имея плотность вероятности $P_n(u)$, для u вычислить однократный интеграл; 2) при помощи $P(x_1, .., x_n)$ произвести n – кратное интегрирование по x_i от $F(u) = F(f(x_1, .., x_n))$. Результат в обоих случаях должен, естественно, получиться одинаковым:

$$\langle F(u) \rangle = \int F(u) \cdot P_n(u)\ du = \int F(f(x_1, .., x_n)) \cdot P(x_1)..P(x_n)\ dx_1..dx_n.$$

Поэтому, чтобы найти распределение для u, необходимо свести n – кратный интеграл к однократному. Тогда функция, стоящая множителем при $F(u)$, и окажется плотностью распределения случайной величины u.

Например, если каждая величина x_i распределена нормальным образом, то сумма их квадратов $u = x_1^2 + .. + x_n^2$ имеет довольно специфическое распределение, качественно различное при $n = 1, 2$ и $n > 2$:



Заметим, что если x_i изменялись в диапазоне от минус до плюс бесконечности, то всегда положительная величина u будет иметь распределение, отличное от нуля, только в интервале $[0..\infty]$. Рассмотрим более подробно вывод этого распределения.

• Пусть n случайных независимых величин распределены нормально с нулевым средним и единичной волатильностью. Найдём плотность вероятности для следующей их комбинации ($0 \leqslant u < \infty$):

$$u = x_1^2 + .. + x_n^2.$$

Вычислим среднее произвольной функции $F(u)$:

$$\langle F(u) \rangle = \int_0^\infty F(u) P_n(u) du = \int_{-\infty}^{\infty} .. \int_{-\infty}^{\infty} F(x_1^2 + .. + x_n^2) e^{-\frac{1}{2}(x_1^2 + .. + x_n^2)} \frac{dx_1 .. dx_n}{(2\pi)^{n/2}}.$$

Первый интеграл является общей формулой для вычисления среднего всегда положительной случайной величины $u > 0$, для которой $P_n(u < 0) = 0$. Второе выражение вычисляет то же среднее при помощи n гауссовых интегралов для каждой из величин x_i.

Введём длину радиус-вектора $r = \sqrt{x_1^2 + .. + x_n^2}$ в $n-$мерном пространстве. Из соображений размерности понятно, что объём n-мерного шара будет пропорционален $V \sim r^n$. В частности, площадь круга ($n = 2$) равна $S = \pi r^2$, а объём шара ($n = 3$): $V = (4\pi/3) r^3$. Поэтому элемент объёма $dV = dx_1 .. dx_n$ в n-мерных сферических координатах равен $dV = r^{n-1} dr d\Omega$, где $d\Omega$ – элемент объема, определяемый остальными, "угловыми" координатами. Так как подынтегральная функция зависит только от r, то интеграл по $d\Omega$ будет равен некоторой константе. Поэтому, учитывая, что $r^2 = u$, а $dr = du/2\sqrt{u}$, приходим к следующему выражению:

$$P_n(u) = C \cdot u^{n/2-1} \cdot e^{-u/2}.$$

Константа C находится из условия нормировки (≺ стр.553). В результате окончательно получаем:

$$\boxed{P_n(u) = \frac{1}{2^{n/2}\Gamma(n/2)} \, u^{n/2-1} e^{-u/2}}, \qquad (2.13)$$

где $\Gamma(z)$ – гамма-функция (стр. 502). Эту плотность вероятности называют χ^2 - *распределением* (хи-квадрат), а параметр n - *числом степеней свободы*.

χ^2 – распределение является частным случаем рассмотренного ранее гамма-распределения (2.8, стр.47), следовательно, его среднее $\langle u \rangle = n$ и волатильность $\sigma_u = \sqrt{2n}$.

Численное значение площади под χ^2 – распределением на интервале $u = [x..\infty]$ можно найти, вызвав в Excel функцию "ХИ2РАСП(x, n)". В частности, "ХИ2РАСП$(0, n)$=1".

• Если одна случайная величина x нормально распределена, а вторая, не зависящая от неё y имеет χ^2 распределение с n степенями свободы, то можно рассмотреть их отношение вида $(-\infty < t < \infty)$:

$$t = \frac{x}{\sqrt{y/n}} = \frac{x}{\sqrt{(x_1^2 + .. + x_n^2)/n}}.$$

Как и выше, вычисляем среднее при помощи произведения плотностей вероятности случайных чисел x и y. Для x это нормальное распределение, а для y – распределение χ^2:

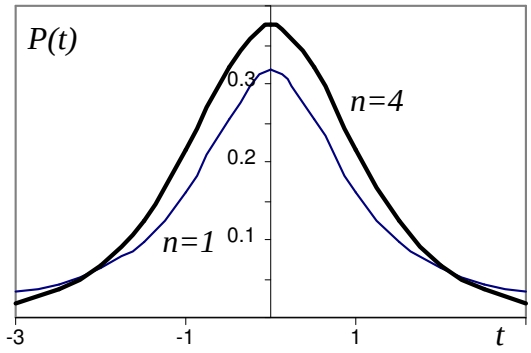
$$\langle F(t)\rangle = \int_{-\infty}^{\infty}\int_{0}^{\infty} F\left(\frac{x}{\sqrt{y/n}}\right) e^{-x^2/2}\,\frac{dx}{\sqrt{2\pi}} \cdot y^{n/2-1}e^{-y/2}\,\frac{dy}{2^{n/2}\Gamma(n/2)}.$$

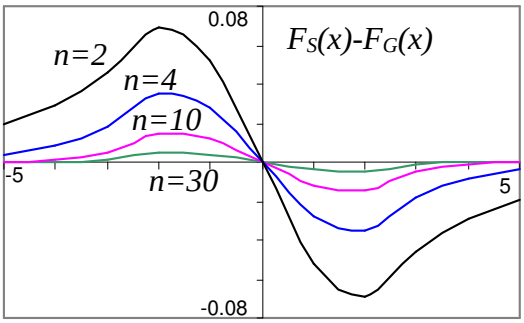
Если сделать замену $t = x/\sqrt{y/n}$ и $s = y$, то двойные интегралы расщепляются и интеграл по s легко вычисляется (≪ стр.554). В результате для величины t получается *распределение Стьюдента* с n степенями свободы:

$$\boxed{P_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\;\Gamma(n/2)}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}}.$$

При больших n выражение $(1 + t^2/n)^{-(n+1)/2}$ стремится к $e^{-t^2/2}$ (см. число Эйлера, стр.445) и, следовательно, распределение Стьюдента стремится к нормальному распределению с $\bar{x} = 0$, $\sigma = 1$. В частном случае $n = 1$ получается распределение Коши (2.9, стр.47).

Ниже на рисунках представлены распределения Стьюдента при $n = 1$ и 4, и разница *интегрального* распределения Стьюдента и Гаусса при различных n:





Среднее значение $\langle t\rangle = 0$, а дисперсия при $n > 2$ равна $\langle t^2\rangle = n/(n-2)$ и бесконечна при $n = 1, 2$. Если $n > 4$, то распределение Стьюдента имеет эксцесс, равный $6/(n-4)$. Среднее значение $\langle t^\mu\rangle$ вычисляется при помощи интеграла (51 стр.503), с заменой в нём $t^2 = z$.

