

С.В. БУЛАШЕВ

СТАТИСТИКА ДЛЯ ТРЕЙДЕРОВ

ББК 60.6

Б 91

Булашев С.В.

Б 91 Статистика для трейдеров. -М.: Компания Спутник+,
2003. - 245с.

ISBN 5-93406-577-7

В этой книге сделана попытка систематизированно рассмотреть практические методы статистики применительно к финансам. Наибольший интерес данная книга может представлять для трейдеров/портфельных менеджеров, то есть специалистов, принимающих самостоятельные решения на финансовых рынках в условиях неопределенности, а также для студентов экономических и финансовых вузов. Изложение материала начинается с базовых понятий, и постепенно переходит к достаточно сложным методам, применяющимся при анализе инвестиционных рисков. В книге содержится большое количество практических алгоритмов вычисления и оптимизации различных финансовых стохастических переменных.

ББК 60.6

ISBN 5-93406-577-7

© Булашев С.В., 2003

ПРЕДИСЛОВИЕ	10
1. ВЕРОЯТНОСТНОЕ ОПИСАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН	13
1.1. Введение.	13
1.2. Случайное событие. Вероятность.	13
1.3. Случайная величина.	14
1.4. Законы распределения случайной величины.	15
1.5. Показатели центра распределения.	16
1.6. Моменты распределения.	17
1.7. Показатели меры рассеяния.	18
1.8. Показатели формы распределения - коэффициент асимметрии.	21
1.9. Показатели формы распределения - эксцесс.	23
1.10. Плотность распределения функции от случайной величины.	24
1.11. Математическое ожидание функции от случайной величины.	26
1.12. Линейное преобразование случайной величины.	27
1.13. Общие свойства случайных величин с произвольным законом распределения.	28
2. АНАЛИТИЧЕСКИЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН	30
2.1. Введение.	30
2.2. Биномиальное распределение.	30
2.3. Распределение Пуассона.	32
2.4. Равномерное распределение.	33
2.5. Нормальное распределение.	34
2.6. Логнормальное распределение.	37
2.7. Распределение Лапласа.	38

2.8.	Распределение Коши.	39
2.9.	Распределение Парето.	40
2.10.	Обобщенное экспоненциальное распределение.	41
2.11.	Поиск интегральной функции распределения путем численного интегрирования плотности распределения.	43
2.12.	Поиск интегральной функции распределения путем разложения плотности распределения в ряд с последующим аналитическим интегрированием этого ряда.	46
2.13.	Моделирование с помощью равномерного распределения случайных чисел с произвольной плотностью распределения.	48
	Приложение 2.1. Гамма-функция Эйлера.	51
3.	СПЕЦИАЛЬНЫЕ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ	53
3.1.	t-распределение Стьюдента.	53
3.2.	χ^2 -распределение.	55
3.3.	F-распределение.	57
4.	ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ПО ВЫБОРКЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ	59
4.1.	Введение.	59
4.2.	Оценки центра распределения.	59
4.3.	Оценка дисперсии и среднеквадратичного отклонения.	62
4.4.	Оценка коэффициента асимметрии и эксцесса.	62
4.5.	Исключение промахов из выборки.	63
5.	СТАТИСТИЧЕСКИЕ ВЫВОДЫ	65
5.1.	Введение.	65
5.2.	Выборочное распределение выборочной средней.	65

5.3.	Доверительный интервал для генеральной средней.	66
5.4.	Выборочное распределение выборочной дисперсии.	67
5.5.	Доверительный интервал для генеральной дисперсии.	68
5.6.	Статистическая проверка гипотез.	69
5.7.	Проверка гипотез о величине генеральной средней.	70
5.8.	Проверка гипотез о величине генеральной дисперсии.	74
6.	ИДЕНТИФИКАЦИЯ ЗАКОНА РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ	79
6.1.	Введение.	79
6.2.	Группировка данных. Оптимальное число интервалов группировки.	79
6.3.	Построение гистограммы распределения.	81
6.4.	Гистограмма логарифмов относительных изменений индекса РТС.	84
6.5.	Использование критериев согласия при идентификации закона распределения случайной величины.	87
7.	КОРРЕЛЯЦИЯ СЛУЧАЙНЫХ ВЕЛИЧИН	91
7.1.	Введение.	91
7.2.	Функция регрессии.	91
7.3.	Линейная корреляция.	93
7.4.	Коэффициент корреляции. Ковариация.	95
7.5.	Математическое ожидание и дисперсия линейной комбинации случайных величин.	96
7.6.	Оценка ковариации и коэффициента корреляции по выборке случайных величин.	97
7.7.	Оценка коэффициентов линейной регрессии по выборке случайных величин.	99
7.8.	Линейная регрессия как наилучшая оценка по методу наименьших квадратов.	99

8.	РЕГРЕССИОННЫЙ АНАЛИЗ	101
8.1.	Введение.	101
8.2.	Выбор вида математической модели.	101
8.3.	Расчет параметров математической модели.	103
8.4.	Сущность метода наименьших квадратов.	105
8.5.	Свойства ошибок метода наименьших квадратов.	106
8.6.	Оценка параметров однофакторной линейной регрессии.	107
8.7.	Коэффициент детерминации.	110
8.8.	Необратимость решений МНК.	114
8.9.	Статистические выводы о величине параметров однофакторной линейной регрессии.	115
8.10.	Статистические выводы о величине коэффициента детерминации.	118
8.11.	Полоса неопределенности однофакторной линейной регрессии.	120
8.12.	Прогнозирование на основе однофакторной линейной регрессии.	121
8.13.	Проверка допущений МНК.	122
8.14.	Сведение нелинейной функциональной зависимости к линейной путем преобразования данных.	128
8.15.	Функция регрессии как комбинация нескольких функций.	130
9.	АНАЛИЗ ФУРЬЕ	132
9.1.	Введение.	132
9.2.	Численный анализ Фурье.	132
9.3.	Амплитудно-частотная характеристика.	133
9.4.	Пример выделения основной гармоник с помощью анализа Фурье.	134

10.	ПРИМЕНЕНИЕ МНК ПРИ ИЗУЧЕНИИ ДИНАМИЧЕСКИХ РЯДОВ	138
10.1.	Введение.	138
10.2.	Модель динамики цен активов.	139
10.3.	Определение тренда.	141
10.4.	Статистические выводы о величине параметров регрессии.	145
10.5.	Полоса неопределенности рассеяния эмпирических данных относительно линии регрессии.	147
10.6.	Проверка допущений МНК.	148
11.	СГЛАЖИВАНИЕ ДИНАМИЧЕСКИХ РЯДОВ	155
11.1.	Введение.	155
11.2.	Типы скользящих средних.	155
11.3.	Простая скользящая средняя.	156
11.4.	Взвешенная скользящая средняя.	156
11.5.	Экспоненциальная скользящая средняя.	157
11.6.	Точки пересечения экспоненциально сглаженных кривых.	160
11.7.	Выбор величины показательного процента для экспоненциальной скользящей средней.	161
11.8.	Экспоненциальная скользящая средняя с переменным показательным процентом.	162
11.9.	Дисперсия скользящих средних.	162
12.	АДАПТИВНОЕ МОДЕЛИРОВАНИЕ ДИНАМИЧЕСКИХ РЯДОВ	165
12.1.	Введение.	165
12.2.	Адаптивное моделирование линейного тренда с помощью экспоненциальных скользящих средних.	165
12.3.	Адаптивное моделирование параболического тренда	169

	с помощью экспоненциальных скользящих средних.	
12.4.	Выбор величины показательного процента при адаптивном моделировании.	173
12.5.	Адаптивное моделирование с переменным показателем процента.	174
13.	МЕХАНИЧЕСКИЕ ТОРГОВЫЕ СИСТЕМЫ	176
13.1.	Введение.	176
13.2.	Механический и интуитивный подход к торговле.	177
13.3.	Свойства МТС.	178
13.4.	Минимальное число сделок.	180
13.5.	Тестирование МТС.	182
13.6.	Отчет о величине торгового счета.	182
13.7.	Сгруппированный отчет о величине торгового счета.	183
13.8.	Отчет о сделках.	184
13.9.	Сводный отчет.	187
13.10.	Математическое ожидание дохода сделки.	192
13.11.	Кумулятивная кривая дохода сделок.	196
13.12.	Вероятность получения убытка в серии последовательных сделок.	197
13.13.	Вероятность разорения в серии последовательных сделок.	201
14.	УПРАВЛЕНИЕ КАПИТАЛОМ	204
14.1.	Введение.	204
14.2.	Ограничение суммы убытка в сделке.	204
14.3.	Ограничение процента убытка в сделке.	205
14.4.	Максимизация средней величины дохода МТС.	206
14.5.	Оптимизация соотношения дохода и риска МТС.	211
14.6.	Анализ соотношения скользящих средних от кумуля	214

тивной кривой дохода сделок.	
14.7. Критерий серий.	216
14.8. Увеличение объема выигрывающей позиции.	218
15. УПРАВЛЕНИЕ РИСКОМ ПОРТФЕЛЯ НА ОС- НОВЕ АНАЛИЗА КОВАРИАЦИЙ АКТИВОВ	222
15.1. Введение.	222
15.2. Корреляция активов и риск портфеля.	222
15.3. Понижение риска портфеля. Диверсификация.	223
15.4. Граница эффективности.	226
15.5. Постановка задачи по оптимизации портфеля.	229
15.6. Введение ограничений на состав и веса активов в портфеле (лимитов).	230
15.7. Численное решение задачи оптимизации портфеля с учетом лимитов методом Монте-Карло.	231
16. УПРАВЛЕНИЕ РИСКОМ ПОРТФЕЛЯ НА ОС- НОВЕ АНАЛИЗА КВАНТИЛЬНЫХ МЕР РИСКА	236
16.1. Введение.	236
16.2. Понятие Value-at-risk и Shortfall-at-risk	237
16.3. Вычисление Value-at-risk и Shortfall-at-risk	239
16.4. Оптимизация портфеля с учетом Value-at-risk и Shortfall-at-risk.	243
РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА	244

ПРЕДИСЛОВИЕ

В последние годы значительно увеличилось количество людей, сфера деятельности которых связана с работой на финансовых рынках. Для этих специалистов необходимо хорошее знание основ теории вероятности и математической статистики, так как результаты решения об инвестировании в различные финансовые инструменты (активы) всегда имеют ту или иную степень неопределенности. В этой книге сделана попытка систематизированно рассмотреть практические методы статистики применительно к финансам. Наибольший интерес данная книга может представлять для трейдеров/портфельных менеджеров, то есть специалистов, принимающих самостоятельные решения на финансовых рынках в условиях неопределенности. Изложение материала начинается с базовых понятий, и постепенно переходит к достаточно сложным методам, применяющимся при анализе инвестиционных рисков. В книге содержится большое количество практических алгоритмов вычисления и оптимизации различных финансовых стохастических переменных.

Данная книга состоит из 16-ти глав.

В 1-й главе рассмотрено понятие вероятности, случайного события, случайной величины, дано определение закона распределения случайной величины, а также изучены основные параметры законов распределения, такие как показатели центра распределения, показатели меры рассеяния, показатели формы распределения.

Во 2-й главе рассказано о наиболее употребительных законах распределения случайных величин и основных параметрах этих законов. Даны методы поиска функции распределения вероятности случайной величины в случае неинтегрируемой плотности вероятности, а также алгоритмы получения последовательностей случайных величин с произвольным законом распределения, что необходимо при моделировании случайных процессов.

В 3-й главе изучены специальные распределения вероятностей, используемые для проверки статистических гипотез и при определении доверительных интервалов для случайных величин.

4-я глава посвящена методам оценки по эмпирической выборке параметров распределения случайной величины, указаны формулы для оценки центра распределения, дисперсии и показателей формы распределения, а также практические приемы удаления аномальных значений (промахов) из выборки.

В 5-й главе рассказано о методах проверки статистических гипотез и методах определения доверительных интервалов для случайных величин.

6-я глава посвящена вопросу о том, как по эмпирической выборке идентифицировать закон распределения случайной величины. Подробно рассмотрена проблема группировки данных, то есть расчет оптимального количества интервалов группировки и оптимальной ширины интервала, а также построения по сгруппированным данным гистограммы распределения таким образом, чтобы максимально возможное сглаживание случайного шума сочеталось с минимальным искажением от сглаживания самого распределения.

В 7-й главе рассмотрено понятие линейной корреляционной связи между случайными величинами.

8-я глава посвящена изучению регрессионного анализа, то есть методам расчета параметров математической модели, связывающей различные стохастические переменные.

В 9-й главе излагается метод аппроксимации эмпирической зависимости тригонометрическим рядом Фурье. Даны формулы, позволяющие по реальной выборке вычислить коэффициенты Фурье, амплитуду и фазу гармоник. Рассказано, как строится амплитудно-частотная характеристика разложения, и как она используется для выделения гармоник с максимальной амплитудой.

В 10-й главе рассмотрено применение регрессионного анализа при изучении динамических (временных) рядов.

В 11-й главе рассказано о методах сглаживания динамических рядов, базирующихся на расчете скользящих средних. Рассмотрены различные типы скользящих средних и даны их сравнительные характеристики.

В 12-й главе изучены методы адаптивного моделирования динамических рядов, которые основаны на экспоненциальном сглаживании (экспоненциальной скользящей средней). Преимуществом этих методов является учет временной ценности данных и, следовательно, постоянное адаптирование к изменяющимся уровням динамического ряда, что имеет решающее значение при моделировании и прогнозировании волатильных рядов.

13-я глава посвящена механическим торговым системам, то есть алгоритмам, которые формализуют правила открытия и закрытия позиций в биржевой торговле. Подробно рассмотрены отчеты о работе торговой системы и даны практические рекомендации о том, как по величине, разбросу и устойчивости показателей системы сделать вывод о ее качестве.

14-я глава является продолжением предыдущей. В ней изучены алгоритмы вычисления доли участвующего в конкретной сделке капитала, которые максимизируют показатели динамики торгового счета.

В 15-й главе рассматриваются вопросы, связанные с оптимизацией портфеля активов. Изучается влияние корреляции между отдельными парами активов на общий риск портфеля, при этом в качестве меры риска принимается дисперсия (или среднеквадратичное отклонение). Рассказано о том, что такое эффективная диверсификация и как общий риск портфеля, составленного из произвольного количества активов, можно разделить на несистематический (диверсифицируемый) риск и рыночный (не диверсифицируемый) риск. Поставлена задача по оптимизации портфеля с учетом ограничений на состав и веса активов в портфеле (лимитов), и приведен алгоритм поиска решений этой задачи методом Монте-Карло.

16-я глава посвящена изучению квантильных мер риска портфеля из произвольного количества активов и управления риском портфеля на основе их анализа.

Все замечания по содержанию и оформлению книги просьба направлять автору по адресу ilion@online.ru

1. ВЕРОЯТНОСТНОЕ ОПИСАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН

1.1. Введение.

Теория вероятностей играет значительную роль во многих областях человеческой деятельности, в том числе в финансах. Это связано с тем, что результаты решения об инвестировании в финансовые инструменты (активы) всегда имеют ту или иную степень неопределенности.

В биржевых торгах по различным активам принимают участие большое количество инвесторов и спекулянтов. Каждый из участников имеет свое представление о том, куда движется рынок, у каждого из них свой горизонт инвестирования и своя технология работы на рынке. Из-за столкновения интересов большого количества людей цены активов приобретают случайный характер. Следствием этого является невозможность точного предсказания будущей цены. Прогноз становится возможным только в вероятностном смысле.

С другой стороны, результаты инвестирования в инструменты с фиксированной доходностью также являются неопределенными из-за того, что существует риск невыполнения эмитентом (заемщиком) своих обязательств.

В этой главе мы рассмотрим на качественном уровне понятие вероятности, случайного события, случайной величины, дадим определение закона распределения случайной величины. Далее будут изучены основные параметры законов распределения, такие как показатели центра распределения, показатели меры рассеяния, показатели формы распределения.

1.2. Случайное событие. Вероятность.

Случайным событием называется такое событие, которое может как произойти, так и не произойти при соблюдении определенного комплекса условий. Будем предполагать, что указанный комплекс условий может быть воспроизведен неограниченное количество раз. *Испытанием* будем называть каждое осуществление этого комплекса условий.

Относительной частотой случайного события называется отношение количества случаев появления этого события M к общему числу проведенных испытаний N .

Опыт показывает, что при многократном повторении испытаний относительная частота M/N случайного события обладает устойчивостью. В разных достаточно длинных сериях испытаний относительные частоты случайного события группируются вокруг некоторого определенного числа. Устойчивость относительной частоты может быть объяснена как проявление объективного свойства случайного события, которое заключается в существовании определенной степени его возможности.

Таким образом, степень возможности случайного события можно описать числом. Это число называется *вероятностью случайного события*. Именно вокруг вероятности группируются относительные частоты данного случайного события. Относительная частота и вероятность случайного события являются безразмерными величинами, которые могут принимать значения от 0 до 1. Вероятность является первичным, базовым понятием, и в общем случае ее нельзя определить через более простые термины.

1.3. Случайная величина.

Случайной величиной называется такая величина, которая принимает те или иные значения с определенными вероятностями. Случайные величины могут быть дискретными и непрерывными.

Дискретной случайной величиной называется такая величина, все возможные значения которой образуют конечную или бесконечную последовательность чисел (x_1, x_2, \dots, x_n) и принятие ей каждого из указанных значений есть случайное событие, характеризующееся соответствующей вероятностью (p_1, p_2, \dots, p_n) . При этом должно соблюдаться условие нормирования, то есть $\sum_n p_n = 1$.

Непрерывной случайной величиной называется такая величина, все возможные значения которой целиком заполняют некоторый промежуток и попадание в любой интервал (x_1, x_2)

есть случайное событие, характеризующееся соответствующей вероятностью $P\{x_1 \leq x \leq x_2\}$. При этом вероятность достоверного события $P\{-\infty \leq x \leq +\infty\} = 1$.

Генеральной совокупностью будем называть все возможные значения, которые может принимать случайная величина.

1.4. Законы распределения случайной величины.

Для характеристики вероятности появления различных значений случайной величины используют *законы распределения вероятностей случайной величины*. При этом различают два вида представления законов распределения: интегральный и дифференциальный.

Интегральным законом, или *функцией распределения вероятностей* случайной величины X , называется функция, значение которой для любого x является вероятностью события, заключающегося в том, что случайная величина X принимает значения, меньшие x , то есть функция $F(x) = P\{X < x\}$. Функция распределения вероятностей $F(x)$ обладает следующими свойствами:

- 1) $0 \leq F(x) \leq 1$ для любого x
- 2) $F(x_1) \leq F(x_2)$, если $x_1 \leq x_2$
- 3) $F(-\infty) = 0, F(+\infty) = 1$

Для случайной величины с непрерывной и дифференцируемой функцией распределения вероятностей $F(x)$ можно найти *дифференциальный закон распределения вероятностей*, выражаемый как производная $F(x)$, то есть $p(x) = dF(x)/dx$. Эта зависимость называется *плотностью распределения вероятностей*. Плотность распределения $p(x)$ обладает следующими свойствами:

- 1) $p(x) \geq 0$ для любого x
- 2) $P\{X < x\} \equiv F(x) = \int_{-\infty}^x p(t)dt$
- 3) $P\{a \leq X < b\} = F(b) - F(a) = \int_a^b p(t)dt$

$$4) \int\limits_{-\infty}^{+\infty} p(x)dx = 1$$

Распределение называется *предельно пологим*, если при $x \rightarrow \pm\infty$ его плотность вероятности $p(x) \cong 1/|x|^{1+\delta}$, где δ - сколь угодно малое положительное число. При более пологих, чем $1/|x|^{1+\delta}$ спадах, площадь под кривой бесконечна, то есть не выполняется условие нормирования, и такие кривые не могут описывать плотность распределения вероятностей.

1.5. Показатели центра распределения.

Координата центра распределения определяет положение случайной величины на числовой оси. Дать однозначное определение этого понятия невозможно. Центр распределения может быть найден несколькими способами:

- как медиана распределения,
- как мода распределения,
- как математическое ожидание.

Медиана

Наиболее общим, а следовательно наиболее фундаментальным, является определение центра распределения согласно принципу симметрии, то есть как такой точки на оси x , слева и справа от которой вероятности появления случайной величины одинаковы и равны 0.5. Такой показатель центра распределения называется *медианой*. В отличие от других показателей центра, медиана существует у любого распределения. Медиану обычно обозначают как *Me*.

Мода

Точка на оси x , соответствующая максимуму кривой плотности распределения, называется *модой*, то есть мода – это наиболее вероятное значение случайной величины. Однако, мода существует не у всех распределений. В качестве примера можно привести равномерное распределение. В этом случае определение центра распределение как моды невозможно. Моду обычно обозначают как *Mo*.

Математическое ожидание

Наиболее часто используемым методом оценки центра распределения является *математическое ожидание*. Преимущественное использование математического ожидания объясняется тем, что это единственная оценка, которую можно выразить аналитически.

Математическое ожидание обозначается как μ и вычисляется по формулам:

- для дискретного распределения

$$M(x) \equiv \mu = \sum_n x_n p_n$$

- для непрерывного распределения

$$M(x) \equiv \mu = \int_{-\infty}^{+\infty} x p(x) dx$$

Необходимо отметить, что математическое ожидание существует только у тех распределений, у которых при $x \rightarrow \pm\infty$ плотность вероятности спадает как $1/|x|^{2+\delta}$ или круче, где δ - сколь угодно малое положительное число. При более пологих, чем $1/|x|^{2+\delta}$ спадах, математическое ожидание не существует, так как определяющий его интеграл расходится.

1.6. Моменты распределения.

Для описания свойств распределений нам понадобится понятие момента распределения. Существуют два типа моментов: начальные и центральные. *Начальным* называется момент распределения, найденный без исключения систематической составляющей. Соответственно, *центральный* является момент, вычисленный с исключением систематической составляющей.

Начальный момент k -го порядка вычисляется по формулам:

- для дискретного распределения

$$M_k = \sum_n x_n^k p_n$$

- для непрерывного распределения

$$M_k = \int_{-\infty}^{+\infty} x^k p(x) dx$$

Первый начальный момент был уже рассмотрен выше - это математическое ожидание.

Центральный момент k -го порядка вычисляется по формулам:

- для дискретного распределения

$$m_k = \sum_n (x_n - \mu)^k p_n$$

- для непрерывного распределения

$$m_k = \int_{-\infty}^{+\infty} (x - \mu)^k p(x) dx$$

Понятие моментов распределения будет использовано при изучении показателей рассеяния случайной величины и показателей формы распределения.

1.7. Показатели меры рассеяния.

Оценив величину центра распределения, нам необходимо иметь представление, как случайная величина рассеяна вокруг этой точки. Для оценки меры рассеяния используются, как правило, два способа:

- квантильное отклонение случайной величины,
- дисперсия и среднеквадратичное отклонение случайной величины.

Квантильное отклонение

Площадь, заключенная под кривой плотности распределения $p(x)$, согласно правилу нормирования, равна единице, то есть отражает вероятность всех возможных событий.

Выберем точку X_1 на оси x таким образом, чтобы площадь под кривой $p(x)$ слева от точки X_1 была бы равна, например, 5% от общей площади, то есть вероятность того, что случайная величина меньше, чем X_1 составляет 0.05. В этом случае говорят, что X_1 - это 5%-ная квантиль распределения. Ее удобно обозначить как $X_1 = X_{0.05}$.

Выберем далее точку X_3 на оси x таким образом, чтобы площадь под кривой $p(x)$ слева от точки X_3 была бы равна 95% от общей площади, то есть вероятность того, что случайная величина

меньше, чем X_3 составляет 0.95. Тогда X_3 - это 95%-ная квантиль распределения. Обозначим ее как $X_3 = X_{0.95}$.

Медиана распределения - это 50%-ная квантиль, так как она делит площадь под кривой $p(x)$ на две равные части. Медиану можно обозначить как $X_2 = X_{0.50}$.

Заметим, что точки $X_1 = X_{0.05}$ и $X_3 = X_{0.95}$ симметричны в том смысле, что

- во-первых, вероятность того, что случайная величина меньше X_1 , и вероятность того, что случайная величина больше X_3 , равны между собой,
- во-вторых, вероятность того, что случайная величина находится в интервале от X_1 до X_2 , и вероятность того, что случайная величина находится в интервале от X_2 до X_3 , также равны между собой.

Интервал значений x между $X_1 = X_{0.05}$ и $X_3 = X_{0.95}$ называют интерквантильным промежутком с 90%-ной вероятностью. Его протяженность $\Delta_{0.90} = X_{0.95} - X_{0.05}$. Половину указанного промежутка, которую будем называть квантильным отклонением с 90%-ной вероятностью, обозначим как $d_{0.90} = \Delta_{0.90} / 2$.

На основании вышеизложенного подхода можно ввести понятие квантильной оценки рассеяния случайной величины, то есть значения рассеяния с заданной доверительной вероятностью. Для симметричных распределений квантильное рассеяние с заданной доверительной вероятностью P - это такой интервал неопределенности $(X_{0.50} - d_P, X_{0.50} + d_P)$, внутри которого лежат $100P$ процентов всех значений случайной величины, а $100(1 - P)$ процентов лежат вне этого интервала.

Так как квантили, ограничивающие доверительный интервал, могут быть различными, при указании квантильной оценки рассеяния обязательно должна быть указана доверительная вероятность такой оценки. Квантильная оценка рассеяния применима для любых законов распределения случайной величины.

При рассмотрении квантильного отклонения, мы не случайно в качестве примера использовали отклонение с 90%-ной доверительной вероятностью. Дело в том, что величина $d_{0.90}$ обладает уникальным свойством, которое заключается в том, что только

квантильное отклонение $d_{0.90}$ имеет однозначное соотношение со среднеквадратичным отклонением σ (которое будет рассмотрено ниже) в виде $d_{0.90} \approx 1.6\sigma$ для очень широкого класса наиболее употребительных законов распределения. Поэтому, при отсутствии данных о виде закона распределения, для оценки квантильного отклонения рекомендуется пользоваться доверительной вероятностью, равной 0.90.

Дисперсия и среднеквадратичное отклонение

Если в качестве показателя центра распределения выбрано математическое ожидание, то в качестве меры рассеяния случайной величины используют дисперсию. *Дисперсия* - это среднее значение квадратов отклонений случайной величины от ее математического ожидания. Дисперсия является вторым центральным моментом распределения.

Дисперсия обозначается как D и вычисляется по формулам:

- для дискретного распределения

$$D = \sum_n (x_n - \mu)^2 p_n$$

- для непрерывного распределения

$$D = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$$

В формуле для дисперсии в качестве центра распределения использовано математическое ожидание. Это не случайно. Дело в том, что использование в качестве центра распределения математического ожидания минимизирует средний квадрат отклонения случайной величины от ее центра. При этом минимум среднего квадрата отклонений как раз и равен дисперсии. Дисперсия и математическое ожидание связаны соотношением:

$$D(x) = M(x^2) - [M(x)]^2$$

Дисперсия имеет размерность квадрата случайной величины. Поэтому для более наглядной характеристики рассеяния используют корень квадратный из дисперсии, который называется *среднеквадратичным отклонением (с.к.о.)*: $\sigma = \sqrt{D}$.

Дисперсия - наиболее широко применяемая оценка рассеяния случайных величин. Это связано с тем, что она обладает свойством аддитивности, то есть дисперсия суммы статистически независимых случайных величин равна сумме дисперсий этих величин, безотносительно к разнообразию законов распределения каждой из суммируемых величин и возможной деформации законов распределения при суммировании. Отметим, что среднеквадратичное отклонение не аддитивно.

Таким образом, для того, чтобы рассеяния случайных величин можно было суммировать аналитически, эти рассеяния должны быть представлены своими дисперсиями, а не квантильными (доверительными) отклонениями.

Однако, конечная дисперсия существует только у тех распределений, у которых при $x \rightarrow \pm\infty$ плотность вероятности спадает как $1/|x|^{3+\delta}$ или круче, где δ - сколь угодно малое положительное число. При более пологих, чем $1/|x|^{3+\delta}$ спадах, определяющий дисперсию интеграл расходится.

1.8. Показатели формы распределения - коэффициент асимметрии.

При изучении формы распределения случайной величины важно выяснить, симметрична ли относительно центра распределения кривая плотности вероятности. Показателем степени несимметричности этой кривой является безразмерная величина, называемая *коэффициентом асимметрии*. Коэффициент асимметрии обозначается как γ или As . Рассмотрим на качественном уровне понятие асимметрии.

В случае, если кривая плотности вероятности имеет крутой левый и пологий правый спад, говорят, что распределение имеет *положительную асимметрию*. В этом случае координаты показателей центра распределения располагаются на оси абсцисс, как правило, следующим образом:
мода < медиана < математическое ожидание.

Если кривая плотности вероятности имеет пологий левый и крутой правый спад, распределение имеет *отрицательную асимметрию*. В этом случае для показателей центра распределения имеем:
математическое ожидание < медиана < мода.

Наконец, у *симметричных* распределений, медиана, мода и математическое ожидание совпадают.

Разумеется, все вышесказанное о соотношении показателей центра, справедливо только для тех распределений, у которых существует мода и/или математическое ожидание. Напомним, что понятие медианы применимо к любому распределению.

Существует несколько методов для оценки коэффициента асимметрии.

Оценка коэффициента асимметрии с помощью квантилей распределения

Рассмотрим, например, интерквантильный промежуток с 90%-ной вероятностью. Напомним, что он образован с помощью 5%-ной и 95%-ной квантилей распределения. Тогда соответствующий коэффициент асимметрии вычисляется по следующей формуле:

$$\frac{(x_{0.95} - x_{0.50}) - (x_{0.50} - x_{0.05})}{x_{0.95} - x_{0.05}} = \frac{x_{0.95} + x_{0.05} - 2x_{0.50}}{x_{0.95} - x_{0.05}}$$

Разумеется, таким способом можно вычислить коэффициент асимметрии на любом интерквантильном промежутке, однако следует сказать, что подобная оценка будет зависеть от выбора интерквантильного промежутка, то есть, например, оценка на 90%-ном и на 50%-ном промежутках будут давать вообще говоря разные результаты. Достоинством данного метода является то, что с его помощью можно рассчитать коэффициент асимметрии для любого распределения.

Оценка коэффициента асимметрии с помощью третьего центрального момента распределения

Если в качестве показателя центра распределения выбрано математическое ожидание, то коэффициент асимметрии рассчитывают, используя третий центральный момент распределения.

В этом случае коэффициент асимметрии - это отношение третьего центрального момента (имеющего размерность куба случайной величины) к среднеквадратичному отклонению (размерность которого совпадает с размерностью случайной величины), возведенному в третью степень.

Коэффициент асимметрии вычисляется по формулам:

- для дискретного распределения

$$\gamma = \frac{\sum (x_n - \mu)^3 p_n}{\sigma^3}$$

- для непрерывного распределения

$$\gamma = \frac{\int_{-\infty}^{+\infty} (x - \mu)^3 p(x) dx}{\sigma^3}$$

1.9. Показатели формы распределения - эксцесс.

Чрезвычайно важным показателем формы распределения является безразмерный показатель, называемый *эксцессом*. Эксцесс обозначается как ε или Ex . Эксцесс характеризует:

- во-первых, остроту пика распределения,
- во-вторых, крутизну спада хвостов распределения.

Если за точку отсчета принять нормальное распределение (которое будет подробно рассмотрено ниже), то распределения плотности вероятности можно условно разделить на три группы:

- островершинные,
- средневершинные,
- плосковершинные.

Островершинные распределения характеризуются более выраженным, чем у нормального распределения, пиком и полого спадающими, "тяжелыми" хвостами.

Средневершинные распределения незначительно отличаются от нормального.

Плосковершинные распределения имеют слабо выраженный пик или совсем не имеют пика и, соответственно, моды. Кроме того, они характеризуются резко спадающими хвостами.

Определим эксцесс как отношение четвертого центрального момента распределения к среднеквадратичному отклонению, возведенному в четвертую степень. Эксцесс вычисляется по формулам:

- для дискретного распределения

$$\varepsilon = \frac{\sum (x_n - \mu)^4 p_n}{\sigma^4}$$

- для непрерывного распределения

$$\varepsilon = \frac{\int_{-\infty}^{+\infty} (x - \mu)^4 p(x) dx}{\sigma^4}$$

Для различных законов распределения эксцесс может иметь значение от 1 до $+\infty$. Нормальное распределение имеет эксцесс, равный трем.

Эксцесс удобно использовать для характеристики остроты пика и крутизны спадов хвостов распределения:

- островершинные распределения имеют эксцесс > 3 ,
- средневершинные распределения имеют эксцесс ≈ 3 ,
- плосковершинные распределения имеют эксцесс < 3 ,

Часто в качестве эксцесса используют отношение четвертого центрального момента к четвертой степени среднеквадратичного отклонения, за вычетом числа три. Однако здесь и далее мы будем рассчитывать эксцесс по приведенным выше формулам.

1.10. Плотность распределения функции от случайной величины.

Пусть X - это случайная величина, имеющая плотность распределения $p_x(x)$. Найдем плотность распределения $p_y(y)$ случайной величины Y , которая является функцией от X .

Пусть функция $y(x)$ монотонно возрастает. Тогда любой интервал (x_1, x_2) взаимно однозначно отображается на интервал (y_1, y_2) . Следовательно, вероятности попадания случайных величин X и Y в соответствующие интервалы равны. В применении к малым интервалам это означает равенство дифференциалов вероятности:

$$p_x(x)dx = p_y(y)dy$$

$$\text{Следовательно } p_y(y) = p_x[x(y)] \cdot \frac{dx}{dy}$$

где $x(y)$ - это функция, обратная функции $y(x)$.

Если функция $y(x)$ монотонно убывает, то положительному значению dx соответствует отрицательное значение dy .

Следовательно, в уравнении равенства дифференциалов нужно заменить dy на $-dy=|dy|$. Это приводит к более общей зависимости:

$$p_y(y) = p_x[x(y)] \cdot \left| \frac{dx}{dy} \right|$$

Для иллюстрации вышесказанного рассмотрим несколько примеров.

1) $y(x) = ax + b, a \neq 0$

В зависимости от знака параметра a эта функция может быть как монотонно возрастающей, так и монотонно убывающей. Переменные x и y определены на всей числовой оси.

$$x(y) = \frac{y-b}{a} \quad \frac{dx}{dy} = \frac{1}{a}$$

$$p_y(y) = \frac{1}{|a|} p_x\left(\frac{y-b}{a}\right)$$

2) $y(x) = x^3$

Эта функция является монотонно возрастающей. Переменные x и y определены на всей числовой оси.

$$x(y) = y^{1/3} \quad \frac{dx}{dy} = \frac{1}{3y^{2/3}}$$

$$p_y(y) = \frac{1}{3y^{2/3}} p_x(y^{1/3})$$

3) $y(x) = \ln(x)$

Эта функция является монотонно возрастающей. Переменная x определена на интервале от 0 до $+\infty$. Переменная y определена на всей числовой оси.

$$x(y) = e^y \quad \frac{dx}{dy} = e^y$$

$$p_y(y) = e^y p_x(e^y)$$

4) $y(x) = e^{-x}$

Эта функция является монотонно убывающей. Переменная x определена на всей числовой оси. Переменная y определена на интервале от 0 до $+\infty$.

$$x(y) = -\ln(y) = \ln(1/y) \quad \frac{dx}{dy} = -\frac{1}{y}$$

$$p_y(y) = \frac{1}{y} p_x(\ln(1/y)) \quad y > 0$$

$$p_y(y) = 0 \quad y \leq 0$$

5) $y(x) = x^2$

Эта функция монотонно убывает на интервале от $-\infty$ до 0 и монотонно возрастает на интервале от 0 до $+\infty$. Переменная x определена на всей числовой оси. Переменная y определена на интервале от 0 до $+\infty$.

$$x < 0: \quad x(y) = -y^{1/2} \quad \frac{dx}{dy} = -\frac{1}{2y^{1/2}}$$

$$x \geq 0: \quad x(y) = y^{1/2} \quad \frac{dx}{dy} = \frac{1}{2y^{1/2}}$$

Следовательно

$$p_y(y) = \frac{1}{2y^{1/2}} p_x(-y^{1/2}) + \frac{1}{2y^{1/2}} p_x(y^{1/2}) \quad y > 0$$

$$p_y(y) = 0 \quad y < 0$$

1.11. Математическое ожидание функции от случайной величины.

Математическое ожидание случайной величины Y , которая является функцией случайной величины X , может быть вычислено без нахождения плотности вероятности этой функции, то есть непосредственно по распределению случайной величины X .

Если обозначить математическое ожидание случайной величины Y как μ_y , то справедливы следующие формулы:

- для дискретного распределения

$$\mu_y \equiv M[y(x)] = \sum_n y(x_n) p_n$$

- для непрерывного распределения

$$\mu_y \equiv M[y(x)] = \int_{-\infty}^{+\infty} y(x)p(x)dx$$

Заметим, что в общем случае $\mu_y \neq y(\mu_x)$.

1.12. Линейное преобразование случайной величины.

В дальнейшем наиболее часто мы будем использовать линейное преобразование случайной величины, то есть преобразование вида $y(x) = ax + b$. В этом случае параметры распределения величин X и Y связаны соотношениями:

$$\mu_y = a\mu_x + b$$

$$D_y = a^2 \cdot D_x$$

$$\sigma_y = |a| \cdot \sigma_x$$

Одним из важнейших примеров линейного преобразования является преобразование случайной величины к стандартному виду (нормирование):

$$t \equiv t(x) = \frac{x - \mu_x}{\sigma_x}$$

То есть случайная величина x с произвольным математическим ожиданием и произвольной дисперсией преобразуется в случайную величину t с нулевым математическим ожиданием и единичной дисперсией и среднеквадратичным отклонением. Величина t называется стандартизированной (нормированной) случайной величиной.

1.13. Общие свойства случайных величин с произвольным законом распределения.

Независимо от закона распределения случайной величины существуют общие свойства распределений вероятностей. К ним можно отнести:

- неравенства, определяющие граничные значения вероятности попадания случайной величины в заданный интервал,
- законы больших чисел, определяющие свойства достаточно большого количества случайных величин.

Неравенство Чебышева

Неравенство Чебышева определяет граничное значение вероятности попадания случайной величины x с произвольным законом распределения, имеющей математическое ожидание μ и дисперсию σ^2 , в заданный интервал вокруг математического ожидания:

$$P\{|x - \mu| \geq \lambda\sigma\} \leq \frac{1}{\lambda^2} \qquad P\{|x - \mu| \leq \lambda\sigma\} \geq 1 - \frac{1}{\lambda^2}$$

Иными словами, вероятность того, что в некотором испытании случайная величина x окажется за пределами интервала, длина которого прямо пропорциональна с.к.о., убывает обратно пропорционально квадрату коэффициента пропорциональности λ .

Неравенство Чебышева определяет важность среднеквадратичного отклонения как характеристики рассеяния случайной величины относительно центра распределения.

Подставив в неравенство Чебышева несколько конкретных значений величины λ , получим, что для любых законов распределения с математическим ожиданием μ и дисперсией σ^2 справедливо:

$$\begin{aligned} \lambda = 1: \quad & P\{|x - \mu| \geq \sigma\} \leq 1 & P\{|x - \mu| \leq \sigma\} \geq 0 \\ \lambda = 2: \quad & P\{|x - \mu| \geq 2\sigma\} \leq 1/4 & P\{|x - \mu| \leq 2\sigma\} \geq 3/4 \\ \lambda = 3: \quad & P\{|x - \mu| \geq 3\sigma\} \leq 1/9 & P\{|x - \mu| \leq 3\sigma\} \geq 8/9 \end{aligned}$$

Законы больших чисел

Невозможно предвидеть, какое значение примет случайная величина в результате отдельного испытания. Однако, при достаточно большом количестве испытаний оценки по выборке параметров распределения случайных величин в достаточной степени утрачивают случайный характер. То же самое можно сказать и в отношении суммы большого количества случайных величин. При увеличении числа слагаемых колебания отдельных величин взаимно сглаживаются и закон распределения суммы приближается к нормальному

распределению. Различные утверждения, относящиеся к этим предельным случаям носят названия законов больших чисел.

Теорема Бернулли

Если в последовательности из N независимых испытаний вероятность p некоторого случайного события остается постоянной, то вероятность того, что отклонение эмпирической частоты этого события M/N от p не превзойдет заранее заданное число $\delta > 0$ стремится к единице:

$$\lim_{N \rightarrow \infty} P\left\{\left|\frac{M}{N} - p\right| < \delta\right\} = 1$$

Теорема Чебышева

Вероятность того, что отклонение среднего арифметического N независимых случайных величин с конечными дисперсиями от среднего арифметического их математических ожиданий не превзойдет заранее заданное число $\delta > 0$ стремится к единице

$$\lim_{N \rightarrow \infty} P\left\{\left|\frac{1}{N} \sum_{k=1}^N x_k - \frac{1}{N} \sum_{k=1}^N \mu_k\right| < \delta\right\} = 1$$

Из теоремы Чебышева следует, что с увеличением числа N среднее арифметическое случайных величин постепенно утрачивает характер случайной величины и все более стремится к константе.

Центральная предельная теорема (теорема Ляпунова)

Распределение суммы N независимых случайных величин с конечными дисперсиями и с произвольными законами распределения стремится к нормальному распределению при $N \rightarrow \infty$, если вклад отдельных слагаемых в сумму мал.

Теорема Ляпунова объясняет широкое распространение нормального закона распределения тем, что рассеяние случайных величин вызывается множеством случайных факторов, влияние каждого из которых ничтожно мало.

2. АНАЛИТИЧЕСКИЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

2.1. Введение.

В этой главе мы рассмотрим наиболее употребительные законы распределения случайных величин, а также основные параметры этих законов. Будут даны методы поиска функции распределения вероятности случайной величины в случае неинтегрируемой плотности вероятности, а также алгоритмы получения последовательностей случайных величин с произвольным законом распределения, что необходимо при моделировании случайных процессов. Особое внимание будет уделено обобщенному экспоненциальному распределению, которое наиболее пригодно при изучении ценообразования активов.

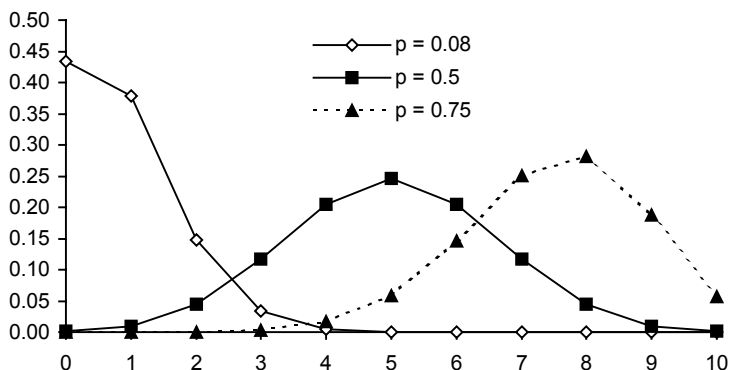
2.2. Биномиальное распределение.

Пусть некоторое событие может иметь только два исхода, которые назовем "успех" и "неудача", при этом вероятность успеха равна p , вероятность неудачи равна $(1 - p)$.

Если проводится серия из N независимых испытаний, то вероятность того, что успех в данной серии повторится x раз, а неудача $(N - x)$ раз, будет равна произведению числа способов, которыми можно выбрать x из N , на вероятность того, что сначала x раз повторится успех, а затем $(N - x)$ раз повторится неудача. Следовательно, вероятность x успехов в N независимых испытаниях равна:

$$p(x) = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x} \quad x = 0, 1, \dots, N$$

Данная формула описывает биномиальный закон распределения случайной величины. Из формулы непосредственно следует, что биномиальный закон полностью характеризуется двумя параметрами: количеством испытаний N и вероятностью успеха p . На рисунке приведена плотность биномиального распределения при $N = 10$ и различных значениях вероятности успеха p . Распределение является дискретным, поэтому точки соединены на графике лишь для наглядности.



Математическое ожидание и дисперсия биномиального распределения составляют:

$$\mu = Np$$

$$D = Np(1 - p)$$

Третий центральный момент данного распределения равен

$$m_3 = Np(1 - p)(1 - 2p)$$

Следовательно, коэффициент асимметрии составляет

$$\gamma = \frac{1 - 2p}{\sqrt{Np(1 - p)}}$$

Знак коэффициент асимметрии зависит от вероятности успеха p :

$\gamma < 0$, если $p > 0.5$

$\gamma = 0$, если $p = 0.5$

$\gamma > 0$, если $p < 0.5$

Если вероятность успеха p фиксирована, то коэффициент асимметрии $\gamma \rightarrow 0$ при количестве испытаний $N \rightarrow \infty$ для любой p .

Четвертый центральный момент данного распределения равен $m_4 = Np(1 - p)[1 + 3(N - 2)p(1 - p)]$. Следовательно, эксцесс составляет

$$\varepsilon = \frac{1 + 3(N - 2)p(1 - p)}{Np(1 - p)}$$

При количестве испытаний $N \rightarrow \infty$ эксцесс биномиального распределения стремится к числу три, то есть к эксцессу нормального распределения.

2.3. Распределение Пуассона.

Распределение Пуассона называют еще распределением редких событий. Ему подчиняются случайные величины, вероятность появления которых в отдельном испытании мала и постоянна.

Распределение Пуассона является предельным случаем биномиального распределения. Его можно применить, когда количество испытаний N достаточно велико, а вероятность успеха p мала, но так что произведение $\mu = Np$ - это некоторое конечное постоянное число.

Если мы обозначим математическое ожидание количества успехов за некоторый промежуток времени (или за некоторое количество испытаний) как μ , то вероятность получить x успехов за этот промежуток времени подчиняется распределению Пуассона:

$$p(x) = \frac{e^{-\mu} \mu^x}{x!} \quad x = 0, 1, 2, \dots$$

Данное распределение зависит от единственного параметра μ , который может принимать конечные положительные значения. Напомним, что величина x - это количество успехов, а потому дискретна.

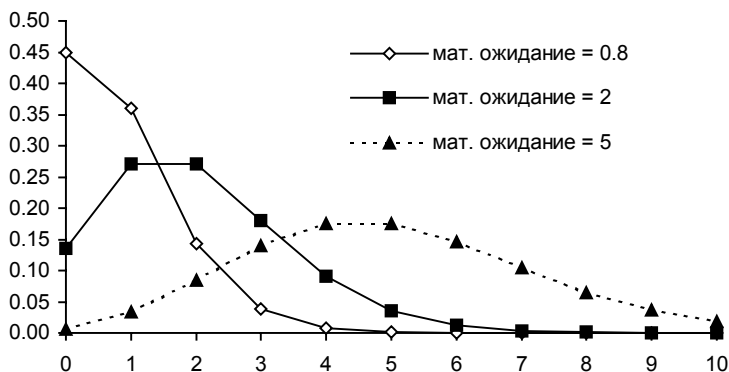
Из формулы для распределения Пуассона непосредственно следует, что $p(x+1)/p(x) = \mu/(x+1)$. Если $\mu < 1$, то $p(x+1) < p(x)$ при любом x и максимальное значение $p(x)$ достигается при $x = 0$. Если же $\mu > 1$, то $p(x)$ сначала растет с увеличением x , достигая максимума при $x \approx \mu$, а затем убывает.

Математическое ожидание и дисперсия распределения Пуассона равны μ .

Третий центральный момент m_3 также равен μ . Следовательно, *коэффициент асимметрии* составляет $\gamma = 1/\sqrt{\mu}$, то есть распределение Пуассона имеет положительную асимметрию. Асимметрия стремится к нулю при $\mu \rightarrow \infty$.

Четвертый центральный момент $m_4 = 3\mu^2 + \mu$. Следовательно, эксцесс составляет $\varepsilon = 3 + 1/\mu$. При $\mu \rightarrow \infty$ эксцесс распределения Пуассона стремится к числу три, то есть к эксцессу нормального распределения.

На рисунке приведена плотность распределения Пуассона при различных значениях математического ожидания. Распределение является дискретным, поэтому точки соединены на графике лишь для наглядности.



2.4. Равномерное распределение.

Если все значения непрерывной случайной величины в некотором интервале от a до b , равновероятны, то аналитически это можно записать в виде:

$$p(x) = 0 \quad x < a, x > b$$

$$p(x) = 1/(b-a) \quad a \leq x \leq b$$

Распределение с такой плотностью вероятности называется равномерным (равновероятным, прямоугольным). Данное распределение часто используют для качественного анализа статистических процессов.

Математическое ожидание и дисперсия равномерного распределения составляют:

$$\mu = (a + b) / 2$$

$$D = (b - a)^2 / 12$$

Медиана распределения совпадает с математическим ожиданием, *моды* не существует.

Коэффициент асимметрии и *эксцесс* равны: $\gamma = 0$, $\varepsilon = 1.8$.

Для равномерного распределения можно в явном виде найти функцию распределения вероятностей:

$$F(x) = 0 \quad x < a$$

$$F(x) = (x - a)/(b - a) \quad a \leq x \leq b$$

$$F(x) = 1 \quad x > b$$

Встроенный в компьютер генератор псевдослучайных чисел выдает числа, равномерно распределенные в интервале от 0 до 1. С их помощью можно моделировать случайные процессы с произвольной функцией распределения. Подробнее о том, как это делается, будет рассказано далее в этой главе.

2.5. Нормальное распределение.

Одним из важнейших распределений, встречающихся в статистике, является нормальное распределение (распределение Гаусса), относящееся к классу экспоненциальных. Плотность вероятности этого распределения:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad -\infty < x < +\infty$$

Распределение имеет вид симметричной колоколообразной кривой, распространяющейся по всей числовой оси. Распределение Гаусса зависит от двух параметров: (μ, σ) .

Математическое ожидание, *медиана* и *мода* данного распределения равны μ , а *дисперсия* σ^2 . Кривая плотности вероятности симметрична относительно математического ожидания. *Коэффициент асимметрии* и *эксцесс* равны $\gamma = 0$, $\varepsilon = 3$.

Часто плотность нормального распределения записывают не как функцию переменной x , а как функцию переменной $z = (x - \mu)/\sigma$, которая имеет нулевое математическое ожидание и дисперсию, равную 1. Плотность вероятности при этом равна:

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Такое распределение называют стандартным нормальным распределением.

Плотность вероятности распределения Гаусса нельзя проинтегрировать для получения интегральной функции распределения вероятностей $F(x)$ в явном виде. $F(x)$ можно найти с использованием:

- численных методов интегрирования функции $p(x)$,
- путем разложения функции $p(x)$ в ряд с последующим аналитическим интегрированием этого ряда.

Широкое применение распределения Гаусса в статистике основано на доказанном в теории вероятностей утверждении, что случайная величина, являющаяся суммой большого числа независимых случайных величин с конечными дисперсиями и с практически произвольными законами распределения, распределена нормально.

То есть условием использования нормального распределения для описания случайной величины являются ситуации, когда изучаемую случайную величину можно представить в виде суммы достаточно большого количества независимых слагаемых, каждое из которых мало влияет на сумму.

Распределение Гаусса можно использовать в качестве первого приближения для описания, например, *логарифмов относительного изменения цен активов*. Однако, только в качестве первого приближения, потому что на практике распределения этих величин отличаются от нормального, то есть имеют как правило более ярко выраженный пик и более "тяжелые" хвосты. Следовательно эти распределения являются островершинными и имеют эксцесс, превышающий три (иногда очень существенно).

Вычисление нормального распределения с помощью Microsoft Excel

Приведем несколько примеров вычисления характеристик нормального распределения. Все используемые функции можно найти в разделе "Статистические функции" электронных таблиц Microsoft Excel.

Пусть случайная величина X подчиняется нормальному распределению с параметрами (μ, σ) .

- 1) Плотность распределения в точке $X = x$:
 $\text{НОРМРАСП}(x, \mu, \sigma, \text{ложь})$
- 2) Вероятность того, что $X \leq x$:
 $\text{НОРМРАСП}(x, \mu, \sigma, \text{истина})$
- 3) Вероятность того, что $X > x$:
 $1 - \text{НОРМРАСП}(x, \mu, \sigma, \text{истина})$
- 4) Если известна вероятность того, что $X \leq x$, то есть $P = P\{X \leq x\}$, то соответствующее значение x можно вычислить как:
 $x = \text{НОРМОБР}(P, \mu, \sigma)$
- 5) Для приведения нормально распределенной случайной величины к стандартному виду, то есть для вычисления $z = (x - \mu) / \sigma$ используется функция:
 $z = \text{НОРМАЛИЗАЦИЯ}(x, \mu, \sigma)$

Пусть случайная величина Z подчиняется стандартному нормальному распределению ($\mu = 0, \sigma = 1$).

- 1) Вероятность того, что $Z \leq z$:
 $\text{НОРМСТРАСП}(z)$
- 2) Вероятность того, что $Z > z$:
 $1 - \text{НОРМСТРАСП}(z)$
- 3) Если известна вероятность того, что $Z \leq z$, то есть $P = P\{Z \leq z\}$, то соответствующее значение z можно вычислить как:
 $z = \text{НОРМСТОБР}(P)$
- 4) Вероятность того, что $-z \leq Z \leq z$:
 $\text{НОРМСТРАСП}(z) - \text{НОРМСТРАСП}(-z)$
- 5) Если известна вероятность того, что $-z \leq Z \leq z$, то есть $P = P\{-z \leq Z \leq z\}$, то соответствующее значение z можно вычислить как:
 $z = \text{НОРМСТОБР}((1 + P) / 2)$ или

$$z = -\text{НОРМСТОБР}((1 - P) / 2)$$

2.6. Логнормальное распределение.

Пусть x - нормально распределенная случайная величина с плотностью распределения:

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad -\infty < x < +\infty$$

Тогда случайная величина y , связанная с величиной x соотношением $y(x) = e^x$ будет распределена логнормально. Заметим, что y может принимать значения только от 0 до $+\infty$. Найдем основные параметры логнормального распределения.

Обозначим неизвестную пока плотность логнормального распределения через $p_y(y)$, которую определим исходя из равенства дифференциалов:

$$p_y(y)dy = p_x(x)dx \Rightarrow p_y(y) = p_x[x(y)] \cdot dx/dy$$

Так как $x(y) = \ln(y)$, и $dx/dy = 1/y$, для плотности вероятности логнормального распределения получается следующая формула:

$$p_y(y) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right) \quad 0 < y < +\infty$$

Параметры логнормального распределения выражаются через параметры соответствующего распределения Гаусса следующим образом:

$$\mu_y = \exp(\mu + \sigma^2 / 2)$$

$$D_y = \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1)$$

$$Me = \exp(\mu)$$

$$Mo = \exp(\mu - \sigma^2)$$

Распределение имеет крутой левый и пологий правый спад, то есть имеет положительную асимметрию.

Как и в случае распределения Гаусса, плотность вероятности логнормального распределения нельзя проинтегрировать

для получения функции распределения вероятностей в явном виде.

Однако, значения интегральной функции логнормального распределения можно найти, используя значения интегральной функции распределения Гаусса, так как они связаны соотношением $F_y(y) = F_x[\ln(y)]$, или в явном виде:

$$\int_0^y p_y(t) dt = \int_{-\infty}^{\ln(y)} p_x(t) dt$$

Логнормальное распределение можно использовать в качестве первого приближения для описания *относительного изменения цен активов*, однако, с теми ограничениями, о которых было сказано при обсуждении распределения Гаусса.

2.7. Распределение Лапласа.

Еще одним типом экспоненциального распределения, наряду с нормальным, является распределение Лапласа, плотность которого выражается формулой:

$$p(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{|x - \mu|}{\sigma/\sqrt{2}}\right) \quad -\infty < x < +\infty$$

Как и распределение Гаусса, распределение Лапласа:

- зависит от двух параметров (μ, σ) ,
- *математическое ожидание, медиана и мода* данного распределения равны μ , а *дисперсия* σ^2 ,
- кривая плотности вероятности симметрична относительно математического ожидания, *коэффициент асимметрии* равен нулю.

Однако *эксцесс* распределения $\varepsilon = 6$, то есть вдвое превышает эксцесс нормального распределения. Следовательно, распределение Лапласа островершинное, то есть имеет высокий пик и "тяжелые" хвосты.

Кроме того, плотность данного распределения интегрируема, и функция распределения может быть получена в явном виде:

$$F(x) = 0.5 \cdot \exp\left(-\frac{|x - \mu|}{\sigma/\sqrt{2}}\right) \quad x \leq \mu$$

$$F(x) = 1 - 0.5 \cdot \exp\left(-\frac{|x - \mu|}{\sigma/\sqrt{2}}\right) \quad x > \mu$$

Распределение Лапласа можно использовать для описания *логарифмов относительного изменения цен активов*, зачастую с большим успехом, чем нормальное распределение. Однако, с еще большей точностью, реальные распределения вероятностей описывает *обобщенное экспоненциальное распределение*, которое будет также рассмотрено в этой главе.

2.8. Распределение Коши.

Распределение Коши является одним из простейших законов распределения. Его плотность выражается формулой:

$$p(x) = \frac{1}{b\pi(1 + [(x - a)/b]^2)}$$

Плотность распределения Коши имеет вид симметричной относительно точки $x = a$ кривой, визуально очень похожей на плотность нормального распределения.

Кроме того $p(x)$ интегрируема, поэтому функцию распределения Коши можно записать в явном виде и не прибегать при ее вычислении к помощи численных методов:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctg[(x - a)/b]$$

Казалось бы, распределение Коши выглядит очень привлекательно для описания и моделирования случайных величин. Однако в действительности это не так. Свойства распределения Коши резко отличны от свойств распределения Гаусса, Лапласа и других экспоненциальных распределений.

Дело в том, что распределение Коши близко к предельно пологому. Напомним, что распределение называется предельно пологим, если при $x \rightarrow \pm\infty$ его плотность вероятности $p(x) \cong 1/|x|^{1+\delta}$, где δ - сколь угодно малое положительное число. При более пологих, чем $1/|x|^{1+\delta}$ спадах, площадь под

кривой бесконечна, то есть не выполняется условие нормирования, и такие кривые не могут описывать плотность распределения вероятностей.

Для распределения Коши не существует даже первого начального момента распределения, то есть *математического ожидания*, так как определяющий его интеграл расходится. При этом распределение имеет и *медиану* и *моду*, которые равны параметру a .

Разумеется, *дисперсия* этого распределения (второй центральный момент) также равна бесконечности. На практике это означает, что оценка дисперсии по выборке из распределения Коши будет неограниченно возрастать с увеличением объема данных.

Из вышесказанного следует, что аппроксимация распределением Коши случайных процессов, которые характеризуются конечным математическим ожиданием и конечной дисперсией, неправомерна.

2.9. Распределение Парето.

Распределение Парето - это усеченное слева распределение, плотность вероятности и функция распределения которого выражаются в виде:

$$x < B: \quad p(x) = F(x) = 0$$

$$x \geq B > 0, \alpha > 0: \quad p(x) = \frac{\alpha}{B} (B/x)^{1+\alpha} \quad F(x) = 1 - (B/x)^\alpha$$

Плотность $p(x)$ распределения равна нулю при $x < B$, имеет максимальное значение при $x = B$ и монотонно убывает при $x > B$.

Распределение Парето можно модифицировать таким образом, чтобы его можно было использовать для описания симметричных распределений вероятностей.

Введя новую переменную $t = x - B$, получим

$$p(t) = \frac{\alpha}{B} \cdot [B/(B+t)]^{1+\alpha} = \frac{\alpha}{B} \cdot [1/(1+t/B)]^{1+\alpha}$$

$$0 \leq t < +\infty, \quad B > 0, \quad \alpha > 0$$

Взяв величину t по модулю, эту формулу можно распространить на всю числовую ось, введя при этом нормировочный коэффициент $1/2$.

$$p(t) = \frac{\alpha}{2B} \cdot [1/(1 + |t|/B)]^{1+\alpha}$$

$$-\infty < t < +\infty, \quad B > 0, \quad \alpha > 0$$

Описываемое последней формулой распределение имеет центр в точке $t = 0$. Для распределения с центром в произвольной точке $t = A$ получим

$$p(t) = \frac{\alpha}{2B} \cdot [1/(1 + |t - A|/B)]^{1+\alpha}$$

$$-\infty < t < +\infty, \quad B > 0, \quad \alpha > 0$$

Итак, мы получили симметричное распределение, зависящее от трех параметров, с помощью которого можно описывать выборки случайных величин, в том числе с пологими спадами. Однако, это распределение обладает недостатками, которые были рассмотрены при обсуждении распределения Коши, а именно, математическое ожидание существует только при $\alpha > 1$, дисперсия конечна только при $\alpha > 2$, и вообще, конечный момент распределения k -го порядка существует при $\alpha > k$.

2.10. Обобщенное экспоненциальное распределение.

Выше в этой главе были рассмотрены два вида экспоненциальных распределений: Гаусса и Лапласа. У них много общего: они симметричны, зависят от двух параметров (μ, σ) , имеют конечные моменты любого порядка. Отличие же состоит в том, что из-за того, что переменная x возводится в разную степень под знаком экспоненты (в квадрат у распределения Гаусса и в первую степень у распределения Лапласа), эксцесс у них разный. Напомним, что эксцесс характеризует остроту пика распределения и крутизну спада хвостов распределения. Возникает вопрос: можно ли записать формулу для плотности вероятности экспоненциального распределения в общем виде, то есть с произвольной положительной степенью переменной x под знаком экспоненты? Оказывается, такая формула существует:

$$p(x) = \frac{\alpha}{2\Gamma(1/\alpha)\lambda\sigma} \exp\left(-\left|\frac{x - \mu}{\lambda\sigma}\right|^\alpha\right) \quad -\infty < x < +\infty$$

$$\lambda = \sqrt{\Gamma(1/\alpha)/\Gamma(3/\alpha)}$$

где $\Gamma(t)$ - это гамма-функция. О том, как вычисляется гамма-функция, рассказано в ПРИЛОЖЕНИИ 2.1 к этой главе.

Распределение с приведенной выше плотностью вероятности мы будем называть *обобщенным экспоненциальным распределением*, которое характеризуется тремя параметрами:

- математическим ожиданием (медианой, модой) μ ,
- среднеквадратичным отклонением σ (дисперсией σ^2),
- показателем степени распределения α .

Показатель степени α характеризует форму распределения:

- при $\alpha < 1$ распределение имеет очень острый пик и очень пологие спады,
- при $\alpha = 1$ распределение тождественно распределению Лапласа,
- при $\alpha = 2$ распределение тождественно распределению Гаусса,
- при $\alpha > 2$ распределение становится похожим на равнобедренную трапецию, то есть имеет плоскую вершину и резко спадающие хвосты,
- при $\alpha \rightarrow \infty$ распределение тождественно равномерному распределению.

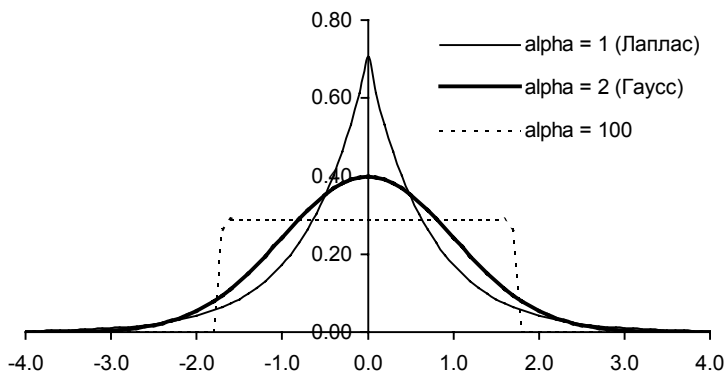
Экссесс распределения однозначно определяется показателем степени α : $\varepsilon = \Gamma(1/\alpha)\Gamma(5/\alpha)/[\Gamma(3/\alpha)]^2$.

Соответственно, из вычисленного по выборке случайных величин значения оценки эксцесса, можно определить оценку показателя α .

Обычно в справочниках распределения Гаусса, Лапласа и равномерное рассматриваются как разные распределения, хотя в излагаемой здесь концепции - это одно и тоже распределение. Единственным параметром, характеризующим форму (а значит и свойства) этих распределений является показатель α .

В дальнейшем, если принята гипотеза о том, что плотность вероятности случайной величины имеет экспоненциальный характер, для описания этой величины будем использовать именно обобщенное экспоненциальное распределение.

На рисунке приведена плотность стандартного обобщенного экспоненциального распределения при различных значениях α .



В общем случае плотность вероятности обобщенного экспоненциального распределения нельзя проинтегрировать для получения функции распределения вероятностей $F(x)$ в явном виде. $F(x)$ можно найти с использованием:

- численных методов интегрирования функции $p(x)$,
- путем разложения функции $p(x)$ в ряд с последующим аналитическим интегрированием этого ряда.

О том, как это делается, будет рассказано в следующих параграфах.