Площадь под $P_n(t)$ в интервале $t = [-x...x]$ может быть найдена в Excel вызовом функции “1-СТЬЮДРАСП$(x, n, 2)$”.

• Рассмотрим теперь две *независимые* случайные величины x и y, имеющие χ^2 распределения с n и m степенями свободы соответственно. Введём новую переменную f, равную их нормированному отношению ($0 \leqslant f < \infty$):

$$f = \frac{x/n}{y/m} = \frac{(x_1^2 + .. + x_n^2)/n}{(y_1^2 + .. + y_m^2)/m}.$$

Среднее значение произвольной функции от переменной f вычисляется при помощи двойного интеграла, содержащего произведение двух χ^2 плотностей вероятности для переменной x и y:

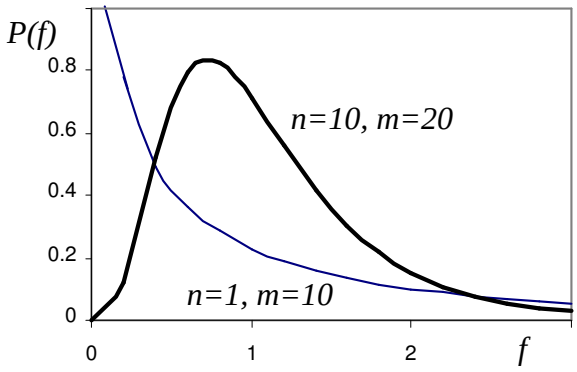
$$\langle F(f)\rangle = \int_0^\infty \int_0^\infty F\left(\frac{x\cdot m}{y\cdot n}\right) \; x^{n/2-1}e^{-x/2}\,\frac{dx}{2^{n/2}\Gamma(n/2)} \;\cdot\; y^{m/2-1}e^{-y/2}\frac{dy}{2^{m/2}\Gamma(m/2)}.$$

Введение переменных $f = (x\cdot m)/(y\cdot n)$ и $g = x + y$ позволяет разделить интеграл (≪ стр.554). Результирующая плотность вероятности для случайной величины f называется *распределением Фишера*:

$$\boxed{P_{n,m}(f) = \mu\,\frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)}\,\frac{(\mu f)^{n/2-1}}{(1+\mu f)^{(n+m)/2}}},$$

где $\mu = n/m$. Несложно видеть, что, если $n = 1$, распределение сингулярно (равно бесконечности) в точке $f = 0$. При $n > 2$ и любых значениях m распределение имеет единственный максимум при $\mu f_0 = (n-2)/(m+2)$. Среднее значение $\langle f\rangle = m/(m-2)$ получается при помощи бета-функции (51 стр.503).

На рисунке представлены графики распределения Фишера при различных комбинациях n и m.



Площадь под $P_{n,m}(f)$ в интервале $f = [x..\infty]$ при помощи Excel вычисляется как “FРАСП(x, n, m)”.

Заметим, что все приведенные в этом разделе функции Excel в различных русифицированных версиях электронной таблицы, вообще говоря, могут иметь несколько различное написание. Поэтому для их уточнения имеет смысл обратиться к подсказке Excel и не забыть включить опцию “Пакет анализа” в меню “Сервис - Надстройки”.

2.8 Производящая функция **

Запишем непрерывное преобразование Фурье (стр.506) для плотности вероятности случайной величины x:

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx}\ P(x)\ dx, \qquad P(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\ \phi(t)\ dt.$$

Функция $\phi(t)$ называется *производящей функцией*. Это название происходит от её очень полезного свойства – быстро получать средние значения степеней x:

$$\frac{1}{i^n}\frac{d^n\phi(t)}{dt^n}\bigg|_{t=0} = \int_{-\infty}^{\infty} x^n\ P(x)\ dx\ =\ \langle x^n\rangle\,.$$

Таким образом, проведя один раз Фурье – интегрирование и найдя производящую функцию, затем можно простым дифференцированием получать значение моментов случайной величины x. Производящая функция является объектом, эквивалентным плотности вероятности или интегральной вероятности, и бывает очень полезна в приложениях. Обратим внимание, что её можно также записать как среднее следующего вида: $\phi(t) = \left\langle e^{itx}\right\rangle$.

Приведём примеры производящих функций для некоторых важных распределений вероятности:

$$\text{Гаусс}: \qquad P(x) = \frac{e^{-(x-x_0)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}, \qquad \phi(t) = e^{ix_0t-\sigma^2t^2/2}.$$

$$\text{Коши}: \qquad P(x) = \frac{a/\pi}{(x-x_0)^2+a^2}, \qquad \phi(t) = e^{ix_0t-a|t|}.$$

$$\text{Гамма}: \qquad P(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\ x^{\alpha-1}\ e^{-\lambda x}, \qquad \phi(t) = \frac{1}{(1-it/\lambda)^\alpha}.$$

Например, для того, чтобы найти среднее значение квадрата x нормального распределения, сначала возьмём первую производную. В точке $t=0$ она равна:

$$\langle x\rangle = \frac{1}{i}\frac{d\phi(t)}{dt}\bigg|_{t=0} = \frac{1}{i}\ (ix_0-\sigma^2t)\ e^{ix_0t-\sigma^2t^2/2}\ \bigg|_{t=0} = x_0.$$

Аналогично, взяв вторую производную, получим $\left\langle x^2\right\rangle = \sigma^2 + x_0^2$, и т.д.

Для нахождения $\phi(t)$ распределения Гаусса необходимо выделить полный квадрат в экспоненте. Функция $\phi(t)$ Коши проще проверяется в обратном направлении, вычисляя по ней $P(x)$. В третьем случае – по формуле для гамма-функции проводится прямое интегрирование.

• Случайная величина y называется *бесконечно делимой*, если её можно представить в виде суммы *независимых* случайных чисел:

$$y = x_1 + .. + x_n.$$

Это важное свойство, которое будет лежать в основе различных моделей стохастической динамики финансовых цен. Например, x_i могут быть ежедневными изменениями цены y.

При помощи производящей функции $\phi(t)$ чисел x_i можно получить производящую функцию, и, следовательно, распределение для случайной величины y. Для этого вычислим среднее значение произвольной функции $F(y)$. Плотность вероятности независимых чисел $P(x_1, .., x_n) = P(x_1) \cdot .. \cdot P(x_n)$, поэтому:

$$\langle F(y) \rangle = \int F(x_1 + .. + x_n)\ P(x_1) \cdot .. \cdot P(x_n)\ dx_1..dx_n.$$

Подставим вместо $P(x_i)$ их фурье-представление при помощи производящей функции. Кроме этого, добавим ещё один интеграл по y, введя δ-функцию Дирака (стр.506). В силу её свойств интегрирование по y вернёт нас к первоначальному выражению, поэтому эти два представления среднего эквивалентны:

$$\langle F(y) \rangle = \frac{1}{(2\pi)^n} \int F(y)\delta(x_1 + .. + x_n - y)\ \phi(t_1) \cdot .. \cdot \phi(t_n) e^{-i\vec{x}\vec{t}}\ dx_1..dx_n\ dt_1..dt_n\ dy.$$

Представив δ - функцию в интегральной форме по переменной t (стр.506)

$$\delta(x_1 + .. + x_n - y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{it\cdot(x_1+..+x_n-y)}\ dt,$$

получим "расщепление" интеграла по каждому из x_k:

$$\langle F(y) \rangle = \frac{1}{2\pi} \int F(y) e^{-ity} \left[\prod_{k=1}^{n} \frac{1}{2\pi} \int\ \phi(t_k) e^{ix_k\cdot(t-t_k)}\ dx_k\ dt_k \right]\ dtdy.$$

Интегралы по x_k приводят к функциям Дирака $\delta(t - t_k)$, интегрирование с которыми по t_k даёт нам окончательный результат:

$$\langle F(y) \rangle = \int F(y) \left[\frac{1}{2\pi} \int \phi^n(t)\ e^{-ity}\ dt \right] dy.$$

Таким образом, производящая функция для случайной переменной y равна n-той степени производящей функции x_i (или их произведению, если они имеют разные распределения).