Вычисление интегральной функции обобщенного экспоненциального распределения в Microsoft Excel

Не приводя доказательства скажем, что интегральная функция обобщенного экспоненциального распределения может быть выражена через интегральную функцию гамма-распределения (встроенная функция Excel) следующим образом:

$$x \geq \mu : F(x) = 0.5 + 0.5 \cdot \text{ГАММАРАСП} \left(\left| \frac{x - \mu}{\lambda \sigma} \right|^\alpha, \frac{1}{\alpha}, 1, \text{ИСТИНА} \right)$$

$$x < \mu : F(x) = 0.5 - 0.5 \cdot \text{ГАММАРАСП} \left(\left| \frac{x - \mu}{\lambda \sigma} \right|^\alpha, \frac{1}{\alpha}, 1, \text{ИСТИНА} \right)$$

2.11. Поиск интегральной функции распределения путем численного интегрирования плотности распределения.

Пусть дана случайная величина, которая подчиняется обобщенному экспоненциальному распределению с параметрами

(μ, σ, α) , то есть плотность вероятности имеет вид:

$$p(x) = \frac{\alpha}{2\Gamma(1/\alpha)\lambda\sigma} \exp\left(-\left|\frac{x-\mu}{\lambda\sigma}\right|^\alpha\right) \quad -\infty < x < +\infty$$

Требуется найти интегральную функцию распределения:

$$F(x) = \int_{-\infty}^x p(z) dz$$

Прежде всего заметим, что задачу можно упростить, перейдя к переменной $t = (x - \mu) / \sigma$. Случайная величина t имеет математическое ожидание, равное нулю, и дисперсию, равную единице. Формулы для плотности вероятности и функции распределения примут вид:

$$p(t) = \frac{\alpha}{2\Gamma(1/\alpha)\lambda} \exp\left(-|t/\lambda|^\alpha\right) \quad -\infty < t < +\infty$$

$$F(t) = \int_{-\infty}^t p(z) dz$$

Для решения поставленной задачи достаточно:

- разбить всю область определения переменной t на N интервалов, при этом узлы разбиения будут образовывать массив $\{T_k\}, k = 0, \dots, N$.
- вычислить в узлах разбиения массив значений интегральной функции $\{F_k\}, k = 0, \dots, N$.

Тогда для произвольного значения переменной t , такого, что $T_k \leq t \leq T_{k+1}$ значение интегральной функции $F(t)$ может быть приближенно определено в виде:

$$F(t) = F_k + (t - T_k)(F_{k+1} - F_k) / (T_{k+1} - T_k)$$

Так как между узлами разбиения величина F аппроксимируется линейной зависимостью, то интервал разбиения должен быть достаточно мал, то есть количество узлов разбиения достаточно велико.

Однако, плотность вероятности определена на всей числовой оси, а разбить бесконечный интервал на конечное количество интервалов конечной длины невозможно. Поэтому приближенно будем считать, что

$$p(t) = \frac{\alpha}{2\Gamma(1/\alpha)\lambda} \exp(-|t/\lambda|^\alpha) \quad -R \leq t \leq R$$

$$p(t) = 0 \quad |t| > R$$

$$F(t) = \int_{-R}^t p(z) dz$$

где величину $2R$ назовем размахом распределения.

Очевидно, интервал от $-R$ до R можно трактовать как интерквантильный промежуток с некоторой близкой к единице доверительной вероятностью. Для обобщенного экспоненциального распределения половину размаха можно задать эмпирически полученной формулой:

$$R = \sqrt{3} + 4.788 \cdot (\varepsilon - 1.8)^{2/3}$$

После введения понятия размаха распределения, все готово для написания алгоритма решения задачи:

1) Задаем входные данные: показатель степени распределения α и количество интервалов разбиения N (целое четное число).

2) Вычисляем эксцесс распределения

$$\varepsilon = \Gamma(1/\alpha)\Gamma(5/\alpha)/[\Gamma(3/\alpha)]^2$$

3) Вычисляем половину размаха распределения

$$R = \sqrt{3} + 4.788 \cdot (\varepsilon - 1.8)^{2/3}$$

4) Вычисляем минимальное и максимальное значение переменной t

$$T_{\min} = -R$$

$$T_{\max} = R$$

5) Вычисляем массив узлов на оси t

$$T_k = T_{\min} + (T_{\max} - T_{\min})k/N \quad k = 0, \dots, N$$

6) Вычисляем номер центра распределения

$$M = N/2$$

7) Вычисляем массив значений плотности вероятности для k от 0 до M

$$p_k = \frac{\alpha}{2\Gamma(1/\alpha)\lambda} \exp(-|T_k/\lambda|^\alpha) \quad k = 1, \dots, M$$

$$p_0 = 0$$

- 8) Вычисляем вспомогательный массив, в котором величины p_k суммируются нарастающим итогом для k от 0 до M

$$S_k = \sum_{i=0}^k p_i \quad k = 0, \dots, M$$

- 9) Так как значение интегральной функции в центре распределения (то есть в узле с номером M) равно 0.5, то можно вычислить левую часть массива, в котором содержится функция распределения

$$F_k = 0.5 \cdot S_k / S_M \quad k = 0, \dots, M$$

- 10) Так как распределение симметрично относительно центра, вычисляем оставшуюся часть массива, в котором содержится функция распределения

$$F_k = 1 - F_{N-k} \quad k = M+1, \dots, N$$

Итак, мы получили массив значений случайной величины $\{T_k\}$ и соответствующий ему массив интегральной функции распределения $\{F_k\}$, то есть задали функцию распределения в табличном виде.

Для произвольного значения переменной t значение интегральной функции $F(t)$ может быть определено по формулам:

$$t < T_0 : F(t) = 0$$

$$T_k \leq t < T_{k+1} \quad k = 0, \dots, N-1 :$$

$$F(t) = F_k + (t - T_k)(F_{k+1} - F_k) / (T_{k+1} - T_k)$$

$$t \geq T_N : F(t) = 1$$

Напомним, что переменная t имеет нулевое математическое ожидание и единичную дисперсию. Переменная x с произвольными математическим ожиданием и дисперсией связана с переменной t соотношением $x = \mu + \sigma t$.

2.12. Поиск интегральной функции распределения путем разложения плотности распределения в ряд с последующим аналитическим интегрированием этого ряда.

Задачу, поставленную в предыдущем параграфе, можно решить путем разложения стоящей под знаком интеграла функции в ряд Тейлора с последующим интегрированием этого

ряда. Итак, нам нужно проинтегрировать обобщенное экспоненциальное распределение ($\mu = 0, \sigma = 1$):

$$p(t) = \frac{\alpha}{2\Gamma(1/\alpha)\lambda} \exp(-|t/\lambda|^\alpha) \quad -\infty < t < +\infty$$

$$F(t) = \int_{-\infty}^t p(z) dz$$

Так как плотность распределения симметрична относительно центра, для нахождения интегральной функции распределения достаточно вычислить

$$\begin{aligned} q(t) &= \int_0^t p(z) dz = \frac{\alpha}{2\Gamma(1/\alpha)\lambda} \int_0^t \exp(-(z/\lambda)^\alpha) dz = \\ &= \frac{\alpha}{2\Gamma(1/\alpha)} \int_0^t \exp(-(z/\lambda)^\alpha) d(z/\lambda) = \frac{\alpha}{2\Gamma(1/\alpha)} \int_0^{t/\lambda} \exp(-y^\alpha) dy \end{aligned}$$

Функцию под знаком интеграла можно разложить в ряд Тейлора следующим образом:

$$\exp(-y^\alpha) = \sum_{k=0}^{\infty} (-1)^k \frac{y^{\alpha k}}{k!}$$

Подставив это разложение под знак интеграла и проведя интегрирование, получим

$$q(t) = \frac{\alpha}{2\Gamma(1/\alpha)} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \cdot \frac{(t/\lambda)^{\alpha k + 1}}{\alpha k + 1}$$

Практически, суммирование производится не до ∞ , а до некоторого $k=N$, такого, что

$$\frac{1}{N!} \cdot \frac{(t/\lambda)^{\alpha N + 1}}{\alpha N + 1} \leq \delta$$

где δ - это некоторое наперед заданное малое положительное число (точность вычисления). То есть мы должны вычислить частичную сумму ряда.

Функция распределения $F(t)$ связана с $q(t)$ соотношениями:

$$t \geq 0: \quad F(t) = 0.5 + q(t)$$

$$t < 0: \quad F(t) = 0.5 - q(|t|)$$

2.13. Моделирование с помощью равномерного распределения случайных чисел с произвольной плотностью распределения.

Встроенный в компьютер генератор псевдослучайных чисел выдает числа, равномерно распределенные в интервале от 0 до 1. Так как любая интегральная функция распределения $F(x)$ имеет область значений от 0 до 1, то с помощью равномерного распределения можно получить случайное число с произвольным законом распределения путем решения обратной задачи, то есть восстанавливая по известному значению $F(x)$ значение x . В качестве примера будем моделировать случайную величину, подчиняющуюся обобщенному экспоненциальному распределению. Для решения этой задачи будем использовать результаты, полученные в двух предыдущих параграфах.

Постановка задачи

Дано случайное число z , равномерно распределенное на интервале от 0 до 1.

Требуется получить число x , подчиняющееся обобщенному экспоненциальному распределению, с параметрами (μ, σ, α) .

Решение путем предварительного численного интегрирования плотности распределения, то есть задания функции распределения в табличном виде.

Для получения искомого числа x , найдем сначала вспомогательное число t , которое подчиняется обобщенному экспоненциальному распределению, с параметрами $(\mu = 0, \sigma = 1)$.

Для этого по методике, изложенной в параграфе 2.11, получим массив значений случайной величины $\{T_k\}$ и соответствующий ему массив интегральной функции распределения $\{F_k\}$, то есть зададим функцию распределения в табличном виде. Для произвольного значения переменной t значение интегральной функции $F(t)$ может быть определено как:

$$t < T_0 : F(t) = 0$$

$$T_k \leq t < T_{k+1} \quad k = 0, \dots, N-1 :$$

$$F(t) = F_k + (t - T_k)(F_{k+1} - F_k) / (T_{k+1} - T_k)$$

$$t \geq T_N : F(t) = 1$$

На интервале $T_0 \leq t \leq T_N$ величины t и F связаны линейно. Если принять, что величина t не может быть меньше T_0 и не может быть больше T_N , то можно получить обратную зависимость $t(F)$ в виде:

$$F = 0: \quad t(F) = T_0$$

$$F = 1: \quad t(F) = T_N$$

$$F_k \leq F < F_{k+1} \quad k = 0, \dots, N-1:$$

$$t(F) = T_k + (F - F_k)(T_{k+1} - T_k)/(F_{k+1} - F_k)$$

Если считать, что полученная с помощью генератора случайных чисел величина z является значением функции распределения F в некоторой точке t , то величина t может быть найдена по приведенным выше формулам, где вместо переменной F подставлена величина z . Искомое число x вычисляется как $x = \mu + \sigma t$.

Решение методом итераций, с использованием вычисления функции распределения через ряд Тейлора.

Как и в предыдущем случае, для получения искомого числа x , найдем сначала вспомогательное число t , которое подчиняется обобщенному экспоненциальному распределению, с параметрами $(\mu = 0, \sigma = 1)$.

В параграфе 2.12 было показано, что можно вычислить функцию распределения $F(t)$ в точке t с требуемой точностью как частичную сумму соответствующего ряда. Теперь нам нужно решить обратную задачу, то есть по известному значению $F(t)$ найти неизвестное значение t . Точнее, в соответствии с условиями поставленной задачи, мы должны решить относительно t уравнение $F(t) - z = 0$.

Для численного решения этого уравнения мы используем *метод деления пополам*. Для того, чтобы приступить к решению этим методом, необходимо задать конечный интервал, в котором должен лежать корень уравнения. В качестве области возможных значений t выберем интервал от $-R$ до R , где $2R$ - это рассмотренный выше размах распределения. Считаем, что $F(-R) = 0, F(R) = 1$.

Численное решение уравнения $F(t) - z = 0$ с заданной точностью означает, что достаточно найти такое t , при котором

$|F(t) - z| \leq \delta$, где δ - это некоторое наперед заданное малое положительное число (точность вычисления).

Решение состоит в последовательном повторении шагов (итераций), до тех пор, пока не будет достигнута необходимая точность:

- 1) Задаем начальные величины граничных значений переменных t и F

$$T_{\min} = -R \quad F_{\min} = 0$$

$$T_{\max} = R \quad F_{\max} = 1$$

- 2) Вычисляем текущее значение t как среднее значение между T_{\min} и T_{\max}

$$t = (T_{\min} + T_{\max}) / 2$$

- 3) Вычисляем по методике из параграфа 2.12 величину $F(t)$

- 4) Проверяем условие $|F - z| \leq \delta$. Если неравенство справедливо, то необходимая точность решения достигнута и текущее значение t является решением.

- 5) В случае $|F - z| > \delta$ изменяем значения величин T_{\min} , T_{\max} , F_{\min} , F_{\max} :

- если $F < z$, то

$$T_{\min} = t, F_{\min} = F$$

T_{\max} , F_{\max} остаются без изменения

- если $F > z$, то

T_{\min} , F_{\min} остаются без изменения

$$T_{\max} = t, F_{\max} = F$$

- 6) Возвращаемся на шаг 2.

После того, как необходимая точность вычисления величины t достигнута, искомое число x находится как $x = \mu + \sigma t$. На практике чаще всего необходимо получить не отдельное случайное число x с заданным законом распределения, а последовательность таких чисел $\{x_k\}$, $k = 0, \dots, N$. Это необходимо, как правило, при моделировании случайных процессов. В этом случае описанные в данном параграфе процедуры нужно повторить соответствующее количество раз.

ПРИЛОЖЕНИЕ 2.1. Гамма-функция Эйлера.

Гамма-функция, обобщающая понятие факториала, является одной из важнейших специальных функций. Для произвольного положительного x , значение $\Gamma(x)$ задается формулой:

$$\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt \quad x > 0$$

В этом приложении мы рассмотрим алгоритм вычисления гамма-функции. Данный алгоритм основан на следующем ее свойстве: $\Gamma(x+1) = x \cdot \Gamma(x)$ для любого $x > 0$. Это свойство позволяет свести вычисление $\Gamma(x)$ от любого x к вычислению гамма-функции на интервале $1 \leq x \leq 2$, на котором ее можно аппроксимировать полиномом пятой степени:

$$\Gamma(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$$

$$a_0 = 3.69764701987851$$

$$a_1 = -6.60156185480728$$

$$a_2 = 6.37763107208549$$

$$a_3 = -3.26329362313704$$

$$a_4 = 0.885309940132118$$

$$a_5 = -0.0957403771932692$$

Область значений величины $x > 0$ можно разбить на три интервала, на каждом из которых $\Gamma(x)$ вычисляется следующим образом:

1) $1 \leq x \leq 2$

В этом случае, $\Gamma(x)$ непосредственно вычисляется с помощью приведенного выше полинома.

2) $0 < x < 1$

В этом случае, $\Gamma(x) = \Gamma(x+1)/x$, и так как $1 < x+1 < 2$, то $\Gamma(x+1)$ вычисляется с помощью полинома.

3) $x > 2$

В этом случае величину x можно представить в виде $x = N + z$, где

N - это целая часть x ($N \geq 2$),

z - это дробная часть x ($0 < z < 1$).

Тогда

$$\begin{aligned}\Gamma(x) &\equiv \Gamma(N+z) = (N+z-1)\Gamma(N+z-1) = \dots = \\ &= \Gamma(1+z) \prod_{k=1}^{N-1} (N+z-k)\end{aligned}$$

и так как $1 < z+1 < 2$, то $\Gamma(z+1)$ вычисляется с помощью полинома.

Вычисление гамма-функции с помощью Microsoft Excel

В Microsoft Excel гамма-функцию можно вычислить, используя следующую комбинацию функций:

$$\Gamma(x) = EXP(ГАММАНЛОГ(x))$$

3. СПЕЦИАЛЬНЫЕ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

3.1. t-распределение Стьюдента.

Плотность распределения Стьюдента описывается формулой:

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + x^2/2\right)^{-(\nu+1)/2} \quad -\infty < x < +\infty$$

Распределение имеет вид колоколообразной кривой, симметричной относительно точки $t = 0$, и зависит от единственного параметра ν , который принято называть числом степеней свободы. Приведем значения основных характеристик распределения Стьюдента:

Математическое ожидание, 0 при $\nu > 1$
медиана, мода

Дисперсия $\frac{\nu}{\nu-2}$ при $\nu > 2$

Коэффициент асимметрии 0

Эксцесс $\frac{3(\nu-2)}{(\nu-4)}$ при $\nu > 4$

При числе степеней свободы $\nu \rightarrow \infty$, распределение Стьюдента стремится к стандартному нормальному распределению, то есть к нормальному распределению с центром 0 и дисперсией 1.

Типичная интерпретация

- 1) Пусть случайная величина X имеет нормальное распределение с математическим ожиданием μ и дисперсией σ^2 .

Если имеется выборка этой случайной величины (x_1, x_2, \dots, x_N) , то состоятельными и несмещенными оценками математического ожидания и дисперсии по выборке будут следующие величины:

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \bar{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2$$

Тогда случайные величины $t = \frac{x_k - \mu}{\sigma}$ и $t = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$ будут

подчиняться распределению Стьюдента с $\nu = N - 1$ степенями свободы.

- 2) Пусть случайные величины X и Y имеют нормальное распределение с математическими ожиданиями и дисперсиями (μ_x, σ_x^2) и (μ_y, σ_y^2) соответственно.

Если имеются выборки этих случайных величин (x_1, x_2, \dots, x_N) и (y_1, y_2, \dots, y_N) , то состоятельной и несмещенной оценкой коэффициента корреляции между этими величинами по выборке будет:

$$\bar{\rho} = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sqrt{\sum_{k=1}^N (x_k - \bar{X})^2} \sqrt{\sum_{k=1}^N (y_k - \bar{Y})^2}}$$

Тогда случайная величина $t = \sqrt{N-2} \cdot \frac{\bar{\rho}}{\sqrt{1-\bar{\rho}^2}}$ будет

подчиняться распределению Стьюдента с $\nu = N - 2$ степенями свободы.

Вычисление распределения Стьюдента с помощью Microsoft Excel

Приведем несколько примеров вычисления характеристик распределения Стьюдента. Все используемые функции можно найти в разделе "Статистические функции" электронных таблиц Microsoft Excel.

Пусть случайная величина X подчиняется распределению Стьюдента с числом степеней свободы ν .

- 1) Вероятность того, что $X \leq x$:

$$1 - \text{СТЮДРАСП}(x, \nu, 1)$$

- 2) Вероятность того, что $X > x$:

$$\text{СТЮДРАСП}(x, \nu, 1)$$

- 3) Вероятность того, что $-x \leq X \leq x$, вычисляется как:

$$P = 1 - \text{СТЮДРАСП}(x, \nu, 2)$$

- 4) Вероятность того, что $|X| > x$, равна:

$$q = \text{СТЮДРАСП}(x, \nu, 2)$$

Величина q - это вероятность того, что случайная величина X попадает в критическую область распределения Стьюдента.

- 5) Если известна вероятность q того, что $|X| > x$, то соответствующее значение x равно:

$$x = \text{СТЮДРАСПОБР}(q, \nu)$$

- 6) Если известна вероятность q того, что $X > x$, то соответствующее значение x равно:

$$x = \text{СТЮДРАСПОБР}(2q, \nu)$$

3.2. χ^2 -распределение.

Плотность χ^2 -распределения задается формулой:

$$x < 0: \quad p(x) = 0$$

$$x > 0: \quad p(x) = \frac{x^{(\nu-2)/2} \cdot \exp(-x/2)}{\Gamma(\nu/2) \cdot 2^{\nu/2}}$$

Плотность зависит от единственного параметра ν , который принято называть числом степеней свободы. Приведем значения основных характеристик распределения:

Математическое ожидание	ν
Мода	$\nu - 2 \quad (\nu \geq 2)$
Дисперсия	2ν
Коэффициент асимметрии	$2\sqrt{2/\nu}$
Эксцесс	$\frac{3\nu + 12}{\nu}$

При числе степеней свободы $\nu \rightarrow \infty$, χ^2 -распределение стремится к нормальному распределению с центром ν и дисперсией 2ν .

Типичная интерпретация

Пусть случайная величина X имеет нормальное распределение с математическим ожиданием μ и дисперсией σ^2 .

Если имеется выборка этой случайной величины (x_1, x_2, \dots, x_N) , то состоятельными и несмещенными оценками математического ожидания и дисперсии по выборке будут следующие величины:

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \bar{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2$$

Тогда случайная величина $\chi^2 = \sum_{k=1}^N ((x_k - \mu) / \sigma)^2$ будет подчиняться χ^2 -распределению с $\nu = N$ степенями свободы, а случайная величина $\chi^2 = (N-1) \cdot (\bar{\sigma}^2 / \sigma^2)$ будет подчиняться χ^2 -распределению с $\nu = N-1$ степенями свободы.

Вычисление χ^2 -распределения с помощью Microsoft Excel

Приведем несколько примеров вычисления характеристик χ^2 -распределения. Все используемые функции можно найти в разделе "Статистические функции" электронных таблиц Microsoft Excel.

Пусть случайная величина X подчиняется χ^2 -распределению с числом степеней свободы ν .

- 1) Вероятность того, что $X \leq x$, вычисляется как:

$$P = 1 - \text{ХИ2РАСП}(x, \nu)$$

- 2) Вероятность того, что $X > x$, равна:

$$q = \text{ХИ2РАСП}(x, \nu)$$

Величина q - это вероятность того, что случайная величина X попадает в критическую область χ^2 -распределения.

- 3) Если известна вероятность P или вероятность q , то соответствующее значение x , определяющее границу интервала $X \leq x$ равно:

$$x = \text{ХИ2ОБР}(q, \nu) \text{ или } x = \text{ХИ2ОБР}(1 - P, \nu)$$

3.3. F-распределение (распределение ν^2).

Плотность F -распределения задается формулой:

$$x < 0: \quad p(x) = 0$$

$$x > 0:$$

$$p(x) = (\nu_1 / \nu_2)^{\nu_1/2} \cdot \frac{\Gamma((\nu_1 + \nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \cdot x^{(\nu_1-2)/2} \cdot (1 + (\nu_1 / \nu_2) \cdot x)^{-(\nu_1 + \nu_2)/2}$$

Плотность F -распределения зависит от двух параметров (ν_1, ν_2) , которые принято называть числом степеней свободы. Приведем значения основных характеристик F -распределения:

Математическое
ожидание

$$\frac{\nu_2}{\nu_2 - 2} \quad \text{при } \nu_2 > 2$$

Мода

$$\frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)} \quad \text{при } \nu_1 > 2$$

Дисперсия

$$\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad \text{при } \nu_2 > 4$$

Типичная интерпретация

Пусть случайные величины X и Y имеют нормальное распределение с дисперсиями σ_x^2 и σ_y^2 соответственно.

Если имеются выборки этих случайных величин (x_1, x_2, \dots, x_N) и (y_1, y_2, \dots, y_M) , то состоятельными и несмещенными оценками дисперсий по выборке будут следующие величины:

$$\overline{\sigma_x^2} = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2 \quad \overline{\sigma_y^2} = \frac{1}{M-1} \sum_{k=1}^N (y_k - \bar{Y})^2$$

Пусть выборочная дисперсия величины X больше выборочной дисперсии величины Y . Тогда случайная величина $F = \overline{\sigma_x^2} / \overline{\sigma_y^2}$ будет подчиняться F -распределению с $\nu_1 = N - 1$, $\nu_2 = M - 1$ степенями свободы.

Вычисление F -распределения с помощью Microsoft Excel

Приведем несколько примеров вычисления характеристик F -распределения. Все используемые функции можно найти в разделе "Статистические функции" электронных таблиц Microsoft Excel.

Пусть случайная величина X подчиняется F -распределению с числом степеней свободы ν_1, ν_2 .

- 1) Вероятность того, что $X \leq x$, вычисляется как:

$$P = 1 - FPACП(x, \nu_1, \nu_2)$$

- 2) Вероятность того, что $X > x$, равна:

$$q = FPACП(x, \nu_1, \nu_2)$$

Величина q - это вероятность того, что случайная величина X попадает в критическую область F -распределения.

- 3) Если известна вероятность P или вероятность q , то соответствующее значение x , определяющее границу интервала $X \leq x$ равно:

$$x = FPACПБР(q, \nu_1, \nu_2) \text{ или}$$

$$x = FPACПБР(1 - P, \nu_1, \nu_2)$$

4. ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ ПО ВЫБОРКЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

4.1. Введение.

Эта глава посвящена методам оценки по эмпирической выборке параметров распределения случайной величины. Будут указаны формулы для оценки центра распределения, дисперсии и показателей формы распределения, а также практические приемы удаления аномальных значений (промахов) из выборки.

4.2. Оценки центра распределения.

По возможности наиболее точная оценка центра распределения по выборке случайных величин исключительно важна, так как центр распределения используется в формулах для вычисления дисперсии, среднеквадратичного отклонения, коэффициента асимметрии и эксцесса распределения. Некорректное определение центра влечет за собой ошибки в определении всех этих величин.

Оценку центра распределения по выборке можно проводить различными способами. Не зная априорно закона распределения случайной величины, невозможно заранее указать наиболее приемлемый способ. К тому же, некоторые из этих оценок чувствительны к наличию *аномальных значений в выборке (промахов)*.

Поэтому для корректной оценки центра распределения мы будем вычислять его пятью различными способами. После этого пять полученных оценок упорядочим по возрастанию и выберем из них в качестве центра распределения срединное, то есть третье по счету, значение.

Выборку случайных величин будем обозначать как $\{x_k\}$, $k = 1, \dots, N$. Упомянутые выше пять оценок центра по выборке следующие:

- медиана $X_{\text{медиана}}$,
- центр 50%-ного интерквантильного промежутка (центр сгибов) $X_{\text{центр_сгибов}}$,
- среднее арифметическое по всей выборке \bar{X} ,

- среднее арифметическое по 50%-ному интерквантильному промежутку $\overline{X}_{50\%}$,
- центр размаха $X_{\text{центр_размаха}}$.

Серединное значение этих оценок будем обозначать как $X_{\text{ЦЕНТР}}$.

Медиана

Перед вычислением медианы выборка $\{x_k\}$ должна быть упорядочена по возрастанию, после чего медиану можно определить следующим образом:

- если объем выборки N является нечетным, то

$$X_{\text{медиана}} = x_{(N+1)/2}$$

- если объем выборки N является четным, то

$$X_{\text{медиана}} = (x_{N/2} + x_{(N/2)+1})/2$$

Медиана нечувствительна к промахам в выборке.

Центр 50%-ного интерквантильного промежутка (центр сгибов)

Перед вычислением этой оценки выборка $\{x_k\}$ также должна быть упорядочена по возрастанию. Обозначим как M четвертую часть от объема выборки, то есть $M = \text{ЦЕЛОЕ}(N/4)$.

Тогда центр сгибов определяется по формуле:

$$X_{\text{центр сгибов}} = (x_{M+1} + x_{N-M})/2$$

Центр сгибов нечувствителен к промахам в выборке.

Среднее арифметическое по всей выборке

Среднее арифметическое (выборочная средняя) является самым распространенным методом оценки центра распределения:

$$\overline{X} = \frac{1}{N} \sum_{k=1}^N x_k$$

Эта величина является несмещенной и состоятельной оценкой математического ожидания (генеральной средней) μ случайной переменной x . Несмещенность заключается в том, что математическое ожидание величины \overline{X} равно μ . Состоятель

ность заключается в том, что при объеме выборки $N \rightarrow \infty$, значение величины $\bar{X} \rightarrow \mu$.

Среднее арифметическое случайных величин само является случайной величиной. Дисперсия и среднеквадратичное отклонение среднего арифметического зависят от дисперсии и среднеквадратичного отклонения самой случайной величины и объема выборки:

$$D(\bar{X}) = D / N = \sigma^2 / N$$

$$\sigma(\bar{X}) = \sigma / \sqrt{N}$$

Это соотношение справедливо для *независимых* данных с конечной дисперсией и с любым законом распределения. Таким образом, с.к.о. среднего значения меньше, чем с.к.о. самой случайной величины в \sqrt{N} раз. Из этого следует, что точность оценки можно повысить путем увеличения объема выборки. Среднее арифметическое не защищено от промахов. Особенно большое влияние на него оказывают промахи при малом объеме выборки. При увеличении объема эта оценка становится все более устойчивой.

Среднее арифметическое по 50%-му интерквантильному промежутку

Перед вычислением этой оценки выборка $\{x_k\}$ должна быть упорядочена по возрастанию. Данная оценка является аналогом предыдущей, но усреднение проводится по усеченной на 25% слева и справа выборке. Если обозначить как M четвертую часть от объема выборки, то есть $M = \text{ЦЕЛОЕ}(N/4)$, то

$$\overline{X}_{50\%} = \frac{1}{N - 2M} \sum_{k=M+1}^{N-M} x_k$$

Среднее арифметическое по 50%-ному интерквантильному промежутку нечувствительно к промахам в выборке.

Центр размаха

Центр размаха определяется как среднее между максимальным и минимальным значением в выборке:

$$X_{\text{центр размаха}} = [\max(x_k) + \min(x_k)] / 2$$

Центр размаха не защищен от промахов в выборке. Более того, в отличие от среднего арифметического, объем выборки оказывает гораздо меньшее влияние на точность этой оценки.

4.3. Оценка дисперсии и среднеквадратичного отклонения.

Оценки дисперсии и среднеквадратичного отклонения по выборке случайной величины $\{x_k\}$, $k=1, \dots, N$ вычисляются по формулам:

$$\overline{D} = \frac{1}{N-1} \sum_{k=1}^N (x_k - \overline{X})^2$$

$$\overline{\sigma} = \sqrt{\overline{D}}$$

В случае небольших выборок и при наличии промахов вместо среднего арифметического \overline{X} следует применять $X_{\text{ЦЕНТР}}$.

Эти оценки называют еще выборочной дисперсией и выборочным с.к.о. Они определяют рассеяние случайной величины, однако сами также являются случайными величинами со своими показателями рассеяния.

Приближенные формулы для вычисления дисперсии и с.к.о. выборочной дисперсии, а также дисперсии и с.к.о. выборочного с.к.о. следующие:

$$D(\overline{D}) \approx \frac{\overline{m_4} - \overline{\sigma}^4}{N} \quad \sigma(\overline{D}) = \sqrt{D(\overline{D})}$$

$$D(\overline{\sigma}) \approx \frac{\overline{m_4} - \overline{\sigma}^4}{4N\overline{\sigma}^2} \quad \sigma(\overline{\sigma}) = \sqrt{D(\overline{\sigma})}$$

где $\overline{m_4}$ - это оценка четвертого центрального момента распределения, которая приведена в следующем параграфе.

4.4. Оценка коэффициента асимметрии и эксцесса.

Оценки третьего и четвертого моментов распределения по выборке $\{x_k\}$, $k=1, \dots, N$ определяются как:

$$\overline{m_3} = \frac{N}{(N-1)(N-2)} \sum_{k=1}^N (x_k - \overline{X})^3$$

$$\begin{aligned} \overline{m_4} = & \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \sum_{k=1}^N (x_k - \bar{X})^4 - \\ & - \frac{3(2N-3)}{N(N-1)(N-2)(N-3)} \sum_{k=1}^N (x_k - \bar{X})^2 \sum_{k=1}^N (x_k - \bar{X})^2 \end{aligned}$$

Следовательно, оценки коэффициента асимметрии и эксцесса можно найти по формулам:

$$\begin{aligned} \overline{\gamma} = & \frac{1}{\sigma^3} \frac{N}{(N-1)(N-2)} \sum_{k=1}^N (x_k - \bar{X})^3 \\ \overline{\varepsilon} = & \frac{1}{\sigma^4} \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \sum_{k=1}^N (x_k - \bar{X})^4 - \frac{3(2N-3)(N-1)}{N(N-2)(N-3)} \end{aligned}$$

В случае небольших выборок и при наличии промахов вместо среднего арифметического \bar{X} следует применять $X_{\text{ЦЕНТР}}$.

Дисперсии оценок коэффициента асимметрии и эксцесса можно оценить как:

$$\begin{aligned} D(\overline{\gamma}) = & \frac{6(N-1)}{(N+1)(N+3)} \\ D(\overline{\varepsilon}) = & \frac{24N(N-2)(N-3)}{(N-1)^2(N+3)(N+5)} \end{aligned}$$

Считается, что если $|\overline{\gamma}| / \sqrt{D(\overline{\gamma})} > 3$, то распределение несимметрично. Если же $|\overline{\gamma}| / \sqrt{D(\overline{\gamma})} < 3$, то асимметрия несущественна и ее наличие может быть объяснено случайностью выборки.

4.5. Исключение промахов из выборки.

Промахами в выборке случайных величин будем называть anomalно отклоняющиеся от центра распределения значения по сравнению с основной массой данных.