Теперь несложно видеть, что сумма нормально распределённых чисел будет снова нормально распределённой случайной величиной со средним $n\bar{x}_0$ и волатильностью $\sigma\sqrt{n}$. Аналогичным свойством обладают распределение Коши с итоговыми параметрами $(x_0 n,\ an)$ и гамма-распределение с $(\lambda, \alpha n)$.

2.9 Распределение Бернули

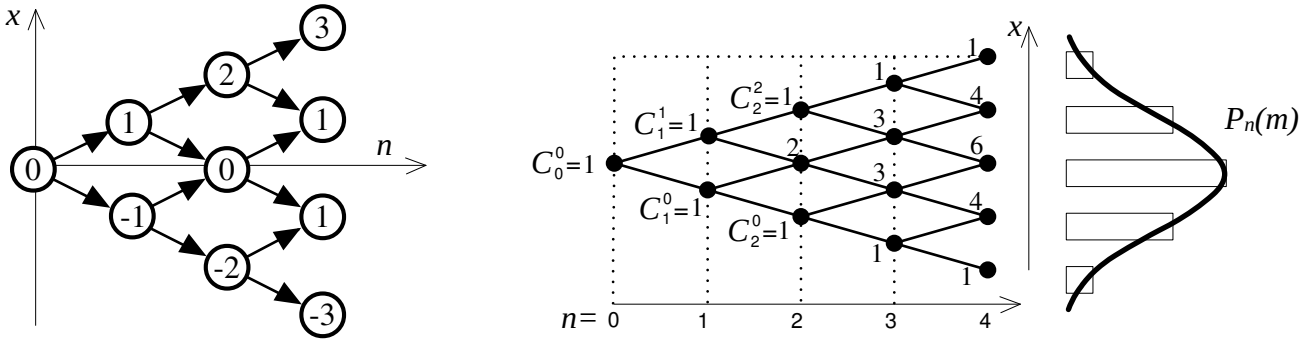
Цены финансовых инструментов удобно представлять в виде непрерывных чисел. Однако на самом деле они дискретны, так как всегда существует минимальное изменение цены (цент, пункт и т.п.). Кроме этого, ряд задач естественным образом описывается в терминах именно дискретных вероятностей.

Рассмотрим простой пример. Пусть в конце торгового дня подводятся его результаты – удачный (нет убытков) или неудачный (получены убытки). Предположим, что результаты трейдера достаточно стабильны и существует вероятность p удачного дня, не зависящая от вчерашнего результата. Зададимся вопросом:

> "*какова вероятность того, что на протяжении n торговых дней m из них окажется удачными?*".

Обозначим удачный день как "1", а неудачный как "0". Если $n = 4$, а $m = 2$, то возможны следующие варианты истории: 1100, 1010, 1001, 0110, 0101, 0011.

Все возможные истории можно представить в виде дерева. Рассмотрим ситуацию, когда трейдер каждый день или зарабатывает один доллар, или теряет его ☹. В этом случае суммарный доход (или убыток) в долларах будет равен разности удачных и неудачных дней $x = m - (n - m) = 2m - n$. Его блуждание во времени можно представить при помощи дерева переходов. Так, в случае четырёх периодов оно выглядит следующим образом (левый рисунок):



Начиная с нулевого дохода, после первого торгового дня результат может быть 1 или -1. На второй день возможны уже три варианта. Если предыдущий день был неудачным, то неудачи могут продолжиться (суммарный "доход" = -2), или будет получен доход, и суммарный результат станет равным нулю. Аналогично для первого удачного дня. Блуждая слева направо по ветвям возможных переходов этого дерева, мы будем получать различные реализации истории.

На правом рисунке цифры возле узлов дерева показывают, сколькими путями можно в этот узел попасть. Так как в любой узел входят только два пути из двух предшествующих узлов, то для того, чтобы определить число способов попадания в него, необходимо сложить цифры двух предшествующих узлов.

Обозначим число способов попадания в узел через C_n^m, где $n = 1, 2, ..$ – номер дня, а $m = 0, 1, .., n$ – номер узла по высоте (число удачных дней). Мы можем записать (m – индекс, а не степень!):

$$C_n^m = C_{n-1}^m + C_{n-1}^{m-1}, \qquad C_n^0 = C_n^n = 1. \tag{2.14}$$

Величины C_n^m называются *биномиальными коэффициентами* и часто встречаются в комбинаторных задачах.

Их значение можно прямо получить при помощи рекуррентных формул (2.14) или вычислить число вариантов расстановки m единиц в последовательности (истории) из n цифр. Первую единицу можно поставить на одно из n мест. *Для каждого* из этих n вариантов вторую единицу можно поставить на одно их оставшихся свободных $n-1$ мест. Поэтому для двух "*различимых*" единиц существует $n \cdot (n-1)$ способов, и т.д. Так как все m единиц на самом деле не различимы, то число полученных комбинаций необходимо разделить на $m!$ повторяющихся вариантов:

$$C_n^m = \frac{n \cdot (n-1) \cdot .. \cdot (n-m+1)}{m!} = \frac{n!}{m!(n-m)!}.$$

Факториал $m! = 1 \cdot 2 \cdot .. \cdot m$ равен числу возможных перестановок m объектов (в нашем случае единиц). Рассуждение аналогично. Первый объект можно поставить на m мест, второй на одно из оставшихся $m-1$ мест, и т.д.