В применении к ценам активов, эти аномалии могут быть вызваны сменой президента или правительства, банкротством крупных компаний, террористическими актами и т.п.

Решение о том, фильтровать промахи или нет, каждый принимает для себя сам. Однако следует учесть, что промахи могут существенно исказить оценку параметров распределения.

В этом параграфе излагается формализованная процедура удаления аномальных величин из выборки. Прежде всего, введем понятие коэффициента цензурирования. Коэффициент цензурирования - это безразмерная величина G , такая, что все значения из выборки $\{x_k\}$, лежащие за пределами интервала

$X_{\text{ЦЕНТР}} - G \cdot \overline{\sigma} \leq x \leq X_{\text{ЦЕНТР}} + G \cdot \overline{\sigma}$, считаются промахами и подлежат исключению из выборки.

Интуитивно понятно, что коэффициент цензурирования должен зависеть от объема выборки и рассчитанного по выборке значения эксцесса. Действительно, такое отклонение от центра, которое является промахом для средневершинного (а тем более плосковершинного) распределения, для островершинного распределения с его длинными "тяжелыми" спадами может безусловно принадлежать выборке.

Эмпирическая формула для коэффициента цензурирования как функции от объема выборки N и эксцесса ε , пригодная к применению для широкого класса распределений следующая:

$$G = 1.55 + 0.8 \cdot \lg(N/10) \cdot \sqrt{\varepsilon - 1}.$$

Теперь все готово для написания алгоритма удаления промахов из выборки:

- 1) Вычислить величину $X_{\text{ЦЕНТР}}$,
- 2) Вычислить оценку среднеквадратичного отклонения $\overline{\sigma}$, при этом в качестве центра распределения использовать $X_{\text{ЦЕНТР}}$,
- 3) Вычислить оценку эксцесса $\overline{\varepsilon}$, при этом в качестве центра распределения использовать $X_{\text{ЦЕНТР}}$,
- 4) Вычислить коэффициент цензурирования G ,
- 5) Исключить из выборки значения, лежащие за пределами интервала $X_{\text{ЦЕНТР}} - G \cdot \overline{\sigma} \leq x \leq X_{\text{ЦЕНТР}} + G \cdot \overline{\sigma}$

После удаления промахов нужно пересчитать параметры распределения. При этом в качестве центра распределения уже можно использовать среднее арифметическое \overline{X} , как состоятельную и несмещенную оценку математического ожидания.

5. СТАТИСТИЧЕСКИЕ ВЫВОДЫ

5.1. Введение.

Какие выводы о некотором параметре генеральной совокупности мы можем сделать, имея выборочное значение этого параметра? Ответ на этот вопрос зависит от того, имеем ли мы априорную информацию о величине генерального параметра.

Если априорная информация о величине генерального параметра отсутствует, то мы можем по выборочному значению *оценить этот параметр*, задав для него *доверительный интервал*, то есть границы, в которых его величина лежит с определенной доверительной вероятностью.

Если есть априорные соображения о величине генерального параметра, то мы можем *проверить гипотезу* о том, соответствует ли выборочная оценка априорному значению генерального параметра.

5.2. Выборочное распределение выборочной средней.

Пусть случайная величина X имеет математическое ожидание μ и генеральную дисперсию σ^2 . Оценками математического ожидания и дисперсии по выборке (x_1, x_2, \dots, x_N) будут выборочная средняя и выборочная дисперсия:

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \bar{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2$$

Рассмотрим случайную величину $t = (\bar{X} - \mu) / (\bar{\sigma} / \sqrt{N})$. Так как $M(\bar{X}) = \mu$ и $\sigma(\bar{X}) = \bar{\sigma} / \sqrt{N}$, то эта случайная величина имеет нулевое математическое ожидание и единичную дисперсию.

Будем считать, что величина t подчиняется распределению Стьюдента с $\nu = N - 1$ степенями свободы, хотя в общем случае это утверждение некорректно. Дело в том, что строго говоря величина t подчиняется распределению Стьюдента только в случае когда выборка (x_1, x_2, \dots, x_N) взята из нормально распределенной совокупности.

5.3. Доверительный интервал для генеральной средней.

Доверительный интервал возможных значений величины t , характеризующийся доверительной вероятностью P или уровнем значимости $q = 1 - P$, это такой интерквантильный промежуток $t_{q/2, \nu} \leq t \leq t_{1-q/2, \nu}$, внутри которого лежат $100P$ процентов всех значений случайной величины t , а $100q$ процентов лежат вне этого промежутка. При этом $100q/2$ процентов лежит слева от $t_{q/2, \nu}$ и $100q/2$ процентов лежит справа от $t_{1-q/2, \nu}$.

Величины $t_{q/2, \nu}$ и $t_{1-q/2, \nu}$ - это квантили распределения Стьюдента с $\nu = N - 1$ степенями свободы, причем, так как это распределение симметрично и имеет нулевое математическое ожидание, то $t_{q/2, \nu} = -t_{1-q/2, \nu}$. Используя последнее равенство и подставив значение $t = (\bar{X} - \mu) / (\bar{\sigma} / \sqrt{N})$ получаем, что

$$-t_{1-q/2, \nu} \leq \frac{\bar{X} - \mu}{\bar{\sigma} / \sqrt{N}} \leq t_{1-q/2, \nu}$$

Отсюда следует, что доверительный интервал для математического ожидания μ через выборочную среднюю и выборочное с.к.о. задается в виде:

$$\bar{X} - t_{1-q/2, \nu} \frac{\bar{\sigma}}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{1-q/2, \nu} \frac{\bar{\sigma}}{\sqrt{N}}$$

Ширина доверительного интервала для математического ожидания очень существенно зависит от объема выборки. Проиллюстрируем это на простом примере. Пусть в двух испытаниях получены одинаковые значения выборочной средней $\bar{X} = 1.2$ и выборочного с.к.о. $\bar{\sigma} = 2.5$. Но в первом случае эти данные были получены по выборке объемом $N = 100$, а во втором случае по выборке объемом $N = 25$. Зададимся уровнем значимости $q = 0.05$.

Вычислим с помощью функций Microsoft Excel доверительные интервалы для математического ожидания:

1) Большая выборка

$$\bar{X} = 1.2 \quad \bar{\sigma} = 2.5 \quad N = 100$$

$$t_{1-q/2, v} = \text{СТЮДРАСПОБР}(q, N - 1) =$$

$$= \text{СТЮДРАСПОБР}(0.05, 99) = 1.984$$

$$1.2 - 1.984 \frac{2.5}{\sqrt{100}} \leq \mu \leq 1.2 + 1.984 \frac{2.5}{\sqrt{100}}$$

$$0.704 \leq \mu \leq 1.696$$

$$\text{Ширина доверительного интервала} = 1.696 - 0.704 = 0.992$$

2) Малая выборка

$$\bar{X} = 1.2 \quad \bar{\sigma} = 2.5 \quad N = 25$$

$$t_{1-q/2, v} = \text{СТЮДРАСПОБР}(q, N - 1) =$$

$$= \text{СТЮДРАСПОБР}(0.05, 24) = 2.064$$

$$1.2 - 2.064 \frac{2.5}{\sqrt{25}} \leq \mu \leq 1.2 + 2.064 \frac{2.5}{\sqrt{25}}$$

$$0.168 \leq \mu \leq 2.232$$

$$\text{Ширина доверительного интервала} = 2.232 - 0.168 = 2.064$$

То есть для данных значений выборочной средней и выборочного с.к.о. увеличение объема выборки в $100/25=4$ раза привело к уменьшению ширины доверительного интервала для математического ожидания в $2.064/0.992=2.08$ раза.

5.4. Выборочное распределение выборочной дисперсии.

Пусть случайная величина X имеет математическое ожидание μ и генеральную дисперсию σ^2 . Оценками математического ожидания и дисперсии по выборке (x_1, x_2, \dots, x_N) будут выборочная средняя и выборочная дисперсия:

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \bar{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2$$

Рассмотрим случайную величину $\chi^2 = (N-1)\bar{\sigma}^2 / \sigma^2$. Эта величина подчиняется χ^2 -распределению с $\nu = N-1$ степенями свободы, если выборочная средняя \bar{X} нормально распределена. Для малых выборок это χ^2 -распределение имеет положительную асимметрию, но с увеличением объема выборки его асимметрия стремится к нулю.

5.5. Доверительный интервал для генеральной дисперсии.

Доверительный интервал возможных значений величины χ^2 , характеризующийся доверительной вероятностью P или уровнем значимости $q = 1 - P$, это такой интерквантильный промежуток $\chi_{q/2, \nu}^2 \leq \chi^2 \leq \chi_{1-q/2, \nu}^2$, внутри которого лежат $100P$ процентов всех значений случайной величины χ^2 , а $100q$ процентов лежат вне этого промежутка. При этом $100q/2$ процентов лежит слева от $\chi_{q/2, \nu}^2$ и $100q/2$ процентов лежит справа от $\chi_{1-q/2, \nu}^2$.

Величины $\chi_{q/2, \nu}^2$ и $\chi_{1-q/2, \nu}^2$ - это квантили χ^2 -распределения с $\nu = N-1$ степенями свободы. Подставив значение $\chi^2 = (N-1)\bar{\sigma}^2 / \sigma^2$ получаем, что

$$\chi_{q/2, \nu}^2 \leq (N-1) \frac{\bar{\sigma}^2}{\sigma^2} \leq \chi_{1-q/2, \nu}^2$$

Отсюда следует, что доверительный интервал для генеральной дисперсии через выборочную дисперсию задается в виде:

$$\frac{(N-1)\bar{\sigma}^2}{\chi_{1-q/2, \nu}^2} \leq \sigma^2 \leq \frac{(N-1)\bar{\sigma}^2}{\chi_{q/2, \nu}^2}$$

Пусть в испытании получено значение выборочного с.к.о. $\bar{\sigma} = 2.5$ по выборке объемом $N = 25$. Зададимся уровнем значимости $q = 0.05$.

Вычислим с помощью функций Microsoft Excel доверительный интервалы для генеральной дисперсии:

$$\bar{\sigma} = 2.5 \quad N = 25$$

$$\nu = N - 1 = 25 - 1 = 24$$

$$q/2 = 0.025 \Rightarrow \chi^2_{0.025, 24} = \text{ХИ2ОБР}(1 - 0.025, 24) = 12.40$$

$$1 - q/2 = 0.975 \Rightarrow \chi^2_{0.975, 24} = \text{ХИ2ОБР}(1 - 0.975, 24) = 39.36$$

$$24 \frac{2.5^2}{39.36} \leq \sigma^2 \leq 24 \frac{2.5^2}{12.4}$$

$$3.81 \leq \sigma^2 \leq 12.10$$

$$\text{Ширина доверительного интервала} = 12.10 - 3.81 = 8.29$$

5.6. Статистическая проверка гипотез.

Статистическая гипотеза - это предположительное суждение о закономерностях, которым подчиняется случайная величина. Мы будем рассматривать гипотезы о величине параметров закона распределения вероятностей и о его виде.

Статистическая проверка гипотез - это система приемов, предназначенных для проверки соответствия эмпирических данных некоторой статистической гипотезе. Процесс проверки базируется на формулировании 2-х гипотез - нулевой и альтернативной:

- *нулевая гипотеза* H_0 - это гипотеза, которая считается верной до тех пор, пока не будет доказано обратное исходя из результатов статистической проверки,
- *альтернативная гипотеза* H_1 - это гипотеза, которая принимается, если в результате статистической проверки отвергается нулевая гипотеза.

Критерий проверки

Правило, по которому принимается или отклоняется нулевая гипотеза, называется *статистическим критерием проверки*. Построение критерия определяется выбором некоторой функции Q от результатов наблюдений, которая служит мерой расхождения между эмпирическими и теоретическими значениями. Функция Q называется *статистикой критерия* и является случайной величиной.

По распределению статистики Q находится такое значение Q_0 , что если гипотеза H_0 верна, то вероятность того, что $Q > Q_0$ равна q , где q - это заданный заранее *уровень значимости*. Если $Q \leq Q_0$, то гипотеза H_0 принимается, а если $Q > Q_0$, то гипотеза H_0 отвергается.

Ошибки 1-го и 2-го рода

При решении вопроса о справедливости гипотезы H_0 могут быть допущены ошибки двух видов:

- *ошибка первого рода* происходит тогда, когда отвергается верная гипотеза H_0 ,
- *ошибка второго рода* происходит тогда, когда принимается ложная гипотеза H_0 .

Уровень значимости

Очевидно, что уровень значимости q - это вероятность ошибки первого рода. Если он чрезмерно велик, то в основном ущерб будет связан с ошибочным отклонением верной гипотезы H_0 , если же он чрезмерно мал, то ущерб будет возникать от ошибочного принятия ложной гипотезы H_0 . На практике в качестве уровня значимости выбирают вероятность в пределах от 0.01 до 0.1.

5.7. Проверка гипотез о величине генеральной средней.

Располагая априорными суждениями о величине генеральной средней (математического ожидания) мы можем проверить гипотезу о том, соответствует ли выборочная средняя априорному значению математического ожидания.

Проверка гипотезы о соответствии выборочной средней априорному значению математического ожидания может быть односторонней (правосторонней или левосторонней) или двусторонней:

- *двусторонняя проверка* используется в том случае, когда необходимо проверить, равна ли выборочная средняя априор

ному значению математического ожидания, и гипотеза формулируется в виде:

$$H_0: \bar{X} = \mu$$

$$H_1: \bar{X} \neq \mu$$

- *правосторонняя проверка* используется в том случае, когда необходимо проверить, что выборочная средняя больше, чем априорное значение математического ожидания, и гипотеза формулируется в виде:

$$H_0: \bar{X} = \mu$$

$$H_1: \bar{X} > \mu$$

- *левосторонняя проверка* используется в том случае, когда необходимо проверить, что выборочная средняя меньше, чем априорное значение математического ожидания, и гипотеза формулируется в виде:

$$H_0: \bar{X} = \mu$$

$$H_1: \bar{X} < \mu$$

Проиллюстрируем проверку гипотез на примерах.

Двусторонняя проверка гипотез

1) Априорная информация

Математическое ожидание $\mu = 1$

2) Результаты испытания

$$N = 100 \quad \bar{X} = 1.2 \quad \bar{\sigma} = 2.5$$

3) Гипотеза

$$H_0: \bar{X} = \mu$$

$$H_1: \bar{X} \neq \mu$$

4) Принятая величина уровня значимости

$$q = 0.05$$

5) Критерий проверки

$$t = \frac{\bar{X} - \mu}{\bar{\sigma} / \sqrt{N}}$$

6) Правило принятия решения

Принять H_0 , если $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $|t| > t_{1-q/2, \nu}$.

7) Расчет границ критической области

$$t_{1-q/2, \nu} = \text{СТЮДРАСПОБР}(q, N-1) = \\ = \text{СТЮДРАСПОБР}(0.05, 99) = 1.984$$

8) Расчет критерия проверки

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} = \frac{1.2 - 1}{2.5 / \sqrt{100}} = 0.8$$

9) Проверка гипотезы

Так как $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$, то критерий проверки $t = 0.8$ не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости выборочная средняя $\bar{X} = 1.2$ статистически незначимо отличается от априорной величины математического ожидания $\mu = 1$.

Правосторонняя проверка гипотез

1) Априорная информация

Математическое ожидание $\mu = 0.7$

2) Результаты испытания

$$N = 100 \quad \bar{X} = 1.2 \quad \sigma = 2.5$$

3) Гипотеза

$$H_0: \bar{X} = \mu$$

$$H_1: \bar{X} > \mu$$

4) Принятая величина уровня значимости

$$q = 0.05$$

5) Критерий проверки

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

6) Правило принятия решения

Принять H_0 , если $t \leq t_{1-q, \nu}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $t > t_{1-q, \nu}$.

7) Расчет границ критической области

$$t_{1-q, \nu} = \text{СТБЮДРАСПОБР}(2q, N-1) = \\ = \text{СТБЮДРАСПОБР}(2 \times 0.05, 99) = 1.66$$

8) Расчет критерия проверки

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} = \frac{1.2 - 0.7}{2.5 / \sqrt{100}} = 2$$

9) Проверка гипотезы

Так как $t > t_{1-q, \nu}$, то критерий проверки $t = 2$ находится в критической области и мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости выборочная средняя $\bar{X} = 1.2$ статистически значимо отличается от априорной величины математического ожидания $\mu = 0.7$.

Левосторонняя проверка гипотез

1) Априорная информация

Математическое ожидание $\mu = 1.5$

2) Результаты испытания

$$N = 100 \quad \bar{X} = 1.2 \quad \bar{\sigma} = 2.5$$

3) Гипотеза

$$H_0: \bar{X} = \mu$$

$$H_1: \bar{X} < \mu$$

4) Принятая величина уровня значимости

$$q = 0.05$$

5) Критерий проверки

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

6) Правило принятия решения

Принять H_0 , если $t \geq -t_{1-q, \nu}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $t < -t_{1-q, \nu}$.

7) Расчет границ критической области

$$\begin{aligned} -t_{1-q, \nu} &= -\text{СТБЮДРАСПОБР}(2q, N-1) = \\ &= -\text{СТБЮДРАСПОБР}(2 \times 0.05, 99) = -1.66 \end{aligned}$$

8) Расчет критерия проверки

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} = \frac{1.2 - 1.5}{2.5 / \sqrt{100}} = -1.2$$

9) Проверка гипотезы

Так как $t \geq -t_{1-q, \nu}$, то критерий проверки $t = -1.2$ не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости выборочная средняя $\bar{X} = 1.2$ статистически незначимо отличается от априорной величины математического ожидания $\mu = 1.5$.

5.8. Проверка гипотез о величине генеральной дисперсии.

Располагая априорными суждениями о величине генеральной дисперсии мы можем проверить гипотезу о том, соответствует ли выборочная дисперсия априорному значению генеральной дисперсии.

Проверка гипотезы для дисперсии может быть односторонней (правосторонней или левосторонней) или двусторонней:

- *двусторонняя проверка* используется в том случае, когда необходимо проверить, равна ли выборочная дисперсия априорному значению генеральной дисперсии, и гипотеза формулируется в виде:

$$H_0: \sigma^2 = \sigma^2$$

$$H_1: \sigma^2 \neq \sigma^2$$

- *правосторонняя проверка* используется в том случае, когда необходимо проверить, что выборочная дисперсия больше,

чем априорное значение генеральной дисперсии, и гипотеза формулируется в виде:

$$H_0: \overline{\sigma}^2 = \sigma^2$$

$$H_1: \overline{\sigma}^2 > \sigma^2$$

- левосторонняя проверка используется в том случае, когда необходимо проверить, что выборочная дисперсия меньше, чем априорное значение генеральной дисперсии, и гипотеза формулируется в виде:

$$H_0: \overline{\sigma}^2 = \sigma^2$$

$$H_1: \overline{\sigma}^2 < \sigma^2$$

Проиллюстрируем проверку гипотез на примерах.

Двусторонняя проверка гипотез

1) Априорная информация

Генеральная дисперсия $\sigma^2 = 4$

2) Результаты испытания

$$N = 25 \quad \overline{\sigma} = 2.5 \quad \overline{\sigma}^2 = 6.25$$

3) Гипотеза

$$H_0: \overline{\sigma}^2 = \sigma^2$$

$$H_1: \overline{\sigma}^2 \neq \sigma^2$$

4) Принятая величина уровня значимости

$$q = 0.05$$

5) Критерий проверки

$$\chi^2 = (N - 1) \frac{\overline{\sigma}^2}{\sigma^2}$$

6) Правило принятия решения

Принять H_0 , если $\chi_{q/2, v}^2 \leq \chi^2 \leq \chi_{1-q/2, v}^2$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки χ^2 попадает в критическую область $\chi^2 < \chi_{q/2, v}^2$ или $\chi^2 > \chi_{1-q/2, v}^2$

7) Расчет границ критической области

$$\chi^2_{q/2, \nu} = \chi^2_{0.025, 24} = \text{ХИ2ОБР}(1 - 0.025, 24) = 12.40$$

$$\chi^2_{1-q/2, \nu} = \chi^2_{0.975, 24} = \text{ХИ2ОБР}(1 - 0.975, 24) = 39.36$$

8) Расчет критерия проверки

$$\chi^2 = (N - 1) \frac{\overline{\sigma^2}}{\sigma^2} = 24 \frac{6.25}{4} = 37.50$$

9) Проверка гипотезы

Так как $\chi^2_{q/2, \nu} \leq \chi^2 \leq \chi^2_{1-q/2, \nu}$, то критерий проверки $\chi^2 = 37.50$ не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости выборочная дисперсия $\overline{\sigma^2} = 6.25$ статистически незначимо отличается от априорной величины генеральной дисперсии $\sigma^2 = 4$.

Правосторонняя проверка гипотез

1) Априорная информация

Генеральная дисперсия $\sigma^2 = 3.6$

2) Результаты испытания

$$N = 25 \quad \overline{\sigma} = 2.5 \quad \overline{\sigma^2} = 6.25$$

3) Гипотеза

$$H_0: \overline{\sigma^2} = \sigma^2$$

$$H_1: \overline{\sigma^2} > \sigma^2$$

4) Принятая величина уровня значимости

$$q = 0.05$$

5) Критерий проверки

$$\chi^2 = (N - 1) \frac{\overline{\sigma^2}}{\sigma^2}$$

6) Правило принятия решения

Принять H_0 , если $\chi^2 \leq \chi^2_{1-q, \nu}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки χ^2 попадает в критическую область $\chi^2 > \chi^2_{1-q, \nu}$

- 7) Расчет границ критической области

$$\chi^2_{1-q, \nu} = \chi^2_{0.95, 24} = \text{ХИ2ОБР}(1 - 0.95, 24) = 36.42$$

- 8) Расчет критерия проверки

$$\chi^2 = (N - 1) \frac{\bar{\sigma}^2}{\sigma^2} = 24 \frac{6.25}{3.6} = 41.67$$

- 9) Проверка гипотезы

Так как $\chi^2 > \chi^2_{1-q, \nu}$, то критерий проверки $\chi^2 = 41.67$ находится в критической области и мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости выборочная дисперсия $\bar{\sigma}^2 = 6.25$ статистически значимо отличается от априорной величины генеральной дисперсии $\sigma^2 = 3.6$.

Левосторонняя проверка гипотез

- 1) Априорная информация

Генеральная дисперсия $\sigma^2 = 9$

- 2) Результаты испытания

$$N = 25 \quad \bar{\sigma} = 2.5 \quad \bar{\sigma}^2 = 6.25$$

- 3) Гипотеза

$$H_0: \bar{\sigma}^2 = \sigma^2$$

$$H_1: \bar{\sigma}^2 < \sigma^2$$

- 4) Принятая величина уровня значимости

$$q = 0.05$$

- 5) Критерий проверки

$$\chi^2 = (N - 1) \frac{\bar{\sigma}^2}{\sigma^2}$$

- 6) Правило принятия решения

Принять H_0 , если $\chi^2 \geq \chi^2_{q, \nu}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки χ^2 попадает в критическую область $\chi^2 < \chi_{q, \nu}^2$

7) *Расчет границ критической области*

$$\chi_{q, \nu}^2 = \chi_{0.05, 24}^2 = \text{ХИ2ОБР}(1 - 0.05, 24) = 13.85$$

8) *Расчет критерия проверки*

$$\chi^2 = (N - 1) \frac{\bar{\sigma}^2}{\sigma^2} = 24 \frac{6.25}{9} = 16.67$$

9) *Проверка гипотезы*

Так как $\chi^2 \geq \chi_{q, \nu}^2$, то критерий проверки $\chi^2 = 16.67$ не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости выборочная дисперсия $\bar{\sigma}^2 = 6.25$ статистически незначимо отличается от априорной величины генеральной дисперсии $\sigma^2 = 9$.

6. ИДЕНТИФИКАЦИЯ ЗАКОНА РАСПРЕДЕЛЕНИЯ ПО ВЫБОРКЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ.

6.1. Введение.

В данной главе будет рассмотрен вопрос о том, как по эмпирической выборке идентифицировать закон распределения случайной величины.

Подробно рассмотрена проблема группировки данных, то есть расчет оптимального количества интервалов группировки и оптимальной ширины интервала, а также построения по сгруппированным данным гистограммы распределения.

Полученное эмпирическое распределение будет аппроксимировано непрерывной аналитической функцией, то есть будет идентифицирован закон распределения случайной величины. Также рассмотрено использование критериев согласия при идентификации закона распределения.

В качестве выборки случайной величины использована выборка, состоящая из логарифмов относительного изменения величины индекса Российской торговой системы (индекса РТС) за период с 1 сентября 1995 года по 31 декабря 2002 года.

6.2. Группировка данных. Оптимальное число интервалов группировки.

Для расчета оценок математического ожидания, дисперсии, среднеквадратичного отклонения, коэффициента асимметрии и эксцесса (на основе моментов распределения) не требуется предварительного упорядочивания и группировки данных. Эти величины могут быть найдены непосредственно по исходной выборке.

Для определения медианы, квантилей распределения, для удаления промахов из выборки данные необходимо расположить в порядке возрастания, то есть упорядочить выборку.

Группировка данных необходима для того, чтобы найти форму распределения, то есть, в конечном итоге идентифицировать закон распределения.

В результате группировки выборка представляется в виде гистограммы, состоящей из L столбцов (интервалов

группировки), каждый из которых имеет ширину d . После нормирования гистограмма представляет собой эмпирическую плотность распределения случайной величины.

Из качественных соображений следует, что должно существовать оптимальное число интервалов группировки.

Действительно, при большом количестве столбцов и поэтому малой ширине столбца, из-за случайности выборки гистограмма будет заполнена очень неравномерно, иметь сильно изрезанный вид, состоять из большого количества всплесков и провалов.

При другой крайности, то есть очень малом числе столбцов большой ширины, гистограмма будет излишне сглаживать распределение, уничтожать его характерные особенности. Например, если выбрать только один интервал группировки с шириной, равной размаху выборки, то любое распределение сведется к прямоугольному. Два столбца выбирать нельзя, так как любое симметричное распределение, как и в предыдущем случае, сведется к прямоугольному. Три столбца также дают мало информации о форме распределения.

Эти сугубо качественные рассуждения показывают, что должно существовать некоторое оптимальное количество интервалов группировки.

Если исходить из предположения, что генеральная совокупность, из которой получена данная конкретная выборка, имеет гладкую кривую плотности вероятности (это справедливо в большинстве случаев), то неравномерности гистограммы являются случайным шумом, обусловленным случайностью выборки. Увеличение ширины столбца и уменьшение количества столбцов фильтруют этот шум. Однако, дальнейшее увеличение ширины столбца начинает сглаживать уже само распределение.

Следовательно, определение оптимального числа интервалов группировки при построении гистограммы является задачей оптимальной фильтрации. При этом *оптимальное количество столбцов гистограммы - это такое количество, при котором максимально возможное сглаживание случайного шума сочетается с минимальным искажением от сглаживания самого распределения.*

Оптимальное число столбцов должно зависеть не только от объема выборки, как это указано в большинстве пособий по статистике. Очевидно, что это число зависит еще и от формы распределения. Действительно, если плосковершинные распределения можно приблизить достаточно малым количеством столбцов, то для островершинных распределений с их длинными, пологими спадами это количество естественно должно быть больше.

Количество интервалов группировки должно быть нечетным числом. При четном числе столбцов область вблизи центра распределения будет описываться двумя симметрично расположенными относительно центра столбцами гистограммы, тем самым пик распределения будет неоправданно сглаживаться. Это особенно критично для островершинных распределений. Как уже говорилось выше, три столбца дают очень мало информации о форме распределения. Поэтому будем считать, что количество столбцов гистограммы должно быть нечетным числом не менее пяти.

Эмпирическая формула для оценки оптимального количества столбцов гистограммы как функции от объема выборки N и эксцесса ε , пригодная к применению для широкого класса распределений следующая:

$$L = \frac{\varepsilon + 1.5}{6} N^{0.4}$$

Вычисленное по этой формуле значение должно быть округлено вниз до ближайшего большего или равного пяти нечетного целого.

Используя значение L , ширину столбца гистограммы можно найти по формуле: $d = \frac{2 \max(|x_k - \bar{X}|)}{L}$

6.3. Построение гистограммы распределения.

Изложим алгоритм построения гистограммы по выборке случайных величин $\{x_k\}$, $k = 1, \dots, N$:

- 1) Упорядочить исходную выборку по возрастанию.
- 2) Вычислить оценки центра распределения:

$X_{\text{медиана}}, X_{\text{центр_сгибов}}, \bar{X}, \overline{X}_{50\%}, X_{\text{центр_размаха}}$

Упорядочить эти оценки по возрастанию и выбрать из них в качестве центра распределения срединное, то есть третье по счету, значение, которое обозначить как $X_{\text{ЦЕНТР}}$.

- 3) Вычислить оценку среднеквадратичного отклонения

$$\bar{\sigma} = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (x_k - X_{\text{ЦЕНТР}})^2}$$

- 4) Вычислить оценку эксцесса

$$\begin{aligned} \bar{\varepsilon} = & \frac{1}{\sigma^4} \cdot \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \sum_{k=1}^N (x_k - X_{\text{ЦЕНТР}})^4 - \\ & - \frac{3(2N-3)(N-1)}{N(N-2)(N-3)} \end{aligned}$$

- 5) Вычислить коэффициент цензурирования

$$G = 1.55 + 0.8 \cdot \lg(N/10) \cdot \sqrt{\bar{\varepsilon} - 1}$$

- 6) Исключить из выборки все значения (промахи), лежащие за пределом интервала

$$X_{\text{ЦЕНТР}} - G \cdot \bar{\sigma} \leq x \leq X_{\text{ЦЕНТР}} + G \cdot \bar{\sigma}$$

Если в выборке присутствовали промахи, то ее объем уменьшился. Обозначим как $\{x_k\}, k = 1, \dots, M$ очищенную от промахов выборку ($M \leq N$). Все дальнейшие операции будут проводиться с очищенной выборкой.

- 7) Заново вычислить параметры распределения

$$\bar{X} = \frac{1}{M} \sum_{k=1}^M x_k$$

$$\bar{\sigma} = \sqrt{\frac{1}{M-1} \sum_{k=1}^M (x_k - \bar{X})^2}$$

$$\bar{\gamma} = \frac{1}{\sigma^3} \cdot \frac{M}{(M-1)(M-2)} \sum_{k=1}^M (x_k - \bar{X})^3$$

$$\varepsilon = \frac{1}{\sigma^4} \cdot \frac{M^2 - 2M + 3}{(M-1)(M-2)(M-3)} \sum_{k=1}^M (x_k - \bar{X})^4 - \frac{3(2M-3)(M-1)}{M(M-2)(M-3)}$$

- 8) Рассчитать оптимальное количество столбцов гистограммы

$$L = \frac{\varepsilon + 1.5}{6} M^{0.4}$$

Полученное число округлить вниз до ближайшего большего или равного пяти нечетного целого.

- 9) Рассчитать левую и правую границы гистограммы

$$X_{\min} = \bar{X} - \max(|x_k - \bar{X}|)$$

$$X_{\max} = \bar{X} + \max(|x_k - \bar{X}|)$$

- 10) Рассчитать ширину столбца гистограммы

$$d = \frac{X_{\max} - X_{\min}}{L} = \frac{2 \max(|x_k - \bar{X}|)}{L}$$

- 11) Рассчитать массив узлов разбиения на оси x

$$X_i = X_{\min} + (i-1) \cdot d$$

$$i = 1, \dots, L+1$$

Интервалы между соседними узлами являются интервалами разбиения.

- 12) Рассчитать количество случайных величин из выборки $\{x_k\}$, $k = 1, \dots, M$, которое попадает в каждый из интервалов разбиения. В результате получится ненормированная гистограмма распределения или гистограмма частот. Она задана в виде массива, который обозначим как $\{s_i\}$, $i = 1, \dots, L$.

- 13) В случае, если есть основания полагать, что плотность вероятности должна быть симметричной, и в подтверждение этого, вычисленный на шаге 7 коэффициент асимметрии незначительно отличается от нуля, то можно провести расчетное симметрирование гистограммы. Центральный столбец остается без изменения, а в симметричных

относительно него парам столбцов количество отсчетов усредняется.

- 14) Вычислить площадь S ненормированной гистограммы. Она должна быть равна произведению ширины столбца d на объем выборки M .
- 15) Нормировать гистограмму путем деления количества отсчетов в каждом столбце на S . Таким образом на этом шаге получена *гистограмма плотности вероятности*:

$$p_i = s_i / S = s_i / (d \cdot M)$$

$$i = 1, \dots, L$$

- 16) Рассчитать значения *интегральной функции распределения* в узлах разбиения

$$F_1 \equiv 0$$

$$F_i = F_{i-1} + p_{i-1} \cdot d$$

$$i = 2, \dots, L + 1$$

Фактически, на этом шаге мы получили функцию распределения в табличном виде, то есть мы имеем массив значений случайной величины $\{X_i\}$ и соответствующий ему массив значений $\{F_i\}, i = 1, \dots, L + 1$.

6.4. Гистограмма логарифмов относительных изменений индекса РТС.

Рассмотрим временной ряд, состоящий из последовательных значений цены некоторого актива $\{P_t\}, t = 0, \dots, T$. Тогда цену в момент времени T можно представить, как

$$P_T = P_0 \frac{P_1}{P_0} \frac{P_2}{P_1} \dots \frac{P_t}{P_{t-1}} \dots \frac{P_T}{P_{T-1}}$$

Движение цены актива - это случайный процесс, вызванный действиями большого количества участников рынка. Предположим, что отношения цен активов в любой момент времени являются случайными величинами с одинаковым законом распределения.

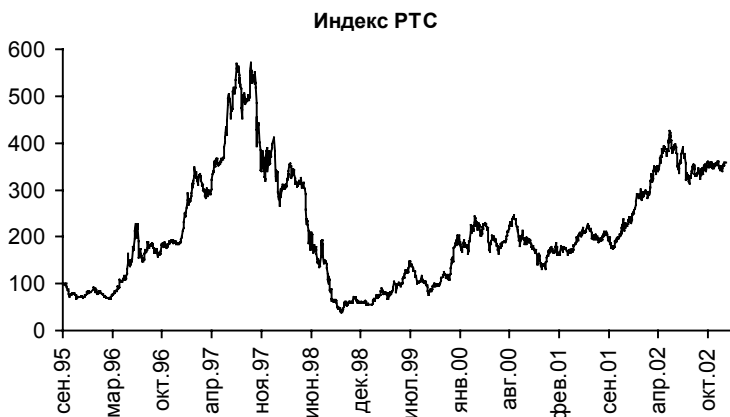
Тогда по выборке этих случайных величин, которая может быть получена из ценового ряда, можно определить их закон распределения.

Но исследовать непосредственно отношение цен представляется не совсем удобным. Дело в том, что так как цена не может упасть ниже нуля, то отношение цен также не может быть меньше нуля. С другой стороны, цена может сколь угодно сильно вырасти, то есть отношение цен может быть неограниченно большим. Этих качественных рассуждений достаточно, чтобы понять, что плотность вероятности отношения цен будет иметь положительную асимметрию. Однако, если мы перейдем к логарифмам отношения цен, ситуация изменится.

$$\ln(P_T) - \ln(P_0) = \ln\left(\frac{P_1}{P_0}\right) + \ln\left(\frac{P_2}{P_1}\right) + \dots + \ln\left(\frac{P_t}{P_{t-1}}\right) + \dots + \ln\left(\frac{P_T}{P_{T-1}}\right)$$

Распределение логарифмов уже может быть симметрично и возможна его аппроксимация одним из аналитических законов распределения, которые были рассмотрены во второй главе.

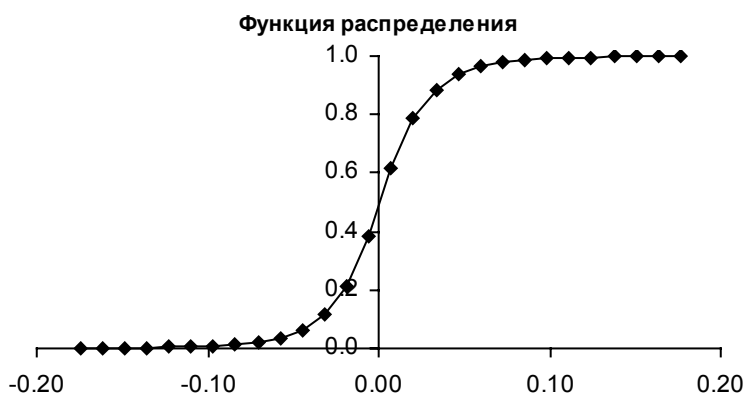
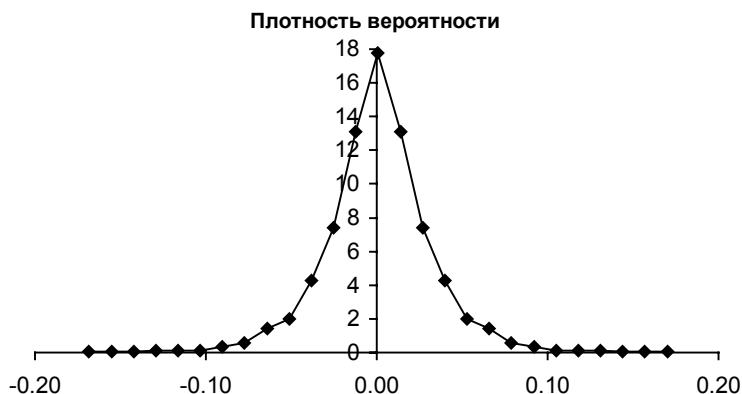
В качестве примера ценового ряда рассмотрим индекс Российской торговой системы (индекс РТС) за период с 1 сентября 1995 года по 31 декабря 2002 года. График этого ряда изображен на рисунке:



Исследуемой выборкой случайных величин будут *натуральные логарифмы отношения цен закрытия* индекса РТС.

Подробный алгоритм вычисления параметров распределения, построения графиков плотности распределения и функции распределения рассмотрен в предыдущем параграфе, поэтому здесь приведем только результаты.

Наименование оценки	Величина
Центр распределения (математическое ожидание)	0.0007
Среднеквадратичное отклонение	0.0324
Коэффициент асимметрии	-0.3064
Экссесс	7.4301



Отметим, что с.к.о. превышает математическое ожидание более чем в 46 раз, то есть исследуемая случайная величина является высоковолатильной. Распределение имеет очень небольшую отрицательную асимметрию, которая вероятно носит случайный характер, поэтому гистограмма плотности вероятности была центрирована.

Экссесс распределения существенно превышает эксцесс нормального распределения, то есть данное распределение является островершинным.

Гистограмма распределения имеет 27 столбцов. Для большей наглядности плотность вероятности приведена не в виде гистограммы, а как плавная линия, проходящая через середины интервалов разбиения.

6.5. Использование критериев согласия при идентификации закона распределения случайной величины.

После построения гистограммы распределения можно выдвинуть гипотезу о том, что данная гистограмма может быть аппроксимирована одним из изученных ранее законов распределения. При этом степень близости гистограммы и принятой аналитической модели может быть проверена с использованием критериев согласия. Здесь будет рассмотрен один из этих критериев - критерий χ^2 Пирсона.

При использовании критерия согласия Пирсона необходимо вычислить величину:

$$\chi^2 = \sum_{i=1}^L \frac{(T_i - s_i)^2}{T_i}$$

где

L - количество столбцов гистограммы,

s_i - фактическая частота попадания в i -й столбец,

T_i - теоретическая частота попадания в i -й столбец.

Для идеально подобранной модели все разности $(T_i - s_i)$ равны нулю и, следовательно, величина χ^2 также равна нулю. Таким образом, ненулевое значение χ^2 является мерой суммарного расхождения между фактическим распределением и моделью.

Насколько велико это расхождение можно проверить, сравнив фактическое значение χ^2 с теоретической величиной $\chi^2_{1-q, \nu}$, которая определяет максимально возможное расхождение между фактическими данными и моделью, соответствующее принятому уровню значимости q .

Уровень значимости q определяет вероятность ошибки 1-го рода, то есть вероятность того, что будет отвергнута не противоречащая эмпирическим данным модель.

Величина ν - это число степеней свободы χ^2 -распределения. Число степеней свободы зависит от количества столбцов гистограммы эмпирических данных L и количества параметров r , описывающих теоретическую модель: $\nu = L - 1 - r$.

Величина $\chi^2_{1-q, \nu}$ - это такая квантиль χ^2 -распределения, что $100(1-q)$ процентов всех значений случайной величины χ^2 лежат слева от $\chi^2_{1-q, \nu}$, а $100q$ процентов всех значений случайной величины χ^2 лежат справа от $\chi^2_{1-q, \nu}$.

Если $\chi^2 \leq \chi^2_{1-q, \nu}$, то считают, что модель не противоречит фактическим данным при заданном уровне значимости.

Если $\chi^2 > \chi^2_{1-q, \nu}$, то считают, что при заданном уровне значимости модель не описывает удовлетворительным образом фактические данные и должна быть отвергнута.

Следует особо подчеркнуть, что при проверке модели по критерию согласия определенным является лишь отрицательный ответ, то есть отклонение модели.

Положительный ответ означает лишь то, что модель не противоречит эмпирическим данным. Это вовсе не означает, что именно этой моделью данные описываются на самом деле, что это наилучшая модель, что нельзя подобрать другую модель для описания данных и т.д. Фактически, положительный ответ при проверке по критерию согласия следует понимать как "возможно эти данные описываются такой-то моделью", и не более того.

Вернемся к полученной в предыдущем параграфе гистограмме натуральных логарифмов относительного изменения цены закрытия индекса РТС.

Гистограмма имеет ярко выраженный пик и достаточно пологие спады. Острровершинность подтверждается еще и значением эксцесса, существенно превышающим эксцесс нормального распределения. Как нам уже известно, распределения с подобными характеристиками могут быть описаны обобщенным экспоненциальным распределением с показателем степени меньше двух.

Выдвинем гипотезу о том, что фактическое распределение описывается моделью

$$p(x) = \frac{\alpha}{2\Gamma(1/\alpha)\lambda\sigma} \exp\left(-\left|\frac{x-\mu}{\lambda\sigma}\right|^\alpha\right) \quad -\infty < x < +\infty$$

где

математическое ожидание $\mu = 0.0007$

среднеквадратичное отклонение $\sigma = 0.0324$

показатель степени $\alpha = 0.87$

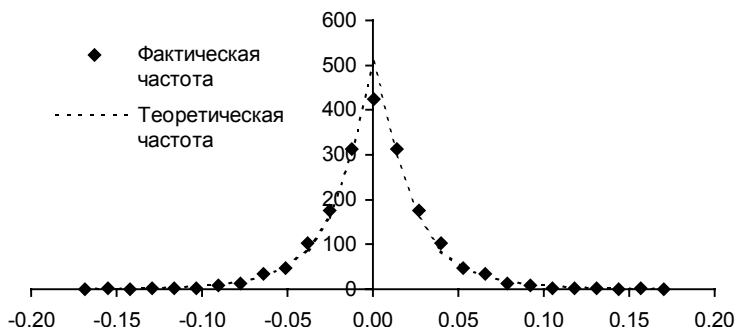
Показатель степени был найден из значения оценки эксцесса распределения, так как для обобщенного экспоненциального распределения показатель степени и эксцесс имеют взаимно однозначное соответствие:

$$\varepsilon = \Gamma(1/\alpha)\Gamma(5/\alpha)/[\Gamma(3/\alpha)]^2$$

Исследуемое эмпирическое распределение имеет 27 столбцов. Аналитическая модель имеет 3 параметра. Следовательно, число степеней свободы для критерия Пирсона равно $\nu = L - 1 - r = 27 - 1 - 3 = 23$.

Фактические и теоретические частоты попадания в столбцы гистограммы дадим для наглядности в графическом виде.

Распределение фактических и теоретических частот



Фактическое значение $\chi^2 = 35.635$. Проверим гипотезу о том, что $\chi^2 \leq \chi^2_{1-q, \nu}$.

$$H_0 : \chi^2 \leq \chi^2_{1-q, \nu}$$

$$H_1 : \chi^2 > \chi^2_{1-q, \nu}$$

Пусть уровень значимости $q = 0.01$. Тогда граница критической области вычисляется как:

$$\chi^2_{1-q, \nu} = \text{ХИ2ОБР}(0.01, 23) = 41.638$$

Так как $\chi^2 \leq \chi^2_{1-q, \nu}$, то исследуемое распределение при заданном уровне значимости можно аппроксимировать обобщенным экспоненциальным распределением.

В заключении следует сказать, что для ликвидных российских акций, торгующихся в РТС, таких как РАО ЕЭС, Лукойл, Сургутнефтегаз, Ростелеком, Мосэнерго, распределения логарифмов относительного изменения цены закрытия также можно описать обобщенным экспоненциальным распределением с соответствующим математическим ожиданием, среднеквадратичным отклонением и показателем степени.

7. КОРРЕЛЯЦИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

7.1. Введение.

Существует два типа зависимостей между переменными: функциональная (строго детерминированная) и статистическая (стохастически детерминированная).

В случае *функциональной* зависимости каждому значению одной переменной соответствует одно или несколько строго заданных значений другой переменной. Функциональная связь двух переменных возможна, если вторая переменная зависит от первой и ни от чего более. На практике таких связей не существует, то есть функциональная связь является упрощающей реальность абстракцией.

В случае *статистической* связи каждому значению одной величины соответствует определенное распределение вероятности другой величины. Это связано с тем, что в любой математической модели на описываемый показатель влияют не только явным образом входящие в модель переменные, но и большое количество факторов, которые существуют в действительности, но не учитываются моделью, причем часть из этих факторов - это случайные величины. Этим можно объяснить случайный характер многих финансовых переменных и взаимосвязей между ними.

Важнейшим частным случаем статистической связи является *корреляционная* связь, когда каждому значению одной переменной соответствует определенное *математическое ожидание* другой переменной, и при изменении значения одной величины математическое ожидание другой величины изменяется закономерным образом. Если же при изменении значения одной переменной закономерным образом изменяется другая статистическая характеристика второй переменной (дисперсия, асимметрия, эксцесс и т.д.), то связь является статистической, но не корреляционной. Данная глава посвящена изучению линейной корреляционной связи между случайными величинами.

7.2. Функция регрессии.

Рассмотрим две непрерывные случайные величины X и Y . Тогда вероятность того, что в некотором испытании величина X

окажется в интервале от x до $x + dx$, а величина Y окажется в интервале от y до $y + dy$ равна $p_{xy}(x, y)dxdy$. Величина $p_{xy}(x, y)$ называется плотностью двумерного распределения вероятностей величин X и Y .

Для двумерного распределения вероятностей плотность распределения координат x и y выражается формулами:

$$p_x(x) = \int_{-\infty}^{+\infty} p_{xy}(x, y)dy$$

$$p_y(y) = \int_{-\infty}^{+\infty} p_{xy}(x, y)dx$$

Случайные величины X и Y находятся в корреляционной зависимости, если:

- каждому значению переменной X соответствует определенное математическое ожидание переменной Y ,
- каждому значению переменной Y соответствует определенное математическое ожидание переменной X .

Рассмотрим условное распределение вероятности переменной Y при фиксированном значении переменной X . Оно описывается условной плотностью распределения:

$$p_{y|x}(x, y) = p_{xy}(x, y) / p_x(x)$$

Используя условную плотность распределения можно найти математическое ожидание случайной величины Y , при условии того, что случайная величина X равна фиксированному значению x (условное математическое ожидание):

$$M_{y|x}(x) = \int_{-\infty}^{+\infty} y \cdot p_{y|x}(x, y)dy$$

Условное математическое ожидание $M_{y|x}(x)$ называют еще *функцией регрессии* Y на X . Функция регрессии обладает важнейшим свойством: среднеквадратичное отклонение случайной величины Y от функции регрессии Y на X меньше, чем ее среднеквадратичное отклонение от любой другой функции от x .

Если функцию регрессии можно удовлетворительным образом аппроксимировать линейной зависимостью, то такая регрессия

называется линейной. Линейная регрессия обладает тем свойством, что если регрессия Y на X линейна, то регрессия X на Y также линейна.

Заметим, что функции регрессии X на Y и Y на X не являются взаимно обратными и соответствующие линии регрессии совпадают только в случае, когда величины Y и X связаны функционально. Если эти величины связаны корреляционно, то линии регрессии X на Y и Y на X различны.

В дальнейшем мы ограничимся рассмотрением только тех случаев, когда функция регрессии является линейной.

7.3. Линейная корреляция.

Корреляционная зависимость между случайными величинами X и Y называется линейной корреляцией, если обе функции регрессии X на Y и Y на X являются линейными.

Пусть математическое ожидание и дисперсия случайной величины X равны μ_x, σ_x^2 , а математическое ожидание и дисперсия случайной величины Y равны μ_y, σ_y^2 .

Выведем уравнение регрессии Y на X , то есть найдем коэффициенты линейной функции $y = ax + b$.

- 1) Выразим коэффициент b через математические ожидания X и Y

$$\mu_y \equiv M(y) = M(ax + b) = aM(x) + b = a\mu_x + b$$

$$b = \mu_y - a\mu_x$$

- 2) Тогда уравнение регрессии можно переписать в виде

$$y = ax + \mu_y - a\mu_x$$

$$y - \mu_y = a \cdot (x - \mu_x)$$

- 3) Найдем коэффициент регрессии a через математическое ожидание произведения случайных величин X и Y

$$M(xy) = M[x(ax + \mu_y - a\mu_x)]$$

$$M(xy) = aM(x^2) + M(x)\mu_y - aM(x)\mu_x$$

$$M(xy) = aM(x^2) + \mu_x\mu_y - a\mu_x^2$$

$$M(xy) = a[M(x^2) - \mu_x^2] + \mu_x \mu_y$$

$$M(xy) = a\sigma_x^2 + \mu_x \mu_y$$

$$a = \frac{M(xy) - \mu_x \mu_y}{\sigma_x^2}$$

- 4) Назовем *коэффициентом корреляции* между X и Y следующую безразмерную и симметричную относительно X и Y величину

$$\rho = M\left(\frac{(x - \mu_x)}{\sigma_x} \cdot \frac{(y - \mu_y)}{\sigma_y}\right) = \frac{M[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y}$$

- 5) Тогда математическое ожидание произведения случайных величин X и Y можно выразить через коэффициент корреляции

$$M(xy) \equiv M[(x - \mu_x + \mu_x)(y - \mu_y + \mu_y)]$$

$$M(xy) = M[(x - \mu_x)(y - \mu_y)] + \mu_x M(y - \mu_y) +$$

$$+ \mu_y M(x - \mu_x) + \mu_x \mu_y$$

$$M(xy) = M[(x - \mu_x)(y - \mu_y)] + \mu_x \mu_y$$

$$M(xy) = \rho \sigma_x \sigma_y + \mu_x \mu_y$$

- 6) Окончательно для коэффициента регрессии Y на X получаем

$$a = \rho \cdot (\sigma_y / \sigma_x)$$

- 7) В итоге уравнение регрессии Y на X приобретает вид

$$y - \mu_y = \rho \cdot (\sigma_y / \sigma_x) \cdot (x - \mu_x)$$

Тангенс угла наклона, под которым эта прямая пересекает ось x равен $\rho \cdot (\sigma_y / \sigma_x)$.

- 8) Аналогично можно получить уравнение регрессии X на Y

$$x - \mu_x = \rho \cdot (\sigma_x / \sigma_y) \cdot (y - \mu_y)$$

Тангенс угла наклона, под которым эта прямая пересекает ось x равен $(1 / \rho) \cdot (\sigma_y / \sigma_x)$.

Заметим, что прямые регрессии Y на X и X на Y пересекают ось x под разными углами. Эти прямые совпадают только тогда, когда модуль коэффициента корреляции $|\rho| = 1$. Обе прямые регрес

сии проходят через центр двумерного распределения вероятностей величин X и Y - точку с координатами (μ_x, μ_y) .

7.4. Коэффициент корреляции. Ковариация.

Рассмотрим подробнее введенный в предыдущем параграфе *коэффициент корреляции*. Было выяснено, что он равен

$$\rho_{xy} = \rho_{yx} = M \left(\frac{(x - \mu_x)}{\sigma_x} \cdot \frac{(y - \mu_y)}{\sigma_y} \right) = \frac{M(xy) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Следовательно, коэффициент корреляции характеризует относительное отклонение математического ожидания произведения двух случайных величин от произведения математических ожиданий этих величин. Так как отклонение имеет место только для зависимых величин, то коэффициент корреляции характеризует степень этой зависимости.

Коэффициент корреляции обладает следующими свойствами:

- 1) Линейные преобразования случайных величин X и Y не изменяют коэффициента корреляции между ними
- 2) Коэффициент корреляции случайных величин X и Y заключен в пределах между -1 и +1, достигая этих крайних значений только в случае линейной функциональной зависимости между X и Y .
- 3) Коэффициент корреляции между независимыми случайными величинами равен нулю.

Обратное утверждение вообще говоря неверно, то есть если коэффициент корреляции равен нулю, то это не означает независимости соответствующих величин. В этом случае говорят, что величины некоррелированы.

Как уже говорилось выше, коэффициент корреляции является безразмерной величиной. Произведение коэффициента корреляции на среднеквадратичные отклонения случайных величин X и Y имеет размерность дисперсии и называется *ковариацией* случайных величин X и Y :

$$\text{cov}(x, y) \equiv \sigma_{xy} = \sigma_{yx} = M[(x - \mu_x)(y - \mu_y)] = M(xy) - \mu_x \mu_y$$

7.5. Математическое ожидание и дисперсия линейной комбинации случайных величин.

В этом параграфе мы рассмотрим правила вычисления математического ожидания и дисперсии многомерной случайной величины, являющейся линейной комбинацией коррелированных случайных величин:

$$M\left(a_0 + \sum_{k=1}^N a_k x_k\right) \text{ и } D\left(a_0 + \sum_{k=1}^N a_k x_k\right)$$

Математическое ожидание

Математическое ожидание обладает следующими свойствами:

- 1) Постоянный множитель можно выносить за знак математического ожидания

$$M(ax) = aM(x) \equiv a\mu_x$$

- 2) Математическое ожидание суммы случайной величины и константы равно сумме математического ожидания этой величины и константы

$$M(x + a) = M(x) + a \equiv \mu_x + a$$

- 3) Математическое ожидание суммы случайных величин равно сумме их математических ожиданий

$$M(x + y) = M(x) + M(y) \equiv \mu_x + \mu_y$$

Следовательно, для линейной комбинации произвольного количества случайных величин получаем

$$M\left(a_0 + \sum_{k=1}^N a_k x_k\right) = a_0 + \sum_{k=1}^N a_k M(x_k) \equiv a_0 + \sum_{k=1}^N a_k \mu_k$$

Дисперсия

Аналогичные свойства для дисперсии следующие:

- 1) Постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат

$$D(ax) = a^2 D(x) \equiv a^2 \sigma_x^2$$

- 2) Дисперсия суммы случайной величины и константы равна дисперсии случайной величины

$$D(x + a) = D(x) \equiv \sigma_x^2$$

- 3) Дисперсия суммы случайных величин равно сумме их дисперсий плюс удвоенное произведение их коэффициента корреляции на среднеквадратичные отклонения

$$\begin{aligned} D(x+y) &= M[(x+y) - (\mu_x + \mu_y)]^2 = \\ &= M(x - \mu_x)^2 + M(y - \mu_y)^2 + 2M[(x - \mu_x)(y - \mu_y)] = \\ &= \sigma_x^2 + \sigma_y^2 + 2\rho_{xy}\sigma_x\sigma_y \end{aligned}$$

Следовательно, для линейной комбинации произвольного количества случайных величин получаем

$$D\left(a_0 + \sum_{k=1}^N a_k x_k\right) = \sum_{k=1}^N a_k^2 \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N a_i a_k \rho_{ik} \sigma_i \sigma_k$$

Если все случайные величины независимы, то так как коэффициенты корреляции для различных случайных величин равны 0, а коэффициент корреляции случайной величины с самой собой равен 1, формула упрощается

$$D\left(a_0 + \sum_{k=1}^N a_k x_k\right) = \sum_{k=1}^N a_k^2 \sigma_k^2$$

Полученные выражения для математического ожидания и дисперсии линейной комбинации произвольного количества коррелированных случайных величин позволяют сделать следующие выводы:

- *математическое ожидание линейной комбинации случайных величин* - это взвешенная сумма математических ожиданий отдельных случайных величин,
- *дисперсия линейной комбинации случайных величин* - это взвешенная сумма ковариаций всех пар случайных величин, при этом вес каждой ковариации равен произведению весов соответствующей пары случайных величин, а ковариация случайной величины с самой собой является дисперсией данной величины.

7.6. Оценка ковариации и коэффициента корреляции по выборке случайных величин.

Для оценки ковариации и коэффициента корреляции между случайными величинами X и Y мы должны располагать двумя соответствующими друг другу выборками этих величин:

$$\{x_k\}, \{y_k\} \quad k = 1, \dots, N$$

Оценка ковариации

В качестве оценки математического ожидания случайных величин X и Y используем средние арифметические значения по соответствующим выборкам:

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N y_k$$

Тогда выборочная ковариация случайных величин X и Y задается формулой:

$$\overline{\sigma_{xy}} = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})$$

Оценка коэффициента корреляции

Для оценки коэффициента корреляции между случайными величинами X и Y нам понадобятся выборочные среднеквадратичные отклонения этих величин:

$$\overline{\sigma_x} = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2} \quad \overline{\sigma_y} = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (y_k - \bar{Y})^2}$$

Тогда выборочный коэффициент корреляции случайных величин X и Y задается формулой:

$$\rho_{xy} = \frac{\overline{\sigma_{xy}}}{\overline{\sigma_x} \cdot \overline{\sigma_y}} = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sqrt{\sum_{k=1}^N (x_k - \bar{X})^2} \sqrt{\sum_{k=1}^N (y_k - \bar{Y})^2}}$$

Дисперсию и с.к.о. выборочного коэффициента корреляции можно оценить как

$$\sigma_{\rho}^2 = \frac{(1 - \rho_{xy}^2)^2}{N-1} \quad \sigma_{\rho} = \frac{1 - \rho_{xy}^2}{\sqrt{N-1}}$$

7.7. Оценка коэффициентов линейной регрессии по выборке случайных величин.

В параграфе 7.3 было получено, что в случае, когда величины X и Y представлены своими генеральными совокупностями, уравнение регрессии Y на X имеет вид:

$$y - \mu_y = \rho \cdot (\sigma_y / \sigma_x) \cdot (x - \mu_x)$$

Следовательно, так как $\rho \cdot (\sigma_y / \sigma_x) \equiv \sigma_{xy} / \sigma_x^2$, то коэффициенты (a, b) линейной регрессии $y = ax + b$ можно представить в виде:

$$a = \sigma_{xy} / \sigma_x^2 \quad b = \mu_y - a\mu_x$$

Переходя к выборочным оценкам получаем:

$$\bar{a} = \frac{\overline{\sigma_{xy}}}{\overline{\sigma_x^2}} = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (x_k - \bar{X})^2} = \frac{\sum_{k=1}^N x_k y_k - N \cdot \bar{X} \cdot \bar{Y}}{\sum_{k=1}^N x_k^2 - N \cdot \bar{X}^2}$$

$$\bar{b} = \bar{Y} - \bar{a} \cdot \bar{X}$$

Аналогичным образом можно получить оценку коэффициентов линейной регрессии X на Y .

7.8. Линейная регрессия как наилучшая оценка по методу наименьших квадратов.

Докажем, что полученные в предыдущем параграфе оценки коэффициентов линейной регрессии Y на X определяют такую прямую линию, что сумма квадратов отклонений величины Y от этой прямой имеет минимальное значение, по сравнению с суммой квадратов отклонений величины Y от любой другой прямой.

Пусть величины X и Y представлены своими выборками:

$$\{x_k\}, \{y_k\} \quad k = 1, \dots, N$$

Предположим, что зависимость величины Y от величины X можно аппроксимировать прямой линией $y = \alpha x + \beta$. Найдем коэффициенты α и β , которые минимизируют сумму квадратов отклонений величины Y от этой прямой:

$$S = \sum_{k=1}^N (y_k - \alpha x_k - \beta)^2$$

Возьмем частные производные S по α и по β , и приравняем их к нулю:

$$\frac{\partial S}{\partial \beta} = -2 \sum_{k=1}^N (y_k - \alpha x_k - \beta) = 0$$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{k=1}^N x_k (y_k - \alpha x_k - \beta) = 0$$

Следовательно:

$$\sum_{k=1}^N y_k - \alpha \sum_{k=1}^N x_k - \beta N = 0$$

$$\sum_{k=1}^N x_k y_k - \alpha \sum_{k=1}^N x_k^2 - \beta \sum_{k=1}^N x_k = 0$$

Из первого уравнения этой системы следует, что

$$\beta = \frac{1}{N} \sum_{k=1}^N y_k - \alpha \frac{1}{N} \sum_{k=1}^N x_k = \bar{Y} - \alpha \cdot \bar{X}$$

Подставив это выражение во второе уравнение системы после несложных преобразований получим:

$$\alpha = \frac{\sum_{k=1}^N x_k y_k - N \cdot \bar{X} \cdot \bar{Y}}{\sum_{k=1}^N x_k^2 - N \cdot \bar{X}^2}$$

Использованный метод поиска коэффициентов α и β называется методом наименьших квадратов. Сравнивая коэффициенты α и β с полученными в предыдущем параграфе выборочными коэффициентами линейной регрессии видим, что они совпадают. Следовательно, утверждение о том, что коэффициенты линейной регрессии Y на X определяют такую прямую линию, что сумма квадратов отклонений величины Y от этой прямой имеет минимальное значение, по сравнению с суммой квадратов отклонений величины Y от любой другой прямой, доказано.

8. РЕГРЕССИОННЫЙ АНАЛИЗ

8.1. Введение.

Различные экономические и финансовые переменные связаны между собой. Если не принимать во внимание случайный характер этих переменных, то для описания связей между ними можно применить функциональный подход, то есть предположить, что связь одной из переменных Y с некоторым количеством других переменных (X_1, \dots, X_M) можно выразить некоторой функцией (математической моделью):

$$Y = f(a_1, \dots, a_L, X_1, \dots, X_M), \text{ где}$$

- (X_1, \dots, X_M) - это набор независимых переменных, которые будем называть *факторами*,
- Y - это зависимая переменная, которую будем называть *откликом*,
- (a_1, \dots, a_L) - это набор констант, которые будем называть *параметрами* математической модели.

В случае, когда отклик Y зависит только от единственного фактора X , модель называется *однофакторной*. Если отклик Y зависит от нескольких факторов (X_1, \dots, X_M) , модель называется *многофакторной*.

Математическая модель, связывающая факторы и отклик, может быть найдена только на основе реальных выборок этих величин. Определение модели включает в себя два этапа:

- выбор вида модели, то есть вида функции f ,
- расчет параметров выбранной модели (a_1, \dots, a_L) .

Первый этап, то есть выбор вида математической модели, является не формализуемой задачей. Это решение принимается с учетом простоты и удобства использования модели, содержательности модели и других соображений. Второй этап, то есть расчет параметров выбранной математической модели, является задачей, которая решается с помощью *регрессионного анализа* реальных выборок факторов и отклика.

8.2. Выбор вида математической модели.

Рассмотрим однофакторную зависимость. Этот случай наиболее прост и может быть изучен графически. Предположим,

что имеется массив значений фактора X и соответствующий ему массив значений отклика Y . Нанесем соответствующие точки $(x_k, y_k), k = 1, \dots, N$ на график. Если фактор и отклик - это реальные статистические данные, то указанные точки никогда не лягут на простую линию (прямую, параболу, гиперболу, экспоненту, синусоиду и т.д.). Всегда будут присутствовать отклонения, связанные со случайным характером рассматриваемых переменных и/или с влиянием неучтенных факторов.

Кроме того часто оказывается, что один и тот же набор точек можно с примерно одинаковой точностью описать различными аналитическими функциями. Следовательно, выбор вида математической модели - это не формализуемая задача. Рациональный выбор той или иной модели может быть обоснован лишь с учетом определенных требований, а именно:

- простоты модели,
- содержательности модели.

Простота модели

Наиболее распространенной ошибкой при описании фактической зависимости является попытка детерминированного описания этой зависимости, то есть включение в математическую модель всех наблюдающихся особенностей конкретной выборки, в том числе и тех, которые в действительности носят случайный характер.

Например, любой набор точек $(x_k, y_k), k = 1, \dots, N$ можно описать абсолютно точно полиномом $(N-1)$ -й степени, зависящим от N параметров $(a_0, a_1, \dots, a_{N-1})$:

$$y = \sum_{k=0}^{N-1} a_k x^k$$

Но на практике получается, что появляющаяся новая $(N+1)$ -я точка уже не будет удовлетворять полученной формуле. То же самое можно сказать обо всех появляющихся далее новых точках. При этом расхождение между реальными данными и моделью будет нарастать с увеличением количества новых данных.

В то же время может оказаться, что исходный набор значений (x_k, y_k) можно приближенно описать какой-либо простой функцией (прямой, параболой, гиперболой, экспонен

той, синусоидой и т.д.), и эта модель, зависящая от небольшого числа параметров, будет устойчива к появлению новых данных. Следовательно, необходимым требованием к математической модели является ее простота.