Если вероятность удачного дня p, то для неудачного она равна $1-p$. Поэтому в каждом из вариантов истории m независимых удачных дней и $n-m$ неудачных произойдут с вероятностью $p^m(1-p)^{n-m}$. Количество различных возможных историй (способов попадания в данное состояние) равно C_n^m. В результате суммарная вероятность m удачных дней (в *любой последовательности*!) будет равна:

$$\boxed{P_n(m) = C_n^m \; p^m(1-p)^{n-m}}.$$

Её называют *распределением Бернули* (или биномиальным распределением). Сумма всех вероятностей должна быть равна единице:

$$\sum_{m=0}^{n} P_n(m) = 1.$$

Это соотношение выполняется автоматически и, как обычно, называется *нормировочным условием.*

• Выше для непрерывных распределений мы ввели производящую функцию в виде фурье - преобразования. В случае дискретных распределений более удобной оказывается *степенная производящая функция*. Для распределения Бернули её можно записать в следующем виде:

$$\phi(s) = \sum_{m=0}^{n} s^m P_n(m) = (1 - p + s \cdot p)^n.$$

В этом несложно убедиться, раскладывая $(1 - p + s \cdot p)^n$ в ряд Тейлора, беря последовательно производные по s. Коэффициенты при степенях s^m окажутся в точности равны биномиальным коэффициентам, умноженным на $p^m(1-p)^{n-m}$. В частности справедлива следующая формула (*биноm Ньютона*):

$$(p + q)^n = \sum_{m=0}^{n} C_n^m \; p^m q^{n-m}.$$

Производные производящей функции по s в точке $s = 1$ позволяют легко вычислить среднее различных степеней m:

$$\langle m \rangle = \sum_{m=0}^{n} m P_n(m) = \phi'(1) = np.$$

Вторая производная: $\phi''(1) = \langle m \cdot (m-1) \rangle = \langle m^2 \rangle - \langle m \rangle = n \cdot (n-1) \cdot p^2$. Следовательно, среднее и дисперсия распределения Бернули имеют следующие значения:

$$\bar{m} = \langle m \rangle = np, \qquad \sigma^2 = \langle m^2 \rangle - \langle m \rangle^2 = np \cdot (1-p). \tag{2.15}$$

Среднее число удачных дней определяет вероятность удачи трейдера $p = \bar{m}/n$.

Аналогичными вычислениями можно получить асимметрию и эксцесс распределения Бернули:

$$asym = \frac{1 - 2p}{\sigma}, \qquad excess = \frac{1 - 6p \cdot (1-p)}{\sigma^2}.$$

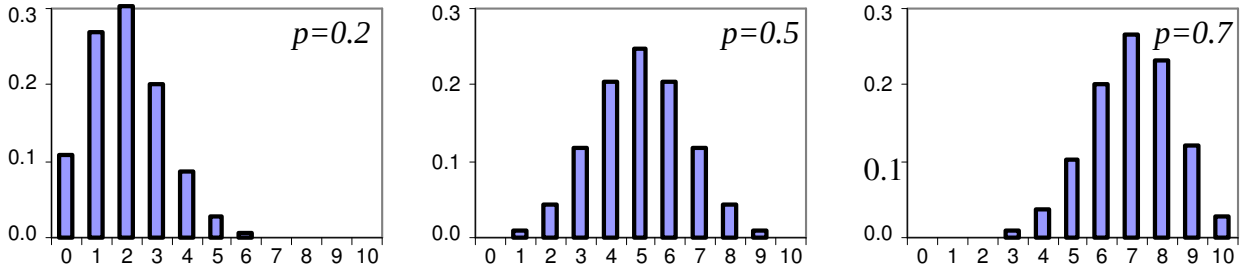
Оно становится симметричным $asym = 0$ для равновероятного случая $p = 1/2$. С ростом n волатильность увеличивается, следовательно, эксцесс стремится к нулю. Как мы увидим ниже, при $n \to \infty$ распределение Бернули стремится к нормальному распределению.

Иногда удобнее при помощи замены $s = e^t$ перейти к производящей функции:

$$\phi(t) = (1 - p + e^t \cdot p)^n = \sum_{m=0}^{n} e^{mt} P_n(m).$$

Её производные по t в точности равны моментам $\langle m^k \rangle = \phi^{(k)}(0)$, однако сами производные становится брать несколько сложнее.

• Нарисуем распределение дискретных вероятностей Бернули при $n = 10$ и различных вероятностях удачи p:



Максимум распределения оказывается в окрестности среднего значения $\langle m \rangle$. Дисперсия достигает наибольшего значения, когда $p = 1/2$, поэтому при существенных отклонениях от $1/2$ распределение становится более узким и высоким, прижимаясь к краю диапазона $m = [0..n]$.

Отметим, что волатильность распределения Бернули (2.15) пропорциональна корню из числа торговых дней n:

$$\sigma = \sqrt{n} \cdot \sigma_0,$$

где $\sigma_0 = \sqrt{p \cdot (1-p)}$ – волатильность одного дня. Это соотношение будет играть ключевую роль при описании стохастических процессов. Оно является достаточно общим результатом, гласящим, что будущая неопределённость увеличивается, как корень квадратный из времени.

Наглядно этот факт можно увидеть на биномиальном дереве, которое с ростом n расширяет число своих веток, а, следовательно, различных возможных будущих состояний и их историй.

• При больших n факториалы чисел в C_n^m растут *очень* быстро, и для их вычисления разумнее использовать формулу Стирлинга (стр.502)

$$n! \;\approx\; \sqrt{2\pi n}\,\left(\frac{n}{e}\right)^n.$$

С её помощью, введя *конечное* относительное отклонение m от его среднего $x = (m - \bar{m})/\sigma$, можно (≪ стр. 555) получить асимптотическое выражение для распределения Бернули, которое при больших n стремится к:

$$P_n(m) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(m-\bar{m})^2/2\sigma^2}, \tag{2.16}$$

где $\bar{m} = np$, а $\sigma^2 = np \cdot (1-p)$.

Таким образом, *нормальное распределение* возникает при достаточно общих предположениях и поэтому часто встречается в природе. Например, если к некоторой величине x случайно прибавляются или вычитаются много небольших приращений Δ и эти воздействия являются *независимыми*, то x оказывается распределённой нормальным образом.

2.10 Распределение Пуассона *

Другое предельное распределение возникает из распределения Бернули в предположении, что вероятность p очень мала, а n, наоборот, велико, но их произведение конечно $n \cdot p = \lambda = const$. Так как $p = \lambda/n$, распределение Бернули можно записать в следующем виде:

$$P_n(m) = C_n^m \cdot p^m \cdot (1-p)^{n-m} = \frac{n \cdot (n-1) \cdot .. \cdot (n-m+1)}{m!} \left(\frac{\lambda}{n}\right)^m \left(1 - \frac{\lambda}{n}\right)^{n-m}.$$

Разделим каждый множитель в числителе биномиального коэффициента на n. В результате имеем:

$$P_n(m) = \frac{\lambda^m}{m!} \cdot 1 \cdot \left(1 - \frac{1}{n}\right) \cdot .. \cdot \left(1 - \frac{m-1}{n}\right) \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^m}.$$

Устремим теперь n к бесконечности, считая λ и m *конечными*. Так как по определению числа Эйлера $(1-\lambda/n)^n \to e^{-\lambda}$, окончательно получаем:

$$\boxed{P_n(m) = \frac{(np)^m}{m!}\, e^{-np}}. \tag{2.17}$$

Эти вероятности называют *распределением Пуассона*. Заметим, что нормировочное условие распределения Пуассона совпадает с разложением в бесконечный ряд Тейлора функции e^{np} (сумма по m от 0 до ∞).

Формулы (2.16), (2.17) являются приближениями к распределению Бернули, которые хорошо работают уже при $n > 10$. Гауссово приближение необходимо использовать при $p \sim 0.5$, а пуассоново – при $p \sim 0$.

Естественно, служить распределением удачных дней вероятности Пуассона не могут ☹. Однако существует другой финансовый процесс, который описывается (2.17). Пусть длительность торгового дня t на бирже разбита на очень большое число n маленьких интервалов времени $\Delta t = t/n$. В каждый из интервалов может с небольшой вероятностью p происходить *одна* торговая сделка. Так как Δt и p очень малы, сделка с большей вероятностью $1-p$ не произойдёт. Вероятность сделки будет тем больше, чем длиннее интервал времени: $p = \lambda \Delta t$, где λ – некоторая константа.

Подставим в распределение Пуассона $n = t/\Delta t$ и $p = \lambda \Delta t$:

$$\boxed{P(m,t) = \frac{(\lambda t)^m}{m!}\, e^{-\lambda t}}. \tag{2.18}$$

Распределение Пуассона в такой "временной" форме описывает множество физических и социальных процессов (вероятности прихода m посетителей в магазин за время t, число звонков на телефонную станцию, вылета электронов из катода и т.д.)

• Распределение Пуассона, описывающее редкие события, встречается достаточно часто, поэтому имеет смысл рассмотреть ещё один способ его вывода.

Обозначим через $P(m,t)$ вероятность m сделок за время t. Как и выше, предположим, что вероятность одной сделки за небольшой интервал времени Δt пропорциональна его длительности $p = \lambda\Delta t$.

Вероятность того, что к моменту $t + \Delta t$ на бирже за день произойдёт m сделок, равна:

$$P(m, t + \Delta t) = P(m-1, t) \cdot \lambda\Delta t + P(m,t) \cdot (1 - \lambda\Delta t).$$

Действительно, существует две возможности. Либо к моменту t было $m-1$ сделок, и с вероятностью $\lambda\Delta t$ произошла ещё одна, либо уже было m сделок, и за Δt не произошло ни одной с вероятностью $1-\lambda\Delta t$. Мы считаем, что интервал Δt очень мал, и более одной сделки произойти не может. Обратим внимание на то, что если бы это было не так, то в первой части стояло бы больше двух слагаемых.

Перепишем это уравнение в следующем виде:

$$\frac{P(m, t + \Delta t) - P(m,t)}{\Delta t} = \lambda \cdot P(m-1, t) \cdot -\lambda \cdot P(m,t).$$

Устремив Δt к нулю, получим дифференциальное уравнение:

$$\frac{\partial P(m,t)}{\partial t} = \lambda \cdot P(m-1,t) - \lambda \cdot P(m,t).$$

Для его решения удобно ввести *производящую функцию*:

$$g(s,t) = \sum_{m=0}^{\infty} s^m \cdot P(m,t).$$

Считая $P(-1,t) = 0$, несложно проверить, что она удовлетворяет простому дифференциальному уравнению:

$$\frac{\partial g(s,t)}{\partial t} = \lambda \cdot (s-1) \cdot g(s,t).$$

Его решением является экспоненциальная функция:

$$g(s,t) = e^{\lambda \cdot (s-1) \cdot t}.$$

Мы выбрали начальное условие $g(s,0) = 1$, так как предполагаем, что при $t = 0$ ни одной сделки ещё нет и, следовательно, $P(0,0) = 1$, а все остальные $P(m,0) = 0$, если $m > 0$. Разлагая выражение для $g(s,t)$ в степенной ряд по s, получаем вероятности (2.18).

При $m \neq 0$ вероятность достигает максимума при $t = m/\lambda$. Вероятность отсутствия событий ($m = 0$) с ростом времени экспоненциально уменьшается $P(0,t) = e^{-\lambda t}$.

2.11 Корреляция

Две случайные величины x и y могут быть связаны между собой некоторой функциональной зависимостью, однако внешние помехи или другие, более глубокие причины приводят к тому, что эта зависимость является не детерминированной (точно определённой), а *стохастической* (случайной).

Важно понимать, что это не всегда означает зависимость y от x или наоборот. Выявление причинно-следственных отношений требует более глубокого анализа. Речь идет только о том, что на графике (x, y) точки эмпирических наблюдений $\{(x_1, y_1), .., (x_n, y_n)\}$ распределены не хаотическим облаком, а "притягиваются" к некоторым функциональным линиям $y = f(x)$.

Подобную зависимость иногда можно описать в виде $y = f(x) + \varepsilon$, где ε – случайная переменная, моделирующая внешний аддитивный шум, который нарушает детерминированный характер функции $f(x)$. При одном и том же значении x каждый раз будут получаться различные значения y, так как величина ε будет при этом непредсказуемо разной. На самом деле это не единственный способ описания стохастических зависимостей, однако он самый простой, и мы начнём именно с него.

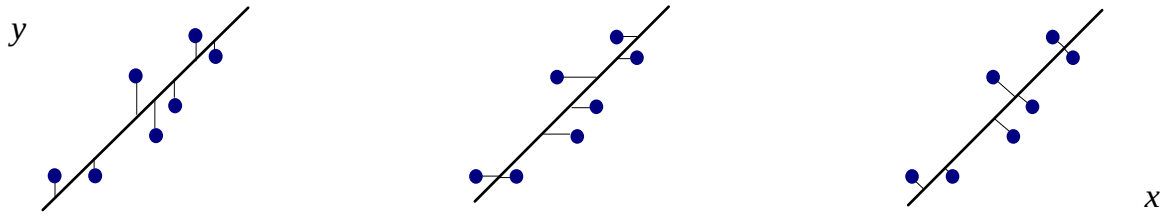
Возможны две интерпретации набора точек $\{(x_1, y_1), .., (x_n, y_n)\}$.

Пусть при различных *заданных* рекламных бюджетах x_i получились некоторые значения суммарных продаж y_i. В этом случае нас интересует связь $y = f(x)$, в которой x является *контролируемой* переменной, которой мы можем управлять, в том числе и при проведении маркетингового исследования. Нас интересует, каково будет значение y, если в следующий раз мы зададим определённое x. При этом связь оказывается стохастической, так как она подвержена внешнему "шуму", воздействующему только на y.

Другая ситуация возникает, например, на фондовом рынке, когда изучается связь между изменением индекса корзины акций x и изменением цены некоторой акции y. В этом случае обе переменные являются случайными, *не контролируемыми* нами величинами, между которыми, тем не менее, существует некоторая зависимость. Мы не знаем, каково будет следующее изменение рыночного индекса, однако с той или иной степенью достоверности можем предсказать, что произойдёт с ценой акции, если индекс поменяется определённым образом.

Если между x и y есть связь, то в обоих случаях на точечной диаграмме (плоскости (x, y)) точки будут "притягиваться" к некоторой линии. Для первого типа данных x_i могут идти с равным шагом, тогда как во втором случае, скорее всего, они будут существенно неравномерны.

• Простейшим видом связи является линейная функция. Для проведения прямой по эмпирическим данным (x_i, y_i) применяют *метод наименьших квадратов*. Считается, что прямая проходит наилучшим образом, если сумма квадратов отклонений от неё по оси y (первый способ) или по оси x (второй способ) минимальны. Возможен и третий способ, в котором минимизируются кратчайшие расстояния до прямой:



Рассмотрим подробнее первый метод. Линейная функция определяется двумя параметрами a и b, нахождение которых и является нашей целью. Сумма квадратов отклонений от прямой по оси y имеет вид:

$$\sum_{i=1}^{n}\big(a + b \cdot x_i - y_i\big)^2 = min$$

и должна быть минимальной.