Содержательность модели

Под содержательностью математической модели будем понимать разумную интерпретируемость результатов, которые могут быть получены при вычислении по этой модели.

Поясним это утверждение на простом примере. Пусть наша задача состоит в том, чтобы описать кривую зависимости цены бескупонной облигации от срока до погашения облигации. В данном случае фактором X является срок до погашения, откликом Y является цена. На эту математическую модель можно наложить очевидные ограничения:

- 1) функция $y = f(x)$ должна быть неотрицательной,
- 2) функция $y = f(x)$ должна быть монотонно убывающей,
- 3) значение функции $y = f(x)$ при $x = 0$ должно быть равно константе (номиналу облигации),
- 4) значение функции $y = f(x)$ при $x \rightarrow \infty$ должно стремиться к нулю.

Приведем примеры функций, которые не удовлетворяют хотя бы одному из ограничений и поэтому не могут быть использованы для построения рассматриваемой модели *из соображений содержательности*:

- линейная функция $y = b - ax$ не удовлетворяет первому и четвертому условию, так как при $x \rightarrow \infty$ величина $y \rightarrow -\infty$,
- гипербола $y = b + a/x$ не соответствует третьему условию, так как при $x \rightarrow 0$ величина $y \rightarrow \infty$.

При этом данные функции могут удовлетворительным образом описывать набор исходных данных (x_k, y_k) .

8.3. Расчет параметров математической модели.

Если выбор вида математической модели - это не формализуемая задача, то расчет параметров уже выбранной математической модели является чисто формальным процессом. В общем

из систем даст свое решение, и все эти решения будут вообще говоря разными. Если их нанести на график, то получится целый пучок аппроксимирующих кривых. Если эти кривые каким-либо образом усреднить, то полученное усредненное решение будет гораздо достовернее описывать истинную зависимость между X и Y , так как оно в значительной степени будет защищено от случайности выборки. Этот эффект усреднения тем больше, чем больше объем выборки N .

Наиболее эффективным методом усреднения решений избыточной системы уравнений является *регрессионный анализ* или *метод наименьших квадратов (МНК)*.

8.4. Сущность метода наименьших квадратов.

Пусть после предварительного анализа принято решение о том, что связь фактора X и отклика Y выражается функцией $Y = f(a_1, \dots, a_L, X)$. Наша задача состоит в том, чтобы по имеющейся выборке, то есть по набору точек $(x_k, y_k), k = 1, \dots, N$ вычислить наилучшие оценки неизвестных параметров модели (a_1, \dots, a_L) . Заметим, что все значения (x_k, y_k) - это не переменные, а конкретные числа.

Между рассчитанными по модели значениями отклика f_k и реальными значениями из выборки y_k будут присутствовать расхождения, которые обозначим как $e_k = y_k - f_k = y_k - f(a_1, \dots, a_L, x_k)$

Метод наименьших квадратов позволяет найти такой набор параметров модели, при котором сумма квадратов всех расхождений между значениями по выборке и вычисленными по модели значениями будет минимальной, то есть

$$S = \sum_{k=1}^N e_k^2 \rightarrow \min$$

$$S = \sum_{k=1}^N [y_k - f(a_1, \dots, a_L, x_k)]^2 \rightarrow \min$$

Величина S является функцией от L переменных (a_1, \dots, a_L) . Минимум этой функции можно найти, приравняв к нулю все ее

Это означает, что отсутствует систематическая ошибка в определении линии регрессии, следовательно оценки параметров регрессии являются *несмещенными*, то есть математическое ожидание оценки каждого параметра равно его истинному значению.

- 3) *Дисперсия ошибки должна быть постоянна: $D(e_k) = \sigma^2$.*

Это означает, что при увеличении объема выборки дисперсия оценок параметров регрессии стремится к нулю, то есть оценки параметров регрессии являются *состоятельными*.

- 4) *Ошибки должны быть независимыми, то есть*

$$\text{cov}(e_k, e_j) = \begin{cases} 0 & k \neq j \\ \sigma^2 & k = j \end{cases}$$

Это означает, что ошибка в одной из величин отклика Y не приводит автоматически к ошибкам в последующих величинах.

Кроме того, в МНК предполагается что *факторы (независимые переменные) не являются случайными величинами*.

8.6. Оценка параметров однофакторной линейной регрессии.

Допустим, что принята гипотеза о том, что связь фактора X и отклика Y выражается линейной функцией $f(x) = ax + b$. Наличие отклонений, связанных со случайным характером рассматриваемых переменных и/или с влиянием неучтенных факторов приведет к тому, что связь между рассчитанными по модели значениями отклика f_k и реальными значениями из выборки y_k будет выражаться в виде:

$$y_k = f_k + e_k = ax_k + b + e_k$$

где e_k - это расхождения между моделью и выборкой.

Оценка параметров линейной регрессии

Вычислим такой набор параметров модели, при котором сумма квадратов всех расхождений между значениями по выборке и вычисленными по модели значениями будет минимальной, то есть

$$S = \sum_{k=1}^N e_k^2 \rightarrow \min$$

$$S = \sum_{k=1}^N [y_k - ax_k - b]^2 \rightarrow \min$$

Величина S является функцией от 2-х переменных (a, b) . Минимум этой функции можно найти, приравняв к нулю ее частные производные по каждому из неизвестных параметров и решив полученную таким образом систему из 2-х уравнений. Так как вычисление параметров мы будем проводить по конечной выборке, то в результате мы получим лишь оценку этих параметров (\bar{a}, \bar{b}) :

$$\begin{cases} \frac{\partial S}{\partial b} = -2 \sum_{k=1}^N [y_k - \bar{a}x_k - \bar{b}] = 0 \\ \frac{\partial S}{\partial a} = -2 \sum_{k=1}^N [y_k - \bar{a}x_k - \bar{b}]x_k = 0 \end{cases}$$

Из 1-го уравнения системы получаем:

$$\sum_{k=1}^N y_k - \bar{a} \sum_{k=1}^N x_k - \bar{b}N = 0 \Rightarrow \bar{b} = \bar{Y} - \bar{a} \cdot \bar{X}$$

Из 2-го уравнения системы получаем:

$$\sum_{k=1}^N y_k x_k - \bar{a} \sum_{k=1}^N x_k^2 - \bar{b} \sum_{k=1}^N x_k = 0 \Rightarrow \sum_{k=1}^N y_k x_k - \bar{a} \sum_{k=1}^N x_k^2 - \bar{b}N\bar{X} = 0$$

Подставив в это уравнение выражение для оценки параметра b найдем оценку параметра a :

$$\bar{a} = \frac{\sum_{k=1}^N x_k y_k - N \cdot \bar{X} \cdot \bar{Y}}{\sum_{k=1}^N x_k^2 - N \cdot \bar{X}^2} \equiv \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (x_k - \bar{X})^2}$$

Из последнего равенства следует, что оценку параметра a можно выразить через ковариацию или коэффициент корреляции переменных X и Y :

$$\bar{a} = \frac{\sigma_{xy}}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x}$$

Параметр a , который еще называют коэффициентом регрессии, численно равен тангенсу угла наклона прямой регрессии к оси x .

Дисперсия оценок параметров линейной регрессии

Так как оценки параметров линейной регрессии получены по случайной выборке, то сами эти оценки являются случайными величинами. Оценка дисперсии параметра a выражается формулой:

$$\bar{\sigma}_a^2 = \frac{\bar{\sigma}_e^2}{\sum_{k=1}^N (x_k - \bar{X})^2}$$

где величина $\bar{\sigma}_e^2$ - это оценка дисперсии случайных отклонений отклика Y от линии регрессии:

$$\bar{\sigma}_e^2 = \frac{1}{N - m - 1} \sum_{k=1}^N e_k^2$$

где m - число факторов (независимых переменных).

В случае парной линейной регрессии

$$\bar{\sigma}_e^2 = \frac{1}{N - 2} \sum_{k=1}^N e_k^2 = \frac{1}{N - 2} \sum_{k=1}^N (y_k - \bar{a}x_k - \bar{b})^2$$

Так как $\bar{b} = \bar{Y} - \bar{a} \cdot \bar{X}$ и так как фактор X предполагается нестохастическим, то для оценки дисперсии параметра b справедливо:

$$\bar{\sigma}_b^2 = \bar{\sigma}_{\bar{Y}}^2 + \bar{X}^2 \cdot \bar{\sigma}_a^2$$

где величина $\bar{\sigma}_{\bar{Y}}^2$ - это оценка дисперсии среднего значения отклика Y :

$$\bar{\sigma}_{\bar{Y}}^2 = \frac{1}{N} \bar{\sigma}_e^2$$

После несложных преобразований для оценки дисперсии параметра b получаем формулу:

$$\overline{\sigma_b^2} = \frac{\overline{\sigma_e^2}}{\sum_{k=1}^N (x_k - \bar{X})^2} \cdot \frac{\sum_{k=1}^N x_k^2}{N}$$

Величину $\overline{\sigma_e^2}$ называют еще *необъясненной дисперсией*.

Чем меньше необъясненная дисперсия (то есть чем меньше отклонения величины Y от линии регрессии), тем меньше ошибки в определении параметров регрессии, и, следовательно, тем точнее модель объясняет фактические данные.

Кроме того, из формул для дисперсии параметров следует, что чем на более широком диапазоне изменения фактора X оценивается регрессия, тем больше величина $\sum (x_k - \bar{X})^2$, а значит меньше дисперсия параметров.

Из тех же самых соображений следует, что чем больше объем выборки N , тем меньше дисперсия параметров.

8.7. Коэффициент детерминации.

Из того, что связь фактора X и отклика Y выражается в виде $y_k = f_k + e_k = ax_k + b + e_k$ следует, что разброс отклика Y может быть объяснен разбросом фактора X и случайной ошибкой e . Необходимо определить индикатор, который бы показывал, насколько разброс Y определяется разбросом X и насколько случайными причинами, то есть насколько хорошо фактические данные описываются функцией регрессии.

В качестве *общей меры разброса* переменной Y естественно использовать сумму квадратов отклонений этой величины от ее среднего значения. Тогда в качестве *объясняемой регрессией меры разброса* переменной Y будем использовать сумму квадратов отклонений прогнозируемых линией регрессии значений от среднего значения величины Y .

Индикатором качества линии регрессии является *коэффициент детерминации*:

$$R^2 = \frac{\sum_{k=1}^N (f_k - \bar{Y})^2}{\sum_{k=1}^N (y_k - \bar{Y})^2} = \frac{\sum_{k=1}^N (ax_k + b - \bar{Y})^2}{\sum_{k=1}^N (y_k - \bar{Y})^2}$$

или

$$R^2 = 1 - \frac{\sum_{k=1}^N e_k^2}{\sum_{k=1}^N (y_k - \bar{Y})^2}$$

В случае однофакторной линейной регрессии коэффициент детерминации равен квадрату коэффициента корреляции величин X и Y .

Иногда при расчете коэффициента детерминации для получения несмещенных оценок дисперсии в числителе и знаменателе делается поправка на число степеней свободы, то есть *скорректированный коэффициент детерминации* вычисляется по формуле:

$$R^2 = 1 - \frac{\frac{1}{N-m-1} \sum_{k=1}^N e_k^2}{\frac{1}{N-1} \sum_{k=1}^N (y_k - \bar{Y})^2}$$

где m - число факторов (независимых переменных).

При добавлении в уравнение регрессии дополнительных объясняющих переменных (факторов) нескорректированный R^2 всегда растет. При этом скорректированный R^2 может уменьшиться за счет увеличения числа m , если новый фактор приводит к небольшому уменьшению необъясненной дисперсии.

В случае парной линейной регрессии скорректированный R^2 вычисляется как:

$$R^2 = 1 - \frac{\frac{1}{N-2} \sum_{k=1}^N e_k^2}{\frac{1}{N-1} \sum_{k=1}^N (y_k - \bar{Y})^2}$$

Коэффициент детерминации может принимать значения от нуля (когда Y не зависит от X) до единицы (когда X полностью определяет Y , то есть между ними существует строгая функциональная зависимость). Чем больше этот коэффициент, тем выше качество линии регрессии.

Запишем формулу для R^2 в компактном виде

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

Отношение ширины полосы рассеяния данных относительно их среднего значения к ширине полосы рассеяния данных относительно линии регрессии называется *числом различных градаций отклика*. Если в качестве меры рассеяния принять соответствующие среднеквадратичные отклонения, то формула для числа различных градаций отклика будет иметь вид:

$$N_{GRAD} = \sigma_y / \sigma_e$$

Как и коэффициент детерминации, число различных градаций является позитивной оценкой корреляционной связи, то есть чем больше N_{GRAD} , тем выше качество уравнения регрессии.

$$R^2 = 1 - 1/(N_{GRAD})^2$$

Негативной оценкой корреляционной связи является *относительная приведенная погрешность*, которая является отношением половины ширины полосы рассеяния данных относительно линии регрессии к ширине полосы рассеяния данных относительно их среднего значения и вычисляется по формуле

$$\gamma = 0.5 \cdot \frac{\sigma_e}{\sigma_y}, \text{ то есть } R^2 = 1 - (2\gamma)^2.$$

Связь между γ и N_{GRAD} задается формулами

$$\gamma = \frac{1}{2 \cdot N_{GRAD}} \quad N_{GRAD} = \frac{1}{2 \cdot \gamma}$$

Приведем таблицу, показывающую связь между коэффициентом детерминации, числом различных градаций отклика и относительной приведенной погрешностью.

N_{GRAD}	γ	R^2	R
1	50.0%	0.00	0.00
1.41	35.4%	0.50	0.71
2	25.0%	0.75	0.87
3	16.7%	0.89	0.94
4	12.5%	0.94	0.97
5	10.0%	0.96	0.98
6	8.3%	0.972	0.986
7	7.1%	0.980	0.990
8	6.3%	0.984	0.992
9	5.6%	0.988	0.994
10	5.0%	0.990	0.995

Отметим следующие важные случаи:

- Коэффициент детерминации $R^2 = 0.5$ ($R \approx 0.71$), то есть только половина разброса отклика Y объясняется уравнением регрессии. В этой ситуации говорят, что влияние сигнала (фактора X) равно влиянию помехи (случайной ошибки e). Поэтому при коэффициенте детерминации меньше чем 0.5, помехи начинают вносить основной вклад в вариацию переменной Y , и такая модель регрессии должна быть отвергнута.
- Если с.к.о. ошибки e ровно в два раза меньше, чем с.к.о. отклика Y , то есть число различных градаций отклика равно 2, то $R^2 = 0.75$ ($R \approx 0.87$). Именно это значение рекомендуется принять в качестве минимально приемлемого значения коэффициента детерминации.

ПРИМЕЧАНИЕ. При оценке величин γ и N_{GRAD} мы предполагали, что мерой ширины полосы рассеяния данных относительно их среднего значения и мерой ширины полосы рассеяния данных относительно линии регрессии являются соответствующие средне-квадратичные отклонения. Если в качестве меры принять доверительные интервалы, то формулы для γ и N_{GRAD} изменятся, так как отклик Y и ошибка уравнения регрессии e - это случайные величины с вообще говоря различными законами распределения. Рас

пределение величины Y , особенно при ярко выраженной линейной зависимости, близко к равномерному. Распределение величины e в большинстве случаев близко к нормальному.

8.8. Необратимость решений МНК.

Если отвлечься от причинно-следственной связи и рассматривать переменные X и Y как равноправные, то по методу наименьших квадратов можно найти линейную регрессию как Y по X так и X по Y .

Пусть линейная регрессия Y по X выражается функцией $Y = a_1 X + b_1$, а линейная регрессия X по Y функцией $X = a_2 Y + b_2$. Оценки параметров a_1 и a_2 выражаются через коэффициент корреляции между переменными X и Y как:

$$a_1 = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (x_k - \bar{X})^2} = \rho \frac{\sigma_y}{\sigma_x}$$

$$a_2 = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (y_k - \bar{Y})^2} = \rho \frac{\sigma_x}{\sigma_y}$$

Тангенс угла наклона функции $Y = a_1 X + b_1$ к оси x равен $a_1 = \rho \cdot (\sigma_y / \sigma_x)$, а тангенс угла наклона функции $X = a_2 Y + b_2$ к оси x равен $1/a_2 = (1/\rho) \cdot (\sigma_y / \sigma_x)$. Это разные величины, следовательно линии регрессии Y на X и X на Y - это разные прямые. Они совпадают только тогда, когда модуль коэффициента корреляции $|\rho| = 1$, то есть когда между переменными X и Y существует строгая функциональная зависимость.

В несовпадении линий регрессии Y на X и X на Y и состоит необратимость решений МНК, то есть нельзя использовать величины (a_2, b_2) для вычисления величин (a_1, b_1) и наоборот:

$$a_1 \neq \frac{1}{a_2} \quad b_1 \neq -\frac{b_2}{a_2} \quad a_2 \neq \frac{1}{a_1} \quad b_2 \neq -\frac{b_1}{a_1}$$

8.9. Статистические выводы о величине параметров однофакторной линейной регрессии.

Полученные в этой главе формулы для выборочных коэффициентов однофакторной линейной регрессии дают лишь оценки истинных значений этих коэффициентов.

Введем обозначения:

- истинные значения параметров линейной регрессии (a, b) ,
- выборочные значения параметров линейной регрессии (\bar{a}, \bar{b}) ,
- выборочные дисперсии параметров $(\overline{\sigma_a^2}, \overline{\sigma_b^2})$.

Выборочное распределение параметров линейной регрессии

При анализе коэффициентов регрессии считают, что случайные величины $t_a = \frac{\bar{a} - a}{\sigma_a}$ и $t_b = \frac{\bar{b} - b}{\sigma_b}$ подчиняются распределению Стьюдента с $\nu = (N - 2)$ степенями свободы, где N - объем выборки. В этих формулах:

$$\bar{a} = \frac{\sum_{k=1}^N (x_k - \bar{X})(y_k - \bar{Y})}{\sum_{k=1}^N (x_k - \bar{X})^2} \quad \bar{b} = \frac{1}{N} \sum_{k=1}^N y_k - \bar{a} \frac{1}{N} \sum_{k=1}^N x_k$$

$$\overline{\sigma_a^2} = \frac{\overline{\sigma_e^2}}{\sum_{k=1}^N (x_k - \bar{X})^2} \quad \overline{\sigma_b^2} = \frac{\overline{\sigma_e^2}}{\sum_{k=1}^N (x_k - \bar{X})^2} \cdot \frac{1}{N} \sum_{k=1}^N x_k^2$$

$$\overline{\sigma_e^2} = \frac{1}{N - 2} \sum_{k=1}^N (y_k - \bar{a}x_k - \bar{b})^2$$

Доверительный интервал для параметров линейной регрессии

Доверительный интервал возможных значений величины t , характеризующийся доверительной вероятностью P или уровнем значимости $q = 1 - P$, это такой интерквантильный

промежуток $t_{q/2, \nu} \leq t \leq t_{1-q/2, \nu}$, внутри которого лежат $100P$ процентов всех значений случайной величины t , а $100q$ процентов лежат вне этого промежутка. При этом $100q/2$ процентов лежит слева от $t_{q/2, \nu}$ и $100q/2$ процентов лежит справа от $t_{1-q/2, \nu}$.

Величины $t_{q/2, \nu}$ и $t_{1-q/2, \nu}$ - это квантили распределения Стьюдента с $\nu = N - 2$ степенями свободы, причем, так как это распределение симметрично и имеет нулевое математическое ожидание, то $t_{q/2, \nu} = -t_{1-q/2, \nu}$.

Подставив значения $t_a = (\bar{a} - a) / \overline{\sigma}_a$ и $t_b = (\bar{b} - b) / \overline{\sigma}_b$ в двойное неравенство $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$ получим доверительные интервалы для истинных значений параметров линейной регрессии (a, b) :

$$\bar{a} - t_{1-q/2, \nu} \overline{\sigma}_a \leq a \leq \bar{a} + t_{1-q/2, \nu} \overline{\sigma}_a$$

$$\bar{b} - t_{1-q/2, \nu} \overline{\sigma}_b \leq b \leq \bar{b} + t_{1-q/2, \nu} \overline{\sigma}_b$$

Гипотезы о величине параметров линейной регрессии

Когда речь идет о линейной регрессии, необходимо знать, насколько значимо отличаются от нуля величины параметров регрессии. Для проверки этого выдвигаются гипотезы:

$$H_0 : \bar{a} = 0 \quad \text{или} \quad H_0 : \bar{b} = 0$$

$$H_1 : \bar{a} \neq 0 \quad \text{или} \quad H_1 : \bar{b} \neq 0$$

Проверка данных гипотез осуществляется в отдельности для каждого из параметров по следующей схеме:

1) Априорные предположения

Истинные значения параметров регрессии равны нулю

$$a = 0$$

$$b = 0$$

2) Результаты испытания

Выборочные коэффициенты регрессии и их выборочные с.к.о.

$$\bar{a}, \bar{\sigma}_a$$

$$\bar{b}, \bar{\sigma}_b$$

при объеме выборки N .

3) *Гипотеза*

$$H_0: \bar{a} = 0 \quad \text{или} \quad H_0: \bar{b} = 0$$

$$H_1: \bar{a} \neq 0 \quad \text{или} \quad H_1: \bar{b} \neq 0$$

4) *Принятая величина уровня значимости*
 $q = 0.05$ или $q = 0.01$

5) *Критерий проверки*

$$t_a = \frac{\bar{a} - a}{\bar{\sigma}_a} = \frac{\bar{a}}{\bar{\sigma}_a}$$

$$t_b = \frac{\bar{b} - b}{\bar{\sigma}_b} = \frac{\bar{b}}{\bar{\sigma}_b}$$

6) *Правило принятия решения*

Принять H_0 , если $-t_{1-q/2, v} \leq t \leq t_{1-q/2, v}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $|t| > t_{1-q/2, v}$.

Граница критической области вычисляется как

$$t_{1-q/2, v} = \text{СТЮДРАСПОБР}(q, N - 2)$$

В качестве критерия проверки t используются t_a и t_b .

7) *Проверка гипотезы*

- Если $-t_{1-q/2, v} \leq t \leq t_{1-q/2, v}$ то критерий проверки t не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости соответствующий параметр регрессии статистически незначимо отличается от нуля.
- В противном случае мы принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости соответствующий параметр регрессии статистически значимо отличается от нуля.

8.10. Статистические выводы о величине коэффициента детерминации.

Коэффициент детерминации является индикатором того, насколько хорошо изменения фактора X объясняют изменения отклика Y . Чем он ближе к единице, тем выше качество уравнения регрессии.

Так как коэффициент детерминации вычисляется по конечной случайной выборке, то он сам является случайной величиной. Проверка значимости коэффициента детерминации - это проверка гипотезы о том, что он значимо отличается от нуля.

$$H_0 : R^2 = 0$$

$$H_1 : R^2 > 0$$

Критерий проверки рассчитывается по формуле:

$$F = \frac{R^2 / m}{(1 - R^2) / (N - m - 1)}$$

где N - объем выборки, m - количество независимых переменных (факторов). Критерий проверки подчиняется F -распределению с m степенями свободы для числителя и $(N - m - 1)$ степенями свободы для знаменателя.

В случае однофакторной линейной регрессии критерий проверки принимает вид:

$$F = \frac{R^2}{(1 - R^2) / (N - 2)}$$

Количество степеней свободы для числителя равно 1, количество степеней свободы для знаменателя равно $(N - 2)$.

Если в действительности переменная Y не зависит от переменной X , то коэффициент детерминации R^2 и критерий проверки F равны нулю. При этом их оценки по случайной выборке могут отличаться от нуля, но чем больше это отличие, тем менее оно вероятно.

Если же критерий проверки F больше некоторого критического значения при заданном уровне доверительной вероятности, то это событие считается слишком маловероятным

и мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 . Это значит, что переменная Y зависит от переменной X .

Проверка гипотезы для однофакторной линейной регрессии проводится по следующей схеме:

1) *Гипотеза*

$$H_0 : R^2 = 0$$

$$H_1 : R^2 > 0$$

2) *Принятая величина уровня значимости*

$$q = 0.05 \text{ или } q = 0.01$$

3) *Критерий проверки*

$$F = \frac{R^2}{1 - R^2} (N - 2)$$

4) *Правило принятия решения*

Принять H_0 , если $F \leq F_{1-q, \nu_1, \nu_2}$.

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки F попадает в критическую область $F > F_{1-q, \nu_1, \nu_2}$.

Здесь F_{1-q, ν_1, ν_2} - это квантиль F -распределения, соответствующая уровню значимости q с $\nu_1 = 1$ степенями свободы для числителя и $\nu_2 = N - 2$ степенями свободы для знаменателя.

Величину F_{1-q, ν_1, ν_2} можно вычислить с помощью электронных таблиц Microsoft Excel:

$$F_{1-q, \nu_1, \nu_2} = \text{FPАСПОБР}(q, \nu_1, \nu_2)$$

5) *Проверка гипотезы*

- Если $F \leq F_{1-q, \nu_1, \nu_2}$, то критерий проверки F не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости изменения фактора X не объясняют изменения отклика Y и регрессионная модель должна быть отвергнута.
- В противном случае мы принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости переменная Y зависит от переменной X .

8.11. Полоса неопределенности однофакторной линейной регрессии.

Так как параметры линейной регрессии зависимы между собой ($\bar{b} = \bar{Y} - \bar{a} \cdot \bar{X}$), то уравнение регрессии можно переписать в виде $f = ax + \bar{b} = \bar{a} \cdot (x - \bar{X}) + \bar{Y}$. Каждая точка на линии регрессии выражается через выборочные значения (\bar{a}, \bar{Y}) , имеющие выборочные дисперсии $(\overline{\sigma_a^2}, \overline{\sigma_Y^2})$, и потому является случайной величиной.

Дисперсия линии регрессии

Так как в МНК предполагается, что фактор X нестохастичен, то дисперсию точки на линии регрессии можно выразить следующим образом:

$$\overline{\sigma_f^2} = (x - \bar{X})^2 \cdot \overline{\sigma_a^2} + \overline{\sigma_Y^2}$$

Из этой формулы следует, что:

- дисперсия величины \bar{Y} влияет на дисперсию точки на линии регрессии *аддитивным* образом, то есть ее вклад постоянен и не зависит от величины фактора X ,
- дисперсия величины a влияет на дисперсию точки на линии регрессии *мультипликативным* образом, то есть ее вклад тем больше, чем больше абсолютное отклонение фактора X от \bar{X} .

С учетом того, что

$$\overline{\sigma_a^2} = \frac{\overline{\sigma_e^2}}{\sum_{k=1}^N (x_k - \bar{X})^2} \quad \overline{\sigma_Y^2} = \frac{1}{N} \overline{\sigma_e^2}$$

для дисперсии точки на линии регрессии получим:

$$\overline{\sigma_f^2} = \overline{\sigma_e^2} \cdot \left(\frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_{k=1}^N (x_k - \bar{X})^2} \right)$$

Доверительный интервал линии регрессии

Аналогично тому, как мы нашли доверительные интервалы для истинных параметров линейной регрессии, мы можем записать доверительный интервал для линии регрессии в виде:

$$\bar{f} - t_{1-q/2, v} \overline{\sigma_f} \leq f \leq \bar{f} + t_{1-q/2, v} \overline{\sigma_f}$$

Ширина доверительного интервала линии регрессии равна $2t_{1-q/2, v} \overline{\sigma_f}$. Эту величину называют еще шириной полосы неопределенности линии регрессии.

8.12. Прогнозирование на основе однофакторной линейной регрессии.

При прогнозировании, то есть при экстраполяции линии регрессии за пределы поля точек, по которым была получена эта линия, мы должны учитывать не только неопределенность положения самой линии регрессии (о чем говорилось в предыдущем параграфе), но и дисперсию случайных отклонений от нее (ошибок МНК).

Дисперсия прогноза

Дисперсию случайной величины $y = f + e$ в произвольной точке x можно выразить следующим образом:

$$\overline{\sigma_{f+e}}^2 = \overline{\sigma_f}^2 + \overline{\sigma_e}^2$$

Используя полученную в предыдущем параграфе формулу для дисперсии линии регрессии получаем:

$$\overline{\sigma_{f+e}}^2 = \overline{\sigma_e}^2 \cdot \left(1 + \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_{k=1}^N (x_k - \bar{X})^2} \right)$$

Доверительный интервал прогноза

Так как математическое ожидание ошибки МНК e равно нулю, то доверительный интервал для прогнозного значения отклика Y в точке x определяется неравенствами:

$$\bar{f} - t_{1-q/2, v} \overline{\sigma_{f+e}} \leq y \leq \bar{f} + t_{1-q/2, v} \overline{\sigma_{f+e}}$$

Назовем величину $\Delta y = 2t_{1-q/2, \nu} \overline{\sigma_{f+e}}$ шириной полосы неопределенности прогноза.

Горизонт прогнозирования

Ширина полосы неопределенности прогноза минимальна при $x = \bar{X}$ и возрастает при увеличении абсолютной величины отклонения переменной x от \bar{X} . Точность прогноза определяется шириной полосы неопределенности.

Пусть мы априорно задаем максимально возможную ширину неопределенности прогноза Δy_{\max} и считаем, что точность прогноза является удовлетворительной, если в точке прогноза $\Delta y \leq \Delta y_{\max}$. При удалении от поля точек, по которым была получена линия регрессии, Δy обязательно достигнет Δy_{\max} . Соответствующее удаление называется *горизонтом прогнозирования*. Дальнейшее удаление приведет к тому, что Δy превысит Δy_{\max} . Интервал значений x , в пределах которого точность прогноза является удовлетворительной, выражается неравенством:

$$|x - \bar{X}| \leq x_{\max}$$

$$\text{где } x_{\max} = \sqrt{\left[\left(\frac{\Delta y_{\max}}{2t_{1-q/2, \nu} \sigma_e} \right)^2 - 1 - \frac{1}{N} \sum_{k=1}^N (x_k - \bar{X})^2 \right]}$$

8.13. Проверка допущений МНК.

Изучая уравнение линейной регрессии мы предполагали, что реальная взаимосвязь фактора X и отклика Y линейна, а отклонения от прямой регрессии случайны, независимы между собой, имеют нулевое математическое ожидание и постоянную дисперсию. Если это не так, то статистический анализ параметров регрессии некорректен и оценки этих параметров не обладают свойствами несмещенности и состоятельности. Например, это может быть, если в действительности связь между переменными нелинейна. Поэтому после получения уравнения регрессии необходимо исследовать его ошибки.

Ошибки метода наименьших квадратов, то есть величины $e_k = y_k - f_k$ должны обладать следующими свойствами:

- 1) Ошибки должны являться реализацией нормально распределенной случайной переменной.
- 2) Математическое ожидание ошибки должно быть равно нулю: $M(e_k) = 0$.
- 3) Дисперсия ошибки должна быть постоянна: $D(e_k) = \sigma^2$.
- 4) Ошибки должны быть независимыми, то есть

$$\text{cov}(e_k, e_j) = \begin{cases} 0 & k \neq j \\ \sigma^2 & k = j \end{cases}$$

После того, как получено уравнение регрессии $y = \bar{a}x + \bar{b} + e$, каждое из этих допущений должно быть проверено.

Проверка гипотезы о том, что ошибки нормально распределены

Идентификация закона распределения случайной величины изучена в главе 6, поэтому здесь мы не будем подробно рассматривать этот вопрос. Кратко можно сказать, что проверка гипотезы о том, что ошибки МНК нормально распределены, проводится в два этапа:

- 1) По выборке (e_1, e_2, \dots, e_N) строится гистограмма распределения случайной величины e .
- 2) Полученная гистограмма проверяется на соответствие нормальному распределению с помощью критерия согласия Пирсона.

Проверка гипотезы о том, что математическое ожидание ошибки равно нулю

Пусть ошибка МНК e имеет математическое ожидание μ_e и генеральную дисперсию σ_e^2 . Состоятельными и несмещенными оценками математического ожидания и дисперсии ошибки будут выборочная средняя и выборочная дисперсия:

$$\bar{e} = \frac{1}{N} \sum_{k=1}^N (y_k - \bar{a}x_k - \bar{b})$$

$$\overline{\sigma_e^2} = \frac{1}{N-2} \sum_{k=1}^N (y_k - \bar{a}x_k - \bar{b})^2$$

Мы должны проверить гипотезу

$$H_0: \bar{e} = 0$$

$$H_1: \bar{e} \neq 0$$

Проверка этой гипотезы осуществляется по следующей схеме:

1) *Априорные предположения*

Математическое ожидание ошибки равно нулю

$$\mu_e = 0$$

2) *Результаты испытания*

Выборочная средняя ошибки и выборочное с.к.о. ошибки

$$\bar{e}, \overline{\sigma_e}$$

при объеме выборки N .