• Представим линейную зависимость в следующем виде:

$$y = a + b \cdot x + \varepsilon,$$

где a и b – некоторые константы, а ε – случайный шум, делающий *детерминированную* функцию $y = a + b \cdot x$ *стохастической*. На финансовых рынках редко удаётся ставить контролируемые эксперименты $\ddot\smile$, поэтому мы будем считать, что x, y, ε – это случайные числа, связанные линейным соотношением. Однако все выведенные ниже формулы будут справедливы и для ситуации контролируемого x. Просто в этом случае знаки усреднения $\langle..\rangle$ нужно понимать, как простое арифметическое среднее по всем эмпирическим парам $\{(x_i, y_i)\}$.

Пусть среднее значение шума равно нулю $\langle\varepsilon\rangle = 0$, так как в противном случае его всегда можно включить в константу a. Чтобы определить параметры наиболее подходящей прямой, необходимо минимизировать дисперсию шума:

$$\sigma_\varepsilon^2 = \big\langle\varepsilon^2\big\rangle = \big\langle(a + b \cdot x - y)^2\big\rangle = min. \tag{2.19}$$

Возьмём частные производные этого выражения по a, b и приравняем их к нулю:

$$\begin{aligned} \partial\sigma_\varepsilon^2/\partial a &= 2\,\langle a + bx - y\rangle = 0, \\ \partial\sigma_\varepsilon^2/\partial b &= 2\,\langle x \cdot (a + bx - y)\rangle = 0. \end{aligned}$$

Таким образом, параметры прямой a и b удовлетворяют линейной системе из двух уравнений:

$$\begin{cases} a + b\,\langle x\rangle = \langle y\rangle \\ a\,\langle x\rangle + b\,\langle x^2\rangle = \langle xy\rangle\,. \end{cases}$$

Она легко решается и даёт наклон прямой, равный:

$$b = \frac{\langle xy\rangle - \langle x\rangle\,\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2} = \frac{\langle (x-\bar{x})(y-\bar{y})\rangle}{\sigma_x^2} = c(x,y)\,\frac{\sigma_y}{\sigma_x},$$

где волатильности по осям x и y равны σ_x и σ_y, а величина

$$\boxed{c(x,y) = \frac{\gamma(x,y)}{\sigma_x\sigma_y} = \frac{\langle (x-\bar{x})(y-\bar{y})\rangle}{\sigma_x\sigma_y}} \tag{2.20}$$

называется *корреляционным коэффициентом* (correlation). Среднее произведение отклонений $\gamma(x,y)$, стоящее в числителе, называют *ковариацией* (covariance).

Уравнение линейной зависимости удобно записать в симметричном виде:

$$\frac{y-\bar{y}}{\sigma_y} = c(x,y)\cdot\frac{x-\bar{x}}{\sigma_x}, \tag{2.21}$$

так, что корреляция оказывается коэффициентом пропорциональности между двумя безразмерными нормированными величинами.

• Если бы при поиске оптимальной прямой минимизация отклонения производилась не по вертикали, а по горизонтали, то мы получили бы такие же соотношения с перестановкой x и y местами:

$$\frac{x-\bar{x}}{\sigma_x} = c(x,y)\cdot\frac{y-\bar{y}}{\sigma_y}. \tag{2.22}$$

Понятно, что наклоны прямых, полученные каждым из методов, несколько отличаются, совпадая только при $c(x,y) \sim \pm 1$. Можно получить регрессионное уравнение и в случае третьего способа, когда минимизируются расстояния точек к прямой (стр.555).

• Иногда в теоретических моделях необходимо тестировать линейную зависимость без свободного члена ($a = 0$). В этом случае, вообще говоря, $\langle\varepsilon\rangle \neq 0$. Тем не менее, мы можем аналогично, при помощи метода наименьших квадратов, минимизировать волатильность σ_ε шума для зависимости $y = bx + \varepsilon$. В результате получается уравнение оптимальной прямой:

$$y = \frac{\langle xy\rangle}{\langle x^2\rangle}\cdot x.$$

В частности, при аппроксимации данных в Excel можно зафиксировать ($a = 0$) и получить наклон прямой и величину ошибки аппроксимации.

• Полученные выше формулы оптимальной с точки зрения "вертикального" метода наименьших квадратов прямой приводят к любопытному соотношению между x и шумом ε. Подставим в определение корреляции линейную зависимость $y = a+bx+\varepsilon$ и её усреднение $\bar{y} = a+b\bar{x}$ (напомним, что мы не различаем в обозначении среднего фигурные скобки и черту сверху, а $\langle\varepsilon\rangle = 0$):

$$c(x,y) = \frac{\langle(x-\bar{x})(y-\bar{y})\rangle}{\sigma_x\sigma_y} = \frac{\langle(x-\bar{x})(b(x-\bar{x})+\varepsilon)\rangle}{\sigma_x\sigma_y} = b\frac{\sigma_x}{\sigma_y} + \frac{\langle(x-\bar{x})\varepsilon\rangle}{\sigma_x\sigma_y},$$

Так как наклон прямой $b = c(x,y)\sigma_y/\sigma_x$, а $\langle\varepsilon\rangle = 0$, то в результате имеем:

$$\langle x\varepsilon\rangle = 0. \tag{2.23}$$

Естественно, это соотношение справедливо как для случайного x, так и для детерминированного набора контролируемых экспериментов.

• Коэффициент корреляции определяет наклон прямой b. Однако важнее то, что он связан с волатильностью шума. Действительно:

$$\sigma_y^2 = \left\langle(y-\bar{y})^2\right\rangle = \left\langle\left(b\cdot(x-\bar{x})+\varepsilon\right)^2\right\rangle = b^2\sigma_x^2 + \left\langle\varepsilon^2\right\rangle,$$

где мы воспользовались полученным выше соотношением $\langle x\varepsilon\rangle = 0$. Так как $b = c(x,y)\sigma_y/\sigma_x$, то для волатильности шума $\varepsilon_\varepsilon^2 = \left\langle\varepsilon^2\right\rangle$ окончательно имеем:

$$\boxed{E = \frac{\sigma_\varepsilon}{\sigma_y} = \sqrt{1-c^2(x,y)}}. \tag{2.24}$$

Поэтому, чем ближе $c^2(x,y)$ к нулю, тем больше величина отклонения данных от прямой. При $c^2(x,y) = 1$, наоборот, отклонения равны нулю и все точки находятся на прямой. Отрицательный коэффициент свидетельствует о том, что прямая имеет отрицательный наклон, т.е. является нисходящей. Корреляция величины сама с собой равна единице: $c(x,x) = 1$. Из (2.24) следует, что корреляционный коэффициент $c(x,y)$ всегда лежит в пределах от -1 до 1. Заметим, что на графиках в Excel обычно приводится величина $R^2 = 1-E^2 = c^2(x,y)$.

• После проведения аппроксимации данных прямой необходимо изучить поведение *остатков модели*: $\varepsilon_i = y_i - a - b\cdot x_i$. Важны не только статистические свойства ε (например, являются ли они гауссовыми), но и их поведение как функция x или y. Если остатки оказываются в горизонтальной полосе, то это свидетельствует об их постоянстве и служит признаком достаточной адекватности модели. Однако эта полоса может иметь переменную ширину, например, расширяясь при увеличении x или y. Это сигнализирует об отсутствии постоянства волатильности ошибки σ_ε. В этом случае говорят о *гетероскедастичности* данных. Иногда полоса ошибок достаточно причудливо выгибается, что, скорее всего, связано с нелинейной зависимостью между x и y.

2.12 Зависит ли y от x?

• Корреляционный коэффициент играет очень важную роль в различных приложениях экономических и финансовых теорий, характеризуя наличие связи между величинами. Поиск такой связи – первоочередная задача любого исследования. Если величины оказываются зависимыми, то вероятность того, что x и y принимают некоторые значения, очевидно, будет уже функцией двух переменных $P(x, y)$. Для вычисления средних необходимо суммировать по всем возможным значениям как x, так и y с весовыми коэффициентами, равными плотности вероятности $P(x, y)$. Если между величинами никакой связи нет (не только линейной), то совместная плотность вероятности $P(x, y)$ равна произведению распределений каждой из величин $P_1(x)P_2(y)$ (стр.453):

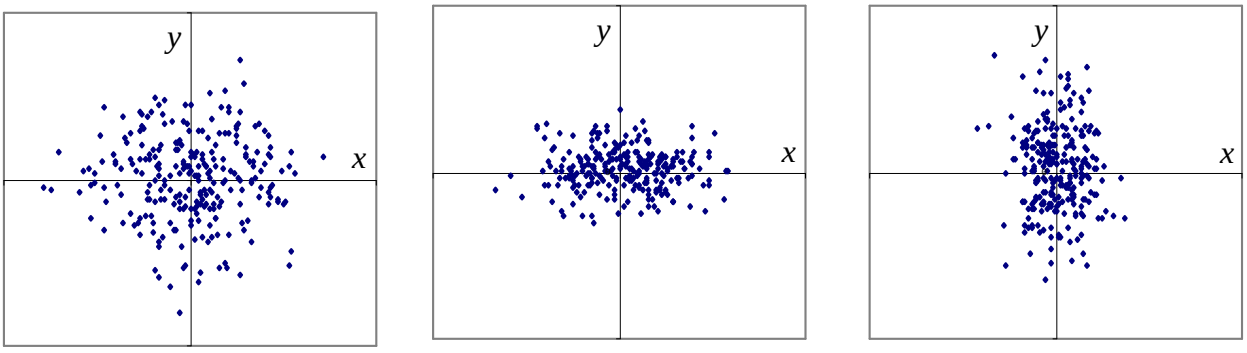
$$\langle f(x)g(y)\rangle = \int_{-\infty}^{+\infty} f(x)g(y)P(x,y)dxdy = \int_{-\infty}^{+\infty} f(x)P_1(x)dx \int_{-\infty}^{+\infty} g(y)P_2(y)dy.$$

Поэтому среднее произведение двух независимых величин равно произведению их средних: $\langle f(x)g(y)\rangle = \langle f(x)\rangle \langle g(y)\rangle$. В частности, $\langle xy\rangle = \langle x\rangle \langle y\rangle$. На самом деле, если задуматься, это достаточно естественный признак отсутствия статистической связи между x и y. В этом случае коэффициент корреляции равен нулю. Действительно, ковариация:

$$\gamma(x, y) = \langle (x - \bar{x})(y - \bar{y})\rangle = \langle x - \bar{x}\rangle \langle y - \bar{y}\rangle = 0,$$

если x и y – независимые случайные величины. Однако, как мы увидим в следующей главе, обратное, вообще говоря, неверно. Из равенства $c(x, y) = 0$ автоматически не следует отсутствие стохастической зависимости.

При построении на двумерной плоскости пар точек, между которым нет стохастической зависимости, мы получим размытое облако:



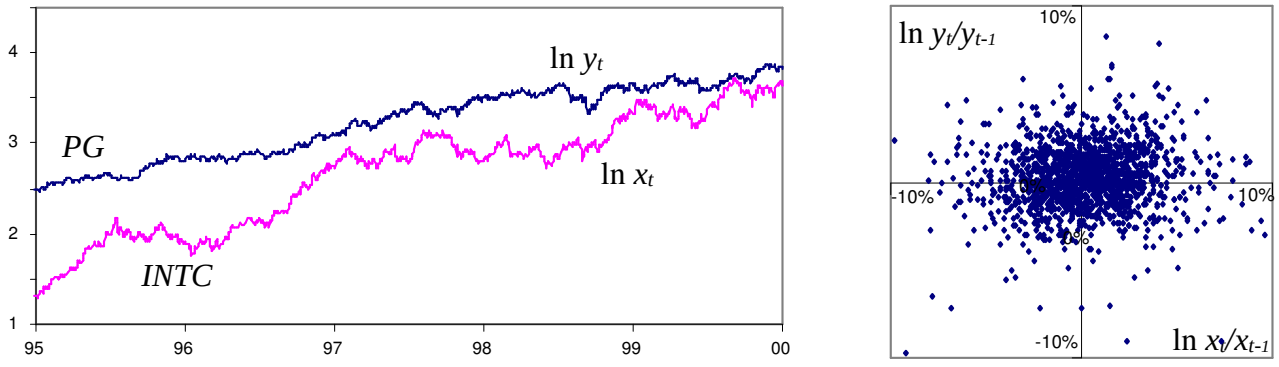
Если распределение x и y имеет максимум, то это облако будет иметь "сгущения" в окрестности точки $(\bar{x}, \bar{y})$. Если волатильность σ_x больше, чем σ_y, то облако будет вытянуто вдоль оси x, иначе – вдоль оси y.

• Корреляция между двумя величинами x, y не всегда означает наличие *причинной связи* $y = f(x)$ или $x = g(y)$. Например, может существовать третья величина z, оказывающая влияние и на x, и на y, синхронизируя их поведение. Так, общий спад мировой экономики может оказывать одинаковое воздействие на две не связанные друг с другом, но экспортно-ориентированные отрасли экономики.

"Ложная" корреляция также возникает, когда две величины имеют явно выраженный восходящий или нисходящий тренд. В этом случае между ними будет появляться заметная корреляция. Однако эта корреляция характеризует лишь наличие детерминированной составляющей роста. Более интересным является возможная стохастическая зависимость синхронности отклонений величин от этой детерминированной динамики.

Например, на фондовом рынке обычно бессмысленно вычислять корреляцию между ценами акций. Существенно информативнее будет корреляция между их *относительными* изменениями (доходностями) за некоторый период.

Рассмотрим в качестве примера две растущие акции PG и INTC за период 1995-2000 г. Динамика цен x_t и y_t приведена на левой диаграмме. Справа дано распределение их ежедневных доходностей $r_t = \ln(x_t/x_{t-1})$.



Несмотря на то, что акции синхронно растут, их доходности между собой связаны очень слабо. В данном случае, если мы вычислим коэффициент корреляции между ценами, то получим значение 0.80, тогда как корреляция между доходностями равна 0.10 (правый рисунок).