3) *Гипотеза*

$$H_0: \bar{e} = 0$$

$$H_1: \bar{e} \neq 0$$

4) *Принятая величина уровня значимости*

$$q = 0.05 \text{ или } q = 0.01$$

5) *Критерий проверки*

$$t = \frac{\bar{e} - \mu_e}{\overline{\sigma_e}} = \frac{\bar{e}}{\overline{\sigma_e}}$$

6) *Правило принятия решения*

Принять H_0 , если $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $|t| > t_{1-q/2, \nu}$.

7) *Проверка гипотезы*

- Если $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$ то критерий проверки t не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости выборочная средняя ошибки \bar{e} статистически незначимо отличается от нуля.

- В противном случае мы принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости в уравнении регрессии присутствует систематическая ошибка, и это уравнение должно быть уточнено.

Проверка гипотезы о том, что дисперсия ошибки постоянна

Упорядочим исходную выборку (x_k, y_k) , $k = 1, \dots, N$ по возрастанию величины x . Обозначим как $N_{1/2}$ половину от объема выборки, то есть $N_{1/2} = \text{ЦЕЛОЕ}(N/2)$. Выберем число $M \leq N_{1/2}$. После этого по упорядоченной по возрастанию величины x выборке рассчитаем отклонения от линии регрессии, первое для $k = 1, \dots, M$ (для меньших значений x), второе для $k = N - M + 1, \dots, N$ (для больших значений x). Для лучшего разграничения между двумя группами наблюдений число M можно выбрать таким образом, чтобы исключить до 20% срединных точек.

В случае постоянства дисперсии ошибок МНК необъясненная дисперсия для меньших значений x должна быть приблизительно равна необъясненной дисперсии для больших значений x , то есть должно быть справедливым следующее равенство:

$$\sum_{k=1}^M e_k^2 \approx \sum_{k=N-M+1}^N e_k^2$$

Обозначим большую из этих сумм как S_1^2 , а меньшую как S_2^2 . Чем ближе к единице отношение S_1^2 / S_2^2 , тем больше оснований рассчитывать на то, что дисперсия ошибок МНК постоянна. Случайная величина $F = S_1^2 / S_2^2$ подчиняется F -распределению Фишера с $\nu_1 = M - 2$, $\nu_2 = M - 2$ степенями свободы. Проверка гипотезы о постоянстве дисперсии ошибок осуществляется по следующей схеме:

1) Гипотеза

$$H_0 : S_1^2 = S_2^2$$

$$H_1 : S_1^2 > S_2^2$$

2) Принятая величина уровня значимости

$$q = 0.05 \text{ или } q = 0.01$$

3) *Критерий проверки*

$$F = \frac{S_1^2}{S_2^2}$$

4) *Правило принятия решения*

Принять H_0 , если $F \leq F_{1-q, \nu_1, \nu_2}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки F попадает в критическую область $F > F_{1-q, \nu_1, \nu_2}$.

5) *Проверка гипотезы*

- Если $F \leq F_{1-q, \nu_1, \nu_2}$, то критерий проверки F не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости дисперсия ошибок уравнения регрессии постоянна.
- В противном случае мы принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости уравнении регрессии не является наилучшим приближением исходных данных.

Непостоянство дисперсии ошибок МНК возникает как правило в том случае, если неправильно выбран вид математической модели зависимости фактора X и отклика Y . Например, если нелинейную зависимость пытаются аппроксимировать линейной функцией.

Проверка гипотезы о том, что ошибки независимы

Одним из предполагаемых свойств уравнения регрессии $y = \bar{a}x + \bar{b} + e$ является то, что ошибки e независимы между собой. На практике проверяется не независимость, а некоррелированность этих величин, которая является необходимым, но недостаточным признаком независимости. При этом проверяется некоррелированность не любых, а соседних величин ошибок, которые можно получить, если исходная выборка $(x_k, y_k) k = 1, \dots, N$ упорядочена по возрастанию величины x .

Рассмотрим например корреляцию ошибок, сдвинутых друг относительно друга на один шаг.

$$e_1 \quad e_2 \quad e_3 \quad \dots \quad e_k \quad \dots \quad e_N$$

$$e_1 \quad e_2 \quad \dots \quad e_{k-1} \quad \dots \quad e_{N-1} \quad e_N$$

Тогда значение выборочного коэффициента корреляции между выборкой (e_2, e_3, \dots, e_N) и выборкой $(e_1, e_2, \dots, e_{N-1})$ запишется в виде:

$$\rho_{k,k+1} = \frac{\sum_{k=1}^{N-1} (e_k - \bar{e})(e_{k+1} - \bar{e})}{\sqrt{\sum_{k=1}^{N-1} (e_k - \bar{e})^2} \sqrt{\sum_{k=1}^{N-1} (e_{k+1} - \bar{e})^2}}$$

Эту величину называют еще *коэффициентом автокорреляции первого порядка*. Так как согласно допущениям МНК математическое ожидание ошибки равно нулю, то формулу можно упростить:

$$\rho_{k,k+1} = \frac{\sum_{k=1}^{N-1} e_k e_{k+1}}{\sqrt{\sum_{k=1}^{N-1} e_k^2} \sqrt{\sum_{k=1}^{N-1} e_{k+1}^2}}$$

Мы можем считать, что автокорреляция отсутствует, если выборочный коэффициент автокорреляции незначимо отличается от нуля, то есть в данном случае мы должны проверить гипотезу:

$$H_0 : \rho_{k,k+1} = 0$$

$$H_1 : \rho_{k,k+1} \neq 0$$

В случае однофакторной линейной регрессии случайная величина $t = \sqrt{N-3} \cdot \frac{\rho_{k,k+1}}{\sqrt{1-\rho_{k,k+1}^2}}$ будет подчиняться

распределению Стьюдента с $\nu = (N-1) - 2$ степенями свободы.

Поэтому гипотеза будет проверяться следующим образом:

1) *Гипотеза*

$$H_0 : \rho_{k,k+1} = 0$$

$$H_1 : \rho_{k,k+1} \neq 0$$

2) *Принятая величина уровня значимости*

$$q = 0.05 \text{ или } q = 0.01$$

3) *Критерий проверки*

$$t = \sqrt{N-3} \cdot \frac{\rho_{k,k+1}}{\sqrt{1-\rho_{k,k+1}^2}}$$

4) *Правило принятия решения*

Принять H_0 , если $-t_{1-q/2, v} \leq t \leq t_{1-q/2, v}$.

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $|t| > t_{1-q/2, v}$.

5) *Проверка гипотезы*

- Если $-t_{1-q/2, v} \leq t \leq t_{1-q/2, v}$, то критерий проверки t не попадает в критическую область и мы принимаем гипотезу H_0 . Это означает, что при заданном уровне значимости выборочный коэффициент автокорреляции первого порядка $\rho_{k,k+1}$ статистически незначимо отличается от нуля. Следовательно, автокорреляция первого порядка ошибок МНК отсутствует.
- В противном случае мы принимаем гипотезу H_1 . Это может означать, что нужно принять другую аналитическую модель зависимости между переменными X и Y .

8.14. Сведение нелинейной функциональной зависимости к линейной путем преобразования данных.

До сих пор мы обсуждали линейную зависимость между фактором X и откликом Y . Когда истинная взаимосвязь между ними носит нелинейный характер, в ряде случаев ее можно свести к линейной путем соответствующего преобразования данных. После этого к преобразованным данным может быть применена линейная регрессия. Преобразованные переменные и параметры мы будем отмечать символом \cap (например \bar{x}).

В этом параграфе мы рассмотрим несколько наиболее употребительных видов нелинейной зависимости.

1) Экспоненциальная функция $y = be^{ax}$

Экспоненциальная функция используется, когда при увеличении фактора X отклик Y растет ($a > 0$) или снижается ($a < 0$) с постоянной относительной скоростью.

Приведение к линейной зависимости $\hat{y} = a\hat{x} + \hat{b}$ осуществляется путем следующего преобразования данных:

$$\hat{y} = \ln(y) \quad \hat{x} = x \quad \hat{a} = a \quad \hat{b} = \ln(b)$$
2) Логарифмическая функция $y = b + a \ln(x)$

Логарифмическая функция используется, когда при увеличении фактора X отклик Y растет ($a > 0$) или снижается ($a < 0$) с уменьшающейся скоростью при отсутствии предельно возможного значения. Преобразование данных:

$$\hat{y} = y \quad \hat{x} = \ln(x) \quad \hat{a} = a \quad \hat{b} = b$$

3) Степенная функция $y = bx^a$

Степенная функция используется когда при увеличении фактора X отклик Y растет или снижается с разной мерой пропорциональности. Преобразование данных:

$$\hat{y} = \ln(y) \quad \hat{x} = \ln(x) \quad \hat{a} = a \quad \hat{b} = \ln(b)$$

4) Логистическая функция $y = \frac{1}{1 + e^{(x-b)/a}}$

Логистическая кривая имеет вид положенной на бок латинской буквы S . Она описывает случай когда при увеличении фактора X отклик Y изменяется (снижается при $a > 0$ или растет при $a < 0$) в пределах от 0 до 1. При этом изменения происходят при $x < b$ с увеличивающейся скоростью и при $x > b$ с уменьшающейся скоростью. Преобразование данных:

$$\hat{y} = \ln(1/y - 1) \quad \hat{x} = x \quad \hat{a} = 1/a \quad \hat{b} = -b/a$$

5) Гиперболическая функция $y = c + \frac{a}{x+b}$

Во многих случаях для аппроксимации нелинейной зависимости очень удобно использовать гиперболу, однако зачастую об этом трудно догадаться. Дело в том, что мы легко узнаем только простую гиперболу, асимптотами которой являются оси координат, то есть $y = a/x$. Если эта гипербола сдвинута вдоль одной из осей или вдоль обеих осей, то ее как правило не узнают.

Проверка того, является ли данная кривая гиперболой со сдвигом только вдоль оси x , то есть $y = a/(x+b)$, проводится путем следующего преобразования данных:

$$\hat{y} = 1/y \quad \hat{x} = x \quad \hat{a} = 1/a \quad \hat{b} = b/a$$

Проверка того, является ли данная кривая гиперболой со сдвигом только вдоль оси y , то есть $y = c + a/x$, проводится путем преобразования данных:

$$\hat{y} = y \quad \hat{x} = 1/x \quad \hat{a} = a \quad \hat{b} = c$$

Особенно сложным является случай, когда гипербола сдвинута одновременно по обеим осям, то есть имеет вид

$$y = c + \frac{a}{x+b}.$$

В этом случае нужно двигаться методом по-

следовательных приближений, то есть

- задавать ряд значений параметра b ,
- вычислять значения $1/(x+b)$,
- строить графики, где по оси абсцисс откладывать $1/(x+b)$, по оси ординат y ,
- выбрать то значение параметра b , при котором график наиболее близок к прямой линии.

8.15. Функция регрессии как комбинация нескольких функций.

На практике может оказаться, что функцию регрессии невозможно описать удовлетворительным образом ни линейной зависимостью, ни любой из перечисленных в предыдущем параграфе нелинейных функций. Тогда стоит попытаться аппроксимировать ее комбинацией этих функций. Делается это следующим образом:

- В общем случае считаем, что зависимость между фактором X и откликом Y нелинейна. Тогда, используя результаты из предыдущего параграфа, преобразуем исходную выборку $(x_k, y_k), k = 1, \dots, N$ таким образом, чтобы в первом приближении можно было считать, что связь между преобразованными данными $(\bar{x}_k, \bar{y}_k), k = 1, \dots, N$ носит линейный характер.
- Вычисляем параметры линейной регрессии.
- Вычисляем ошибки МНК $e_k, k = 1, \dots, N$.
- Проверяем свойства ошибок МНК. Если ошибки не удовлетворяют допущениям МНК, то полученная аппроксимация является слишком грубой.
- Дальнейшее уточнение модели можно сделать, если в качестве зависимой переменной использовать полученные ошибки, то есть выборка приобретает вид $(\bar{x}_k, e_k), k = 1, \dots, N$. Эту выборку необходимо обработать по той же схеме. Процесс продолжается до тех пор, пока на определенном шаге ошибки не станут удовлетворять допущениям МНК. При этом надо помнить, что нельзя излишне переусложнять модель, и что полученные по модели результаты должны разумным образом интерпретироваться.

9. АНАЛИЗ ФУРЬЕ

9.1. Введение.

В этой главе излагается метод аппроксимации эмпирической зависимости тригонометрическим рядом Фурье. Даны формулы, позволяющие по реальной выборке вычислить коэффициенты Фурье, амплитуду и фазу гармоник. Рассказано, как строится амплитудно-частотная характеристика разложения, и как она используется для выделения гармоник с максимальной амплитудой.

9.2. Численный анализ Фурье.

Пусть выборка значений фактора X и отклика Y задана в виде массива $(x_n, y_n), n = 0, \dots, N$, содержащего $N+1$ точку, причем все значения фактора X упорядочены по возрастанию и равноотстоят друг от друга. Будем считать, что величина X изменяется в интервале $(0, X_{\max})$, следовательно выборка фактора X задается рядом $x_n = X_{\max} \cdot n / N$.

Если принято решение о том, что связь переменных X и Y носит периодический характер, то аппроксимировать зависимость Y от X на интервале $(0, X_{\max})$ необходимо тригонометрическим рядом, то есть функцией вида:

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^M \left[a_m \cos\left(m \frac{2\pi x}{X_{\max}}\right) + b_m \sin\left(m \frac{2\pi x}{X_{\max}}\right) \right]$$

Данная функция зависит от $(2M+1)$ параметра $(a_0, a_1, \dots, a_M, b_1, \dots, b_M)$. Так как количество неизвестных параметров $2M+1$ не должно превышать объем выборки $N+1$, то $M \leq N/2$.

Наилучшим приближением будет тригонометрический ряд с таким набором параметров, который минимизирует сумму квадратов отклонений этого ряда от выборочных значений отклика Y , то есть

$$S = \sum_{n=1}^N [y_n - f(x_n)]^2 \rightarrow \min$$

Без доказательства приведем формулы для определения искомых параметров:

$$a_m = \frac{2}{N} \sum_{n=0}^{N-1} y_n \cos\left(m \frac{2\pi n}{N}\right) \quad 0 \leq m \leq M$$

$$b_m = \frac{2}{N} \sum_{n=0}^{N-1} y_n \sin\left(m \frac{2\pi n}{N}\right) \quad 1 \leq m \leq M$$

Определенные по этим формулам параметры называют *коэффициентами Фурье*, а тригонометрический ряд с такими коэффициентами является *рядом Фурье*. Тогда аппроксимация величины Y рядом Фурье в точке $x_n = X_{\max} n / N$ будет равна:

$$f_n = \frac{a_0}{2} + \sum_{m=1}^M \left[a_m \cos\left(m \frac{2\pi n}{N}\right) + b_m \sin\left(m \frac{2\pi n}{N}\right) \right]$$

При увеличении количества гармоник M эта аппроксимация все точнее описывает выборочные значения величины Y , и наконец при $M = N/2$ для любого n становится справедливым равенство $y_n = f_n$.

Однако, наша задача состоит не в том, чтобы с абсолютной точностью аппроксимировать исходную выборку, то есть включить в математическую модель все наблюдающиеся особенности конкретной выборки, в том числе и те, которые в действительности носят случайный характер. Нам нужно найти всего несколько наиболее значимых гармоник, то есть гармоник, имеющих максимальную амплитуду. Для этого необходимо построить и проанализировать амплитудно-частотную характеристику разложения.

9.3. Амплитудно-частотная характеристика.

Введем параметры (R_m, θ_m) , которые назовем *амплитуда* и *фаза* соответственно. Эти величины связаны с параметрами (a_m, b_m) следующими соотношениями:

$$R_m = \sqrt{a_m^2 + b_m^2} \quad \theta_m = -\arctg \frac{b_m}{a_m} \quad -\pi < \theta_m \leq \pi$$

$$a_m = R_m \cos \theta_m \quad b_m = -R_m \sin \theta_m$$

Тогда, заменив параметры (a_m, b_m) , разложение Фурье можно переписать в виде

$$f_n = \frac{a_0}{2} + \sum_{m=1}^M \left[R_m \cos \theta_m \cos \left(m \frac{2\pi n}{N} \right) - R_m \sin \theta_m \sin \left(m \frac{2\pi n}{N} \right) \right]$$

$$f_n = \frac{a_0}{2} + \sum_{m=1}^M R_m \cos \left(m \frac{2\pi n}{N} + \theta_m \right)$$

Назовем *частотой* колебаний величину $\omega_m = m/N$. Полный набор частот называется *спектром* разложения. Тогда окончательно получаем

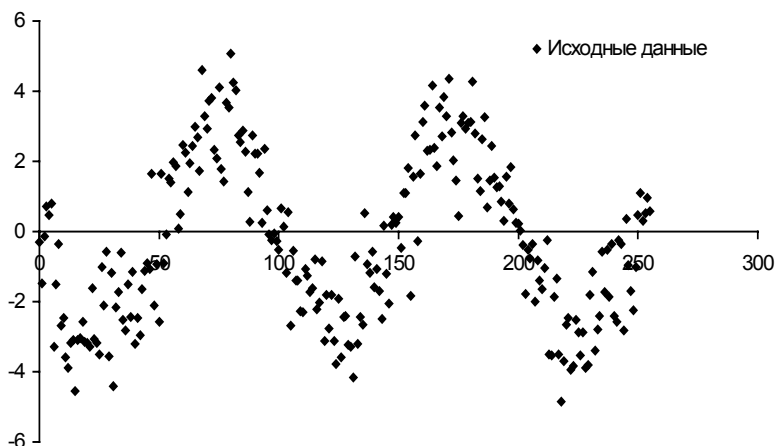
$$f_n = \frac{a_0}{2} + \sum_{m=1}^M R_m \cos(2\pi\omega_m n + \theta_m)$$

Смысл приведенных выше преобразований состоит в том, чтобы перейти от ряда из синусов и косинусов к ряду из одних косинусов. Если теперь построить график, где по оси абсцисс отложена частота, а по оси ординат отложена амплитуда, то есть график в координатах (ω_m, R_m) , то наглядно будет видно, при каких значениях частоты наблюдаются максимумы амплитуды. Такой график называется *амплитудно-частотной характеристикой (АЧХ)*. С помощью АЧХ мы получаем возможность выбрать из разложения Фурье только самые значимые гармоники и пренебречь остальными. Заметим, что период колебания связан с частотой соотношением $T_m = 1/\omega_m$.

При необходимости аналогичным образом можно построить *фазочастотную характеристику (ФЧХ)*, то есть график в координатах (ω_m, θ_m) .

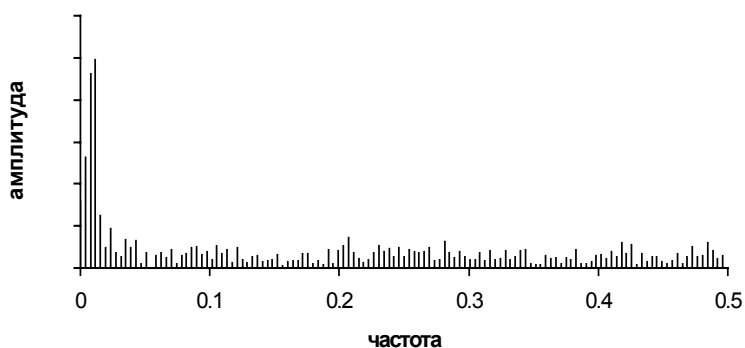
9.4. Пример выделения основной гармоники с помощью анализа Фурье.

Рассмотрим выделение основной гармоники с помощью анализа Фурье на примере выборки, состоящей из 256-ти точек $(x_n, y_n), n = 0, \dots, 255$. График исходных данных приведен на рисунке.



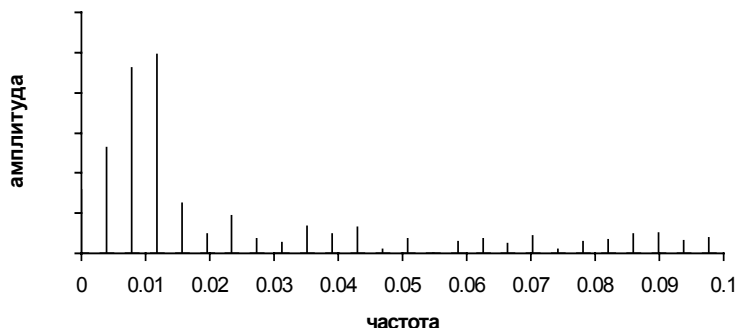
Этот график дает основания предположить, что связь переменных X и Y носит периодический характер. По методике, изложенной в предыдущих 2-х параграфах, представим аппроксимирующую функцию рядом Фурье и построим амплитудно-частотную характеристику.

АЧХ



Максимум амплитуды находится в начальной части спектра. Рассмотрим подробнее этот участок.

АЧХ



При ближайшем рассмотрении оказывается, что максимум амплитуды приходится на частоту $\omega \approx 0.01$ (период $T \approx 100$). Учитывая, что $\omega_m = m / N$, рассчитаем для этого значения частоты коэффициенты разложения Фурье:

$$a_m = a(\omega_m) = -0.1049$$

$$b_m = b(\omega_m) = -2.8918$$

Используя эти данные, вычисляем амплитуду и фазу основной гармоники:

$$R_m = 2.8934$$

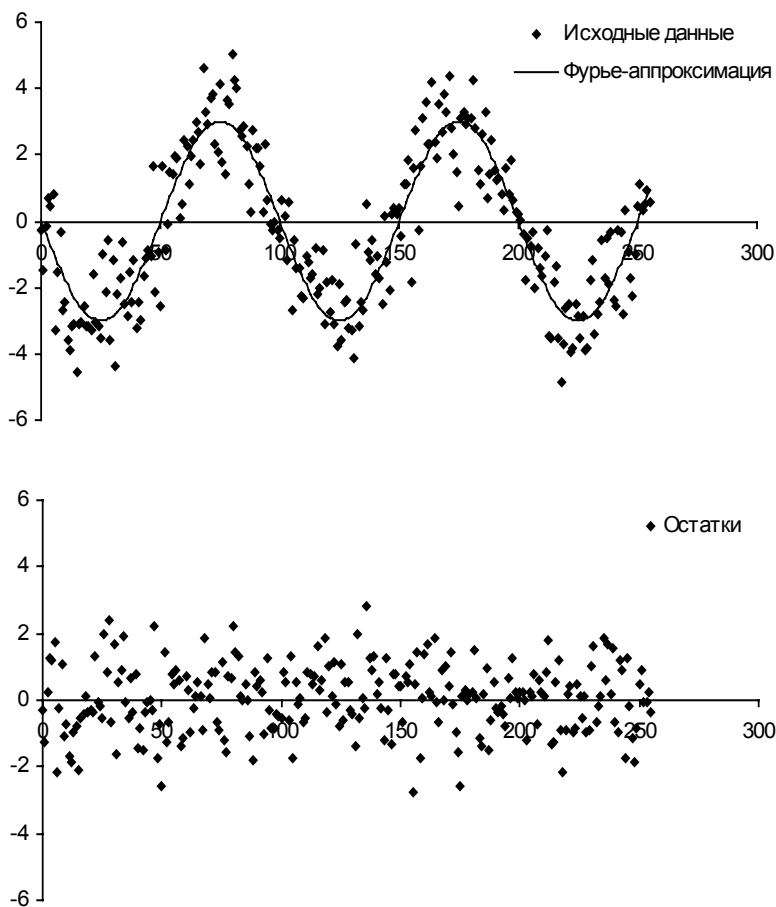
$$\theta_m = 1.6070$$

Таким образом, Фурье-аппроксимация исходных данных и ошибки модели будут вычисляться по формулам

$$f_n = R_m \cos(2\pi\omega_m n + \theta_m) = 2.8934 \cdot \cos(2\pi \cdot 0.01 \cdot n + 1.607)$$

$$e_n = y_n - f_n$$

Приведем график исходных данных вместе с Фурье-аппроксимацией и график остатков (ошибок модели).



Очевидно, что ошибки аппроксимации носят непериодический характер. В противном случае нужно было бы повторить всю процедуру, используя в качестве исходной выборки эти ошибки, и повторять ее до тех пор, пока не будут выделены все значимые гармоники.

На практике, при изучении динамики цен активов не рекомендуется использовать для аппроксимации этих рядов более трех гармоник Фурье.

10. ПРИМЕНЕНИЕ МНК ПРИ ИЗУЧЕНИИ ДИНАМИЧЕСКИХ РЯДОВ

10.1. Введение.

Аналитическая аппроксимация динамического временного ряда, содержащего цены некоторого актива в последовательные моменты времени, представляет собой математическую модель развития во времени этого динамического ряда и описывает присущие ему статистические характеристики.

Аналитическая аппроксимация содержит в себе некоторую условность, связанную с тем, что цена актива рассматривается как функция времени. На самом деле цена зависит не от того, сколько времени прошло с начального момента, а от того, какие факторы на нее влияли, в каком направлении и с какой интенсивностью они действовали. Зависимость от времени можно рассматривать как внешнее выражение суммарного воздействия этих факторов. Удовлетворительным образом аппроксимировать динамический ряд с помощью метода наименьших квадратов возможно лишь тогда, когда воздействие всех влияющих факторов однородно на всем рассматриваемом промежутке времени.

В случае, если динамический ряд цены актива удастся аппроксимировать аналитической функцией времени с соблюдением допущений МНК, становится возможной экстраполяция этой функции, то есть прогноз цены в будущие моменты времени. Однако при этом стоит помнить, что при прогнозе неявным образом предполагается, что те же самые условия, в которых формировались цены в прошлом, будут существовать и в будущем. Использование экстраполяции в изменившихся условиях будет приводить к ошибкам, выходящим за рамки обычных для МНК погрешностей, связанных с шириной полосы неопределенности линии регрессии. Долгосрочные прогнозы сопряжены с большими ошибками, чем краткосрочные. Во-первых, это связано с расширением полосы неопределенности линии регрессии при удалении от центра тяжести эмпирических данных, по которым эта линия была получена. Во-вторых, это связано с возрастанием влияния новых факторов при увеличении периода прогноза.

Для того, чтобы динамический ряд можно было эффективным образом аппроксимировать с применением МНК, этот ряд должен удовлетворять следующим условиям:

- быть достаточно длинным,
- быть как можно менее волатильным.

При этом следует сказать, что применение МНК при изучении временных рядов имеет следующие особенности:

- для адаптирования регрессионной модели к изменяющимся условиям необходимо периодически пересчитывать параметры модели с учетом новых данных, а иногда возможно пересматривать саму модель,
- при расчете параметров регрессии все эмпирические данные входят с одинаковым весом, хотя интуитивно понятно, что более поздние данные имеют большую ценность.

10.2. Модель динамики цен активов.

Биржевые цены активов формируются как результат совместных действий большого количества участников рынка и, как следствие этого, в них присутствует случайная составляющая.

Рассмотрим временной ряд, состоящий из последовательных значений цены некоторого актива P_1, P_2, \dots, P_t . Цена не может быть отрицательной, но может принимать сколь угодно большие положительные значения. Следовательно, и отношение цен в последовательные моменты времени P_k / P_{k-1} также не может оказаться ниже нуля, но может быть сколь угодно большим. Значит плотность вероятности цен активов и плотность вероятности отношения цен должны иметь положительную асимметрию.

Ситуация меняется при переходе к логарифмам отношения цен, то есть к величине $\Delta y_k = \ln(P_k / P_{k-1})$. Распределение логарифмов уже может быть симметрично и возможна его аппроксимация одним из аналитических законов распределения, которые были рассмотрены во второй главе (как правило обобщенным экспоненциальным распределением). При этом логарифм цены в произвольный момент времени складывается из логарифма цены в начальный момент времени (эта величина пред

полагается нестохастической) и суммы логарифмов отношения цен:

$$\ln(P_t) = \ln(P_0) + \sum_{k=1}^t \ln(P_k / P_{k-1})$$

Если величины $\Delta y_k = \ln(P_k / P_{k-1})$ независимы и имеют конечную дисперсию, то согласно центральной предельной теореме величина $y_t = \ln(P_t)$ будет нормально распределена при любом законе распределения Δy_k . Так как логарифм цены распределен нормально, то цена подчиняется логнормальному распределению.

Итак, если все случайные величины Δy_k независимы и подчиняются одному и тому же закону распределения с математическим ожиданием μ и дисперсией σ^2 , то случайная величина $\ln(P_t)$ будет иметь нормальное распределение с математическим ожиданием μt и дисперсией $\sigma^2 t$. Следовательно, логарифм цены в произвольный момент времени можно записать как $\ln(P_t) = \ln(P_0) + \mu t + \sigma \sqrt{t} z$

где случайная величина z подчиняется стандартному нормальному распределению.

Рисковые активы имеют положительное математическое ожидание дохода, следовательно $\mu > 0$. Величина μ определяет тренд актива, то есть воздействие на цену постоянно действующих систематических факторов.

Величина σ определяет волатильность актива, то есть воздействие на цену множества случайных факторов.

Отношение ожидаемого дохода к ожидаемому риску за единицу времени μ / σ характеризует степень устойчивости роста цены актива. Чем выше это отношение, тем привлекательнее при прочих равных условиях инвестиции в данный актив.

Наряду с влиянием постоянно действующих факторов и случайных колебаний, цена актива может испытывать воздействие причин, характеризующихся циклическими колебаниями. Возникновение циклов связано с изменением оценки инвесто

рами ожидаемого дохода актива. С учетом периодических компонент модель динамики цены можно представить в виде

$$\ln(P_t) = \ln(P_0) + \mu t + \sum_{m=1}^M R_m \cos(2\pi t / T_m + \theta_m) + \sigma \sqrt{t} z$$

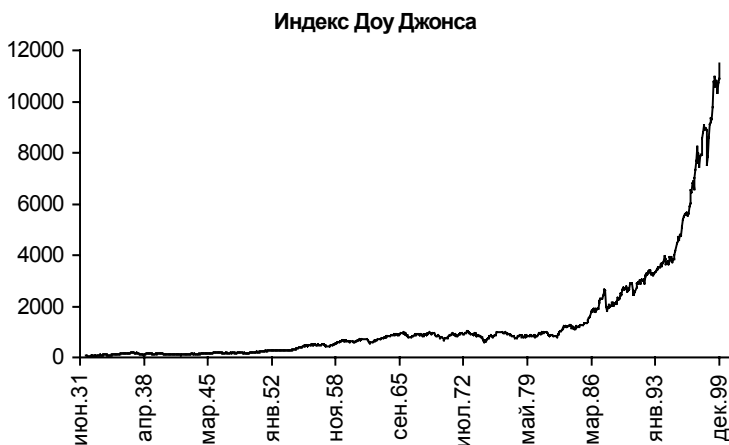
где T_m - период колебания, R_m - амплитуда колебания, θ_m - начальная фаза. Существует эмпирическое правило, которое называют принципом пропорциональности, согласно которому амплитуды колебаний прямо пропорциональны их периодам. Для выделения отдельных гармоник из временного ряда цены актива используют анализ Фурье.

С учетом вышесказанного, исследование динамики цены актива должно включать в себя следующие этапы:

- определение тренда,
- определение циклических компонент,
- составление прогноза цены актива.

10.3. Определение тренда.

В качестве исходных данных рассмотрим цены закрытия по индексу Доу Джонса на последний торговый день месяца за период с 1932 по 1999 год.



Рассмотрим тот же график в полулогарифмическом масштабе.



Полулогарифмический график дает основания полагать, что тренд логарифма цены закрытия можно в первом приближении описать линейной функцией времени.

Для построения регрессионной модели в качестве фактора (независимой переменной) будем использовать номер месяца. При этом первый месяц в выборке (январь 1932 года) получает номер 0, последний месяц в выборке (декабрь 1999 года) получает номер 815, то есть $t_k = 0, \dots, 815$. Объем выборки $N = 816$ точек.

Откликом (зависимой переменной) является логарифм цены закрытия $y_k = \ln(P_k)$. Эмпирическая зависимость отклика от фактора приведена на рисунке:



Оценка параметров линейной регрессии

Примем гипотезу о том, что связь фактора и отклика выражается линейной функцией $f(t) = at + b$. Оценки параметров линейной регрессии проводятся по формулам:

$$a = \frac{\sum_{k=1}^N t_k y_k - N \cdot \bar{T} \cdot \bar{Y}}{\sum_{k=1}^N t_k^2 - N \cdot \bar{T}^2} \quad b = \bar{Y} - a \cdot \bar{T}$$

где

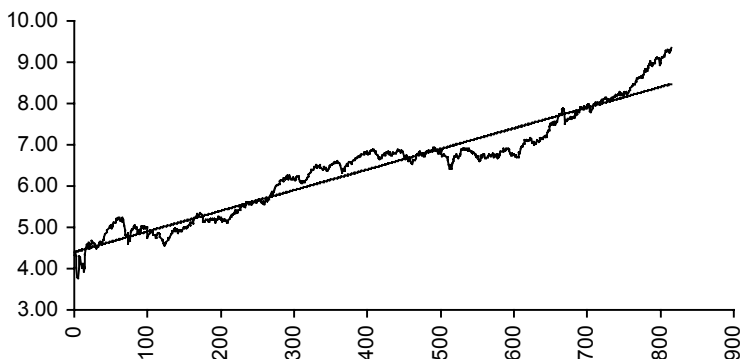
$$\bar{T} = \frac{1}{N} \sum_{k=1}^N t_k \quad \bar{Y} = \frac{1}{N} \sum_{k=1}^N y_k$$

Вычисленные значения параметров составляют:

$$a = 0.005, b = 4.402$$

Эмпирическая зависимость и линейная аппроксимация изображены на рисунке:

Эмпирическая зависимость и линейная аппроксимация



При этом график ошибок аппроксимации

$$e_k = y_k - at_k - b = y_k - 0.005t_k - 4.402$$

имеет вид:



Дисперсия оценок параметров линейной регрессии

Оценка дисперсии случайных отклонений отклика Y от линии регрессии (необъясненная дисперсия) вычисляется по формуле:

$$\sigma_e^2 = \frac{1}{N-2} \sum_{k=1}^N e_k^2 = \frac{1}{N-2} \sum_{k=1}^N (y_k - at_k - b)^2$$

Вычисленные значения необъясненной дисперсии и соответствующее с.к.о. равны:

$$\sigma_e^2 = 0.098 \quad \sigma_e = 0.313$$

Оценка дисперсии параметров a и b выражаются формулами:

$$\sigma_a^2 = \frac{\sigma_e^2}{\sum_{k=1}^N (x_k - \bar{X})^2} \quad \sigma_b^2 = \frac{\sigma_e^2}{\sum_{k=1}^N (x_k - \bar{X})^2} \cdot \frac{\sum_{k=1}^N x_k^2}{N}$$

Расчетные значения этих величин по выборке составляют:

$$\sigma_a^2 = 2.2 \cdot 10^{-9} \quad \sigma_a = 4.7 \cdot 10^{-5}$$

$$\sigma_b^2 = 4.8 \cdot 10^{-4} \quad \sigma_b = 2.2 \cdot 10^{-2}$$

Коэффициент детерминации

Качество линии регрессии характеризуется коэффициентом детерминации:

$$R^2 = 1 - \frac{\sum_{k=1}^N e_k^2}{\sum_{k=1}^N (y_k - \bar{Y})^2}$$

В рассматриваемом случае эта величина равна $R^2 = 0.9348$. Так как среднеквадратичные отклонения отклика Y и ошибок аппроксимации e связаны соотношением $\sigma_e = \sqrt{1 - R^2} \cdot \sigma_y$, то получаем, что с.к.о. ошибок приблизительно в четыре раза меньше с.к.о. отклика: $\sigma_e = 0.255\sigma_y$.

10.4. Статистические выводы о величине параметров регрессии.

Необходимо убедиться, что значения параметров регрессии значимо отличаются от нуля. Для проверки этого выдвигаются гипотезы:

$$H_0 : a = 0 \quad H_0 : b = 0$$

$$H_1 : a \neq 0 \quad H_1 : b \neq 0$$

1) Примем величину уровня значимости $q = 0.05$

2) Рассчитаем критерии проверки

$$t_a = \frac{a}{\sigma_a} = \frac{0.005}{4.7 \cdot 10^{-5}} = 106.4$$

$$t_b = \frac{b}{\sigma_b} = \frac{4.402}{2.2 \cdot 10^{-2}} = 200$$

3) Правило принятия решения

Принять H_0 , если $-t_{1-q/2, v} \leq t \leq t_{1-q/2, v}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки t попадает в критическую область $|t| > t_{1-q/2, v}$.

4) Расчет границ критической области

$$t_{1-q/2, \nu} = \text{СТБЮДРАСПОБР}(q, N - 2) = \\ = \text{СТБЮДРАСПОБР}(0.05, 814) = 1.96$$

5) *Проверка гипотезы*

Так как критерии проверки для обоих параметров регрессии находятся в критической области, мы принимаем гипотезу H_1 . Это означает, что при заданном уровне значимости параметры регрессии статистически значимо отличаются от нуля.

Статистические выводы о величине коэффициента детерминации

Убедимся в том, что коэффициент детерминации значимо отличается от нуля. Для проверки этого выдвигается гипотеза:

$$H_0 : R^2 = 0$$

$$H_1 : R^2 > 0$$

1) *Примем величину уровня значимости*

$$q = 0.05$$

2) *Рассчитаем критерий проверки*

$$F = \frac{R^2}{(1 - R^2)/(N - 2)} = \frac{0.9348}{(1 - 0.9348)/814} = 11671$$

3) *Правило принятия решения*

Принять H_0 , если $F \leq F_{1-q, \nu_1, \nu_2}$.

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки F попадает в критическую область $F > F_{1-q, \nu_1, \nu_2}$.

F_{1-q, ν_1, ν_2} - это квантиль F -распределения, соответствующая уровню значимости q с $\nu_1 = 1$ степенями свободы для числителя и $\nu_2 = N - 2$ степенями свободы для знаменателя.

4) *Расчет границ критической области*

$$F_{1-q, \nu_1, \nu_2} = \text{FRАСПОБР}(q, \nu_1, \nu_2) = \\ = \text{FRАСПОБР}(0.05, 1, 814) = 3.85$$

5) *Проверка гипотезы*

Так как критерий проверки для коэффициента детерминации находится в критической области, мы принимаем гипотезу H_1 .

Это означает, что при заданном уровне значимости изменения отклика y объясняются изменением фактора t .

10.5. Полоса неопределенности рассеяния эмпирических данных относительно линии регрессии.

Дисперсия случайной величины $y = f + e$ в произвольной точке t вычисляется по формуле:

$$\sigma_{f+e}^2 = \sigma_e^2 \cdot \left(1 + \frac{1}{N} + \frac{(t - \bar{T})^2}{\sum_{k=1}^N (t_k - \bar{T})^2} \right)$$

где

$$\sigma_e = 0.313 \quad N = 816 \quad \bar{T} = 407.5$$

$$\sum_{k=1}^N (t_k - \bar{T})^2 = 45\,278\,140$$

В данном случае на большом диапазоне изменения t без существенной потери точности вторым и третьим слагаемым в скобках можно пренебречь, то есть $\sigma_{f+e}^2 \approx \sigma_e^2$.

Величина $\Delta y = 2t_{1-q/2, \nu} \sigma_{f+e}$ называется шириной полосы неопределенности. Зададимся доверительной вероятностью $P = 0.95$ ($q = 0.05$). Тогда квантиль распределения Стьюдента равна

$$\begin{aligned} t_{1-q/2, \nu} &= \text{СТЮДРАСПОБР}(q, \nu) = \\ &= \text{СТЮДРАСПОБР}(0.05, 814) = 1.96 \end{aligned}$$

Ширина полосы неопределенности составит

$$\Delta y = 2 \cdot 1.96 \cdot 0.313 = 1.226$$

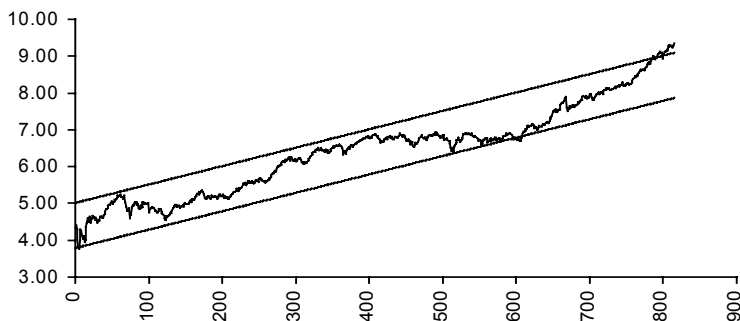
Следовательно, с вероятностью $P = 0.95$ случайная величина $y = f + e$ будет лежать в пределах:

$$f - \Delta y / 2 \leq y \leq f + \Delta y / 2$$

$$(0.005 t + 4.402) - 0.613 \leq y \leq (0.005 t + 4.402) + 0.613$$

Эмпирическая зависимость и ее полоса неопределенности изображены на рисунке:

Эмпирическая зависимость и полоса неопределенности



Приведем также график ошибок МНК и его полосу неопределенности:

Ошибки линейной аппроксимации и полоса неопределенности



Количество точек, находящееся внутри полосы неопределенности, равно 774, что составляет $(774/816) \cdot 100\% = 94.85\%$ от объема выборки. Это соответствует доверительной вероятности $P = 0.95$.

10.6. Проверка допущений МНК.

Для того, чтобы мы могли сказать, что модель адекватна эмпирическим данным, ошибки e должны обладать следующими свойствами:

- 1) Ошибки должны являться реализацией нормально распределенной случайной переменной.
- 2) Математическое ожидание ошибки должно быть равно нулю: $M(e_k) = 0$.

- 3) Дисперсия ошибки должна быть постоянна: $D(e_k) = \sigma^2$.
- 4) Ошибки должны быть независимыми, то есть
- $$\text{cov}(e_k, e_j) = \begin{cases} 0 & k \neq j \\ \sigma^2 & k = j \end{cases}$$

Проверка гипотезы о том, что ошибки нормально распределены

Оценки основных параметров распределения величины e приведены в таблице:

Наименование оценки	Величина
Центр распределения \bar{e}	0.01
Среднеквадратичное отклонение σ_e	0.313
Коэффициент асимметрии γ	0.063
С.к.о. коэффициента асимметрии σ_γ	0.085

Для проверки гипотезы о том, что ошибки нормально распределены, нам необходимо построить гистограмму выборочного распределения величины e .

Оптимальное число столбцов гистограммы можно найти, округлив вниз до ближайшего большего или равного пяти нечетного целого величину, определенную по формуле:

$$L = \frac{\varepsilon + 1.5}{6} N^{0.4}$$

Вычисленное значение $L = 9$. Таким образом, область изменения величины e разбивается на 9 интервалов, в каждом из которых необходимо рассчитать эмпирические частоты попадания в соответствующий интервал.

При использовании критерия согласия Пирсона необходимо вычислить величину:

$$\chi^2 = \sum_{i=1}^L \frac{(T_i - s_i)^2}{T_i}$$

где

L - количество столбцов гистограммы,

S_i - эмпирическая частота попадания в i -й столбец,

T_i - теоретическая частота попадания в i -й столбец.

Приведем таблицу эмпирических частот.

Номер интервала	Левая граница	Правая граница	Эмпирическая частота S_i
1	-0.8732	-0.6791	5
2	-0.6791	-0.4851	46
3	-0.4851	-0.2911	87
4	-0.2911	-0.0970	166
5	-0.0970	0.0970	206
6	0.0970	0.2911	139
7	0.2911	0.4851	124
8	0.4851	0.6791	22
9	0.6791	0.8732	21

Так как отношение коэффициента асимметрии к его среднеквадратичному отклонению меньше трех

$$\gamma / \sigma_\gamma = 0.063 / 0.085 = 0.738 < 3$$

то несимметричность носит случайный характер и распределение частот можно расчетным образом симметризовать относительно центрального пятого столбца:

Номер интервала	Левая граница	Правая граница	Эмпирическая частота S_i
1	-0.8732	-0.6791	13.00
2	-0.6791	-0.4851	34.00
3	-0.4851	-0.2911	105.50
4	-0.2911	-0.0970	152.50
5	-0.0970	0.0970	206.00
6	0.0970	0.2911	152.50
7	0.2911	0.4851	105.50
8	0.4851	0.6791	34.00
9	0.6791	0.8732	13.00

Вычислим теоретические частоты попадания в соответствующий интервал для нормального распределения с ($\mu = 0.01, \sigma = 0.313$) и рассчитаем величину χ^2 :

Номер интервала	Левая граница	Правая граница	Эмпирическая частота S_i	Теоретическая частота T_i	$\frac{(T_i - S_i)^2}{T_i}$
1	-0.8732	-0.6791	13.00	10.03	0.88
2	-0.6791	-0.4851	34.00	37.08	0.26
3	-0.4851	-0.2911	105.50	94.29	1.33
4	-0.2911	-0.0970	152.50	165.03	0.95
5	-0.0970	0.0970	206.00	198.88	0.25
6	0.0970	0.2911	152.50	165.03	0.95
7	0.2911	0.4851	105.50	94.29	1.33
8	0.4851	0.6791	34.00	37.08	0.26
9	0.6791	0.8732	13.00	10.03	0.88
ИТОГО					$\chi^2 = 7.09$

Зададимся уровнем значимости $q = 0.05$. Тогда с учетом того, что количество степеней свободы равно

$$\nu = L - 1 - r = 9 - 1 - 2 = 6$$

граница критической области вычисляется как:

$$\chi^2_{1-q, \nu} = \text{ХИ2ОБР}(0.05, 6) = 12.59$$

Так как $\chi^2 \leq \chi^2_{1-q, \nu}$, то распределение отклонений от линии регрессии можно аппроксимировать нормальным распределением при заданном уровне значимости.

Проверка гипотезы о том, что математическое ожидание ошибки равно нулю

Проверка гипотезы осуществляется по схеме:

1) Априорные предположения

Математическое ожидание ошибки равно нулю

$$\mu_e = 0$$

2) Результаты испытания

Выборочная средняя ошибки и выборочное с.к.о. ошибки

$$\bar{e} = 0.01$$

$$\sigma_e = 0.313$$

при объеме выборки $N = 816$.

3) Гипотеза

$$H_0: \bar{e} = 0$$

$$H_1: \bar{e} \neq 0$$

4) Принятая величина уровня значимости
 $q = 0.05$

5) Расчет критерия проверки

$$t = \frac{\bar{e} - \mu_e}{\sigma_e} = \frac{0.01}{0.313} = 0.032$$

6) Правило принятия решения

Принять H_0 , если $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$

В противном случае принять H_1 .

7) Расчет границ критической области

$$\begin{aligned} t_{1-q/2, \nu} &= \text{СТБЮДРАСПОБР}(q, N - 2) = \\ &= \text{СТБЮДРАСПОБР}(0.05, 814) = 1.96 \end{aligned}$$

8) Проверка гипотезы

Так как $-t_{1-q/2, \nu} \leq t \leq t_{1-q/2, \nu}$ то мы принимаем гипотезу H_0 , то есть при заданном уровне значимости выборочная средняя ошибки \bar{e} статистически незначимо отличается от нуля.

Проверка гипотезы о том, что дисперсия ошибки постоянна

Для проверки этой гипотезы разделим эмпирические данные на две группы по 350 точек: с 1-й по 350-ю и с 467-й по 816-ю точки. Серединные точки с 351-й по 466-ю (14.2% от объема выборки) исключаем для лучшего разграничения между группами. Рассчитаем суммы квадратов ошибок для каждой из этих групп:

$$S_1 = \sum_{k=467}^{816} e_k^2 = 50.37 \qquad S_2 = \sum_{k=1}^{350} e_k^2 = 19.67$$

Проверка гипотезы о постоянстве дисперсии осуществляется по схеме:

1) Гипотеза

$$H_0: S_1^2 = S_2^2$$

$$H_1: S_1^2 > S_2^2$$

2) Принятая величина уровня значимости

$$q = 0.01$$

3) Расчет критерия проверки

$$F = \frac{S_1^2}{S_2^2} = \frac{50.37}{19.67} = 2.561$$

4) Правило принятия решения

Принять H_0 , если $F \leq F_{1-q, v_1, v_2}$

В противном случае принять H_1 , то есть H_1 принимается, когда критерий проверки F попадает в критическую область $F > F_{1-q, v_1, v_2}$.

5) Расчет границ критической области

$$\begin{aligned} F_{1-q, v_1, v_2} &= F_{\text{РАСПОБР}}(q, v_1, v_2) = \\ &= F_{\text{РАСПОБР}}(0.01, 350 - 2, 350 - 2) = 1.284 \end{aligned}$$

6) Проверка гипотезы

Даже при уровне значимости $q = 0.01$ критерий проверки F попадает в критическую область $F > F_{1-q, v_1, v_2}$, то есть мы отклоняем гипотезу H_0 и принимаем гипотезу H_1 . Следовательно дисперсия ошибок регрессии не постоянна.

Проверка гипотезы о том, что ошибки независимы

На практике проверяется не независимость, а некоррелированность ошибок, которая является необходимым, но недостаточным условием независимости. Для этого нужно рассчитать коэффициент автокорреляции первого порядка

$$\rho_{k, k+1} = \frac{\sum_{k=1}^{N-1} e_k e_{k+1}}{\sqrt{\sum_{k=1}^{N-1} e_k^2} \sqrt{\sum_{k=1}^{N-1} e_{k+1}^2}}$$

Для рассматриваемого здесь случая эта величина равна $\rho_{k, k+1} = 0.987$. Очевидно, что коэффициент автокорреляции значительно отличается от нуля и ошибки уравнения высокоррелированы.

Выводы

Следует признать, что аппроксимация линейной функцией логарифма цены актива является неудовлетворительной, так как не соблюдаются два из четырех допущений МНК.

Не приводя доказательств скажем, что попытка уточнить модель путем введения циклических компонент не приводит к улучшению качества ошибок регрессии.

На практике при изучении динамических рядов цен активов используют *методы адаптивного моделирования*, о которых будет рассказано в следующих главах.

11. СГЛАЖИВАНИЕ ДИНАМИЧЕСКИХ РЯДОВ

11.1. Введение.

Целью сглаживания динамического ряда является фильтрация случайных колебаний уровней этого ряда и выявление наиболее устойчивой тенденции движения. Мы будем рассматривать методы сглаживания, базирующиеся на вычислении скользящих средних. Любое скользящее среднее - это метод определения среднего уровня динамического ряда за некоторый период времени. Термин "скользящее" подразумевает, что среднее значение каждый раз заново вычисляется в последовательные моменты времени. В этой главе под динамическим рядом мы, как правило, будем понимать ряд, состоящий из цен активов.

11.2. Типы скользящих средних.

В общем виде формула для вычисления любой скользящей средней (moving average) имеет вид:

$$MA = \sum_k w_k y_k$$

где $\{y_k\}$ - массив цен актива,

$\{w_k\}$ - массив весов, с которыми цены входят в формулу.

При этом для набора весов должно соблюдаться правило нормирования:

$$\sum_k w_k = 1$$

Скользящая средняя характеризуется:

- *объектом вычисления*, то есть тем динамическим рядом, который необходимо сгладить,
- *периодом* скользящей средней,
- *типом* скользящей средней, который определяет алгоритм вычисления набора весов $\{w_k\}$.

Различают три основных типа скользящих средних:

- простая скользящая средняя (SMA - simple moving average),
- взвешенная скользящая средняя (WMA - weighted moving average),
- экспоненциальная скользящая средняя (EMA - exponential moving average).

11.3. Простая скользящая средняя.

Простая скользящая средняя порядка T - это средняя арифметическая цен за период времени $[t - T + 1, t]$, то есть

$$SMA_t = \frac{1}{T} \sum_{k=t-T+1}^t y_k$$

Внутри интервала $t - T + 1 \leq k \leq t$ все веса, с которыми входят цены при расчете скользящей средней одинаковы и равны $w_k = 1/T$. За пределами этого интервала, то есть при $k < t - T + 1$ веса равны нулю.

Первым недостатком SMA является равенство весов в пределах интервала расчета, так как интуитивно понятно, что последние данные должны иметь большую ценность, то есть входить в формулу для расчета с большим весом.

Второй недостаток SMA становится понятным при рассмотрении рекуррентной формулы для ее вычисления:

$$SMA_t = SMA_{t-1} + \frac{1}{T} y_t - \frac{1}{T} y_{t-T}$$

Очевидно, что SMA на каждую цену реагирует дважды: первый раз, когда цена входит в интервал расчета, и второй раз, когда цена выбывает из него. Вторая реакция никак не связана с текущей динамикой и, следовательно, нежелательна.

Традиционно, скользящую среднюю соотносят с последней точкой интервала расчета, то есть с моментом времени t , хотя, строго говоря, это некорректно. Вычисленное значение SMA нужно ставить в соответствие с точкой на оси времени, имеющей координату

$$\bar{t} = \frac{1}{T} \sum_{k=t-T+1}^t k = t - \frac{(T-1)}{2}$$

то есть с точкой, сдвинутой влево по оси времени от момента t на величину $\Delta t = (T-1)/2$.

11.4. Взвешенная скользящая средняя.

Взвешенная скользящая средняя придает больший вес последним данным. Она рассчитывается путем умножения каждой

цены в пределах периода времени $[t - T + 1, t]$ на соответствующий вес. В простейшем случае при линейно убывающих весах от момента t до момента $t - T + 1$ формула имеет вид:

$$WMA_t = \frac{2}{T(T+1)} \sum_{k=t-T+1}^t [k - (t - T)] \cdot y_k$$

Цена в момент времени $k = t$ входит в формулу для расчета с максимальным весом $w = 2/(T + 1)$, а цена в момент времени $k = t - T + 1$ входит в формулу для расчета с минимальным весом $w = 2/(T \cdot (T + 1))$.

При отсутствии специализированных программ технического анализа, для расчета линейно взвешенной скользящей средней может быть полезна рекуррентная формула

$$WMA_t = WMA_{t-1} + \frac{2}{T} y_t - \frac{2}{T(T+1)} y_{t-T} - \frac{2}{(T+1)} SMA_t$$

Из этой формулы следует, что реакция WMA на выбытие цены из интервала расчета менее выражена, чем у SMA, и эта реакция тем меньше, чем больше период скользящей средней.

11.5. Экспоненциальная скользящая средняя.

Как и в случае взвешенной средней, экспоненциальная скользящая средняя придает больший вес последним данным, однако при расчете используется вся история цен. Рекуррентная формула для ее вычисления имеет вид:

$$EMA_t = \alpha \cdot y_t + (1 - \alpha) \cdot EMA_{t-1}$$

$$0 < \alpha \leq 1$$

Показательный процент α определяет степень сглаживания. Чем больше α , тем меньше степень сглаживания. При $\alpha = 1$ экспоненциальная скользящая средняя равна цене.

ЕМА лишена недостатка, присущего SMA и WMA, связанного с фиксированным интервалом расчета скользящей средней.

Формулу для вычисления ЕМА можно записать в явном виде, если предположить, что в нулевой момент времени скользящая средняя совпадает с ценой ($EMA_0 = y_0$):

$$\begin{aligned}
EMA_t &= \alpha \cdot y_t + (1-\alpha) \cdot EMA_{t-1} = \\
&= \alpha \cdot y_t + \alpha \cdot (1-\alpha) \cdot y_{t-1} + (1-\alpha)^2 \cdot EMA_{t-2} = \\
&= \alpha \cdot y_t + \alpha \cdot (1-\alpha) \cdot y_{t-1} + \alpha \cdot (1-\alpha)^2 \cdot y_{t-2} + (1-\alpha)^3 \cdot EMA_{t-3} = \\
&= \alpha \cdot y_t + \alpha \cdot (1-\alpha) \cdot y_{t-1} + \alpha \cdot (1-\alpha)^2 \cdot y_{t-2} + \dots + (1-\alpha)^t \cdot y_0
\end{aligned}$$

Следовательно

$$EMA_t = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i \cdot y_{t-i} + (1-\alpha)^t \cdot y_0$$

или (эквивалентная форма записи)

$$EMA_t = \alpha \sum_{k=1}^t (1-\alpha)^{t-k} \cdot y_k + (1-\alpha)^t \cdot y_0$$

Вычисленное значение ЕМА нужно ставить в соответствие с точкой на оси времени, имеющей координату

$$\bar{t} = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i (t-i) = t\alpha \sum_{i=0}^{t-1} (1-\alpha)^i - \alpha \sum_{i=0}^{t-1} (1-\alpha)^i i$$

Суммы в последней формуле вычисляются как

$$\begin{aligned}
\sum_{i=0}^{t-1} (1-\alpha)^i &= \frac{1 - (1-\alpha)^t}{\alpha} \\
\sum_{i=0}^{t-1} (1-\alpha)^i \cdot i &= \frac{(1-\alpha) - (1+\alpha t - \alpha)(1-\alpha)^t}{\alpha^2}
\end{aligned}$$

После несложных преобразований получаем, что

$$\bar{t} = t - \frac{(1-\alpha)}{\alpha} + \frac{(1-\alpha)^{t+1}}{\alpha}$$

При достаточно большом t , т.к. $(1-\alpha) < 1$, то $(1-\alpha)^{t+1} \approx 0$, значит можно пренебречь последним слагаемым и написать приближенное выражение $\bar{t} \approx t - (1-\alpha)/\alpha$.

Период ЕМА

Момент времени \bar{t} сдвинут влево по оси времени от момента t на величину $\Delta t = (1-\alpha)/\alpha$. Если по аналогии с простой скользящей средней обозначить эту величину как

$\Delta t = (T - 1) / 2$, где T является периодом, то связь периода и показательного процента задается выражением:

$$\frac{(1 - \alpha)}{\alpha} = \frac{(T - 1)}{2}$$

Отсюда следуют формулы для конвертирования показательного процента в период и наоборот:

$$T = \frac{2}{\alpha} - 1 \quad \alpha = \frac{2}{T + 1}$$

С учетом этих соотношений можно переписать рекуррентную формулу для ЕМА:

$$EMA_t = \frac{2}{T + 1} \cdot y_t + \frac{T - 1}{T + 1} \cdot EMA_{t-1}$$

ЕМА произвольного порядка

До сих пор мы рассматривали экспоненциальную скользящую среднюю первого порядка, то есть сглаживанию подвергался непосредственно исходный динамический ряд:

$$EMA_t^{(1)} = \alpha \cdot y_t + (1 - \alpha) \cdot EMA_{t-1}^{(1)}$$

При обозначении ЕМА первого порядка верхний индекс обычно опускается.

Экспоненциальная скользящая средняя произвольного n -го порядка задается формулой:

$$EMA_t^{(n)} = \alpha \cdot EMA_t^{(n-1)} + (1 - \alpha) \cdot EMA_{t-1}^{(n)}$$

DEMA

Рассмотрим ошибку ЕМА, то есть величину $e_t = y_t - EMA_t$. Если прибавить к значению экспоненциальной скользящей средней цены значение экспоненциальной скользящей средней ошибки, то такая величина называется двойной экспоненциальной скользящей средней:

$$\begin{aligned} DEMA_t &= EMA_t + EMA(e_t) = EMA_t + EMA(y_t - EMA_t) = \\ &= 2 \cdot EMA_t - EMA(EMA_t) \equiv 2 \cdot EMA_t^{(1)} - EMA_t^{(2)} \end{aligned}$$

ТЕМА

Рассмотрим ошибку $DEMA_t$, то есть величину $e_t = y_t - DEMA_t$. Тогда тройная экспоненциальная скользящая средняя вычисляется по формуле:

$$TEMA_t = DEMA_t + EMA(e_t) = DEMA_t + EMA(y_t - DEMA_t)$$

После преобразований получим, что

$$\begin{aligned} TEMA_t &= 3 \cdot EMA_t - 3 \cdot EMA(EMA_t) + EMA(EMA(EMA_t)) \equiv \\ &\equiv 3 \cdot EMA_t^{(1)} - 3 \cdot EMA_t^{(2)} + EMA_t^{(3)} \end{aligned}$$

11.6. Точки пересечения экспоненциально сглаженных кривых.

Часто в момент времени t ("сегодня") необходимо знать, какая цена должна быть в момент времени $t+1$ ("завтра"), чтобы произошло пересечение цены y с какой-либо экспоненциально сглаженной кривой или пересечение двух различных экспоненциально сглаженных кривых. Приведем соответствующие формулы для некоторых наиболее важных случаев.

- 1) Пересечение цены y и EMA 1-го порядка

$$y_{t+1} = EMA_t^{(1)}$$

- 2) Пересечение цены y и EMA 2-го порядка

$$y_{t+1} = \frac{EMA_t^{(2)} + \alpha \cdot EMA_t^{(1)}}{1 + \alpha}$$

- 3) Пересечение цены y и $DEMA$

$$y_{t+1} = \frac{(1 - \alpha) \cdot ((2 - \alpha) \cdot EMA_t^{(1)} - EMA_t^{(2)})}{1 - \alpha \cdot (2 - \alpha)}$$

или

$$y_{t+1} = \frac{(1 - \alpha) \cdot (DEMA_t - \alpha \cdot EMA_t^{(1)})}{1 - \alpha \cdot (2 - \alpha)}$$

- 4) Пересечение двух EMA 1-го порядка различных периодов

$$y_{t+1} = \frac{(1 - \alpha_2) \cdot EMA2_t^{(1)} - (1 - \alpha_1) \cdot EMA1_t^{(1)}}{\alpha_1 - \alpha_2}$$

$EMA1_t^{(1)}$ характеризуется показательным процентом α_1 ,

$EMA2_t^{(1)}$ характеризуется показательным процентом α_2 .

- 5) Пересечение двух ЕМА 2-го порядка различных периодов

$$y_{t+1} = [\alpha_2 \cdot (1 - \alpha_2) \cdot EMA2_t^{(1)} + (1 - \alpha_2) \cdot EMA2_t^{(2)} - \\ - \alpha_1 \cdot (1 - \alpha_1) \cdot EMA1_t^{(1)} - (1 - \alpha_1) \cdot EMA1_t^{(2)}] / [\alpha_1^2 - \alpha_2^2]$$

$EMA1_t^{(1)}$ и $EMA1_t^{(2)}$ характеризуются показательным процентом α_1 ,

$EMA2_t^{(1)}$ и $EMA2_t^{(2)}$ характеризуются показательным процентом α_2 .

- 6) Пересечение ЕМА 1-го порядка (показательный процент α_1) и ЕМА 2-го порядка (показательный процент α_2)

$$y_{t+1} = [\alpha_2 \cdot (1 - \alpha_2) \cdot EMA2_t^{(1)} + (1 - \alpha_2) \cdot EMA2_t^{(2)} - \\ - (1 - \alpha_1) \cdot EMA1_t^{(1)}] / [\alpha_1 - \alpha_2^2]$$

11.7. Выбор величины показательного процента для экспоненциальной скользящей средней.

Для того, чтобы оценить, насколько хорошо подобрана величина показательного процента α , необходимо рассмотреть ошибки, возникающие при прогнозировании уровня цены в момент времени $t + 1$ ("завтра") значением ЕМА в момент времени t ("сегодня"). Введем обозначения:

- y_t - цена в момент времени t ,
- α - показательный процент сглаживания ряда цен,
- Y_t - ЕМА для ряда цен, т.е. $Y_t = \alpha \cdot y_t + (1 - \alpha) \cdot Y_{t-1}$,
- f_t - прогноз цены, причем $f_{t+1} = Y_t$,
- e_t - ошибка прогноза, т.е. $e_t = y_t - f_t$,
- β - показательный процент сглаживания ряда квадратов ошибок прогноза,
- Q_t - ЕМА для ряда квадратов ошибок прогноза, т.е. $Q_t = \beta \cdot e_t^2 + (1 - \beta) \cdot Q_{t-1}$.

Оптимизация величины показательного процента α - это подбор такого его значения, чтобы при фиксированном β добиться того, чтобы $Q_t \rightarrow \min$. Обычно величину β выбирают в пределах от 0.1 до 0.2, что приблизительно соответствует периоду сглаживания в пределах от 10 до 20.

11.8. Экспоненциальная скользящая средняя с переменным показательным процентом.

На нестабильных рынках имеет смысл использовать ЕМА с переменным показательным процентом, который по мере получения новых данных постоянно подстраивается к текущей рыночной ситуации. Введем обозначения:

- y_t - цена в момент времени t ,
- α_t - переменный показательный процент сглаживания ряда цен,
- Y_t - ЕМА для ряда цен, т.е. $Y_t = \alpha_t \cdot y_t + (1 - \alpha_t) \cdot Y_{t-1}$,
- f_t - прогноз цены, причем $f_{t+1} = Y_t$,
- e_t - ошибка прогноза: $e_t = y_t - f_t$,
- β - показательный процент сглаживания ошибок прогноза и модулей ошибок прогноза,
- E_t - ЕМА ошибок прогноза: $E_t = \beta \cdot e_t + (1 - \beta) \cdot E_{t-1}$,
- A_t - ЕМА модулей ошибок: $A_t = \beta \cdot |e_t| + (1 - \beta) \cdot A_{t-1}$.

Значение