Заметим, что вычисление корреляций между ценами акций имеет смысл только в том случае, если они испытывают нерегулярные колебания с большими периодами, без явно выраженного повышающего или понижающего тренда. В этом случае корреляционный коэффициент даёт дополнительную информацию о связи этих колебаний. Может оказаться, что ежедневные доходности акций между собой связаны очень плохо, однако на длительных интервалах времени их цены испытывают существенно скоррелированные колебания. Необходимо, правда, обязательно убедиться, что эта скоррелированность не обусловлена детерминированной составляющей роста.

• Рассмотрим теперь стохастические зависимости между реальными экономическими и финансовыми величинами.

Пусть нас интересует прогноз значения некоторого макроэкономического параметра y_t. Скорее всего, он будет зависеть от всей предыстории y и предыстории множества других макроэкономических величин x_t. Подобные уравнения для x_t, y_t, .. образуют достаточно сложную, вообще говоря, нелинейную модель.

Как для записи этих уравнений, так и для создания рефлекторного понимания (иногда иллюзорного) экономических закономерностей исследователи ищут простые линейные связи между различными величинами. Для этого сложную многофакторную модель заменяют на линейную зависимость между текущим y и значением x, отстоящим на s периодов в прошлое:

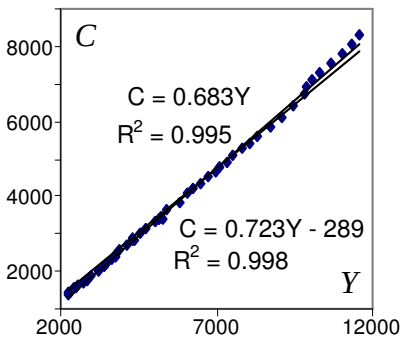
$$y_t = a + b \cdot x_{t-s} + \varepsilon_t.$$

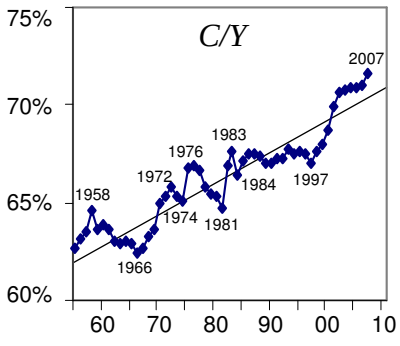
Часть шума ε_t обусловлена внешними факторами, однако другая его составляющая возникает из-за сильного упрощения модели. Рассмотрим два примера.

• Чем больше произведено валового внутреннего продукта Y в стране, тем, скорее всего, больше будет потрачено C населением. Поэтому в большинстве экономических моделей предполагается линейная *функция потребления*:

$$C = C(Y) = a + b \cdot Y.$$

Ежегодные данные в фиксированных ценах по ВВП и потреблению в США за период 1955 - 2007 г. имели следующий вид (левый график):

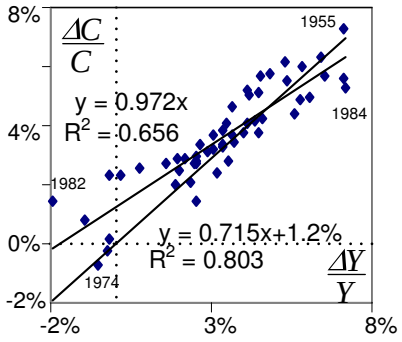


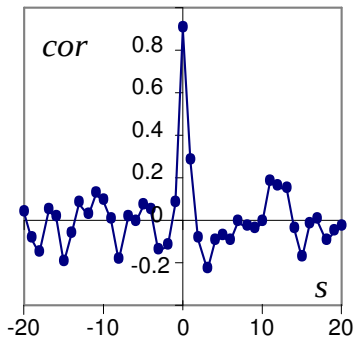


Подобное представление линейной зависимости малоинформативно и создаёт впечатление очень тесной линейной связи. Так как можно положить $a = 0$, посмотрим, как изменялось со временем отношение C/Y (правый график). Диаграмма уже заметно информативнее и зависимость стала динамической.

Ещё более тонкие детали функции потребления проявляются при анализе связи относительных изменений. В линейной модели с $a = 0$: $C = b \cdot Y$, поэтому должно выполняться равенство относительных изменений: $\Delta C/C = \Delta Y/Y$.

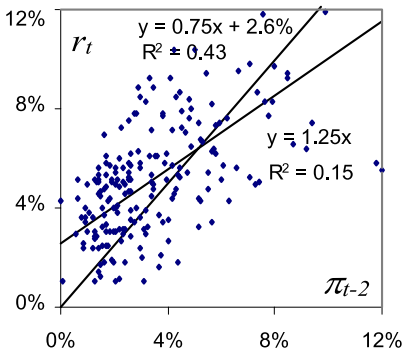
Нанесём эти величины (в процентах) на точечную диаграмму (левый график). Видно, что линейная связь стала менее выраженной:

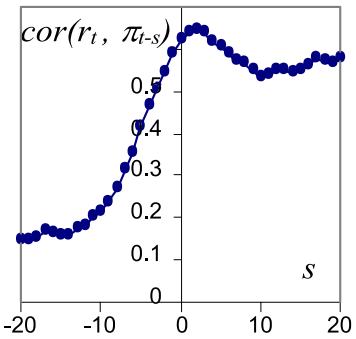




Правая диаграмма показывает зависимость корреляционного коэффициента между изменениями s *лет назад*: $cor = c(\Delta C_t/C_t, \Delta Y_{t-s}/Y_{t-s})$. Острый максимум корреляции 0.67 при $s = 0$ говорит о том, что американцы тут же тратят то, что получают ☺.

• В качестве второго примера рассмотрим зависимость между средней за квартал процентной ставкой r_t и квартальными изменениями индекса цен (дефлятор) $\pi_t = \Delta P_t/P_t$ Америки (левый график). Вычислим $cor = c(r_t,\ \pi_{t-s})$, со сдвигом (лагом) в s кварталов (правый график). Падение коэффициента корреляции при $s < 0$ означает, что именно процентная ставка зависит от инфляции, а не наоборот. Это связано с тем, что центробанк поднимает учётную ставку постепенно, наблюдая за произведенным эффектом. Левый график построен для процентной ставки r_t и инфляции π_{t-2}:





Стоит обратить внимание на два любопытных факта. Во-первых, нет существенного пика на графике корреляции в окрестности некоторого значения лага s. В этом смысле он качественно отличается от коррелограммы для $C(Y)$. Это сигнализирует о недостаточной адекватности дискретной *локальной* модели. Во-вторых, мы на самом деле видим явное проявление регулятивного воздействия на экономику, которое является не экономической, а политической закономерностью. Обратная же экономическая связь $s < 0$, на основании которой ☺ и принимается политическое решение по повышению процентной ставки, имеет достаточно слабый корреляционный коэффициент.

Математика финансовых рынков

Сергей С. Степанов

Последняя версия находится на сайте http://synset.com. Все обнаруженные ошибки и замечания просьба присылать по почте: phys@synset.com. (c) 2009-2012, Печать: 2012-06-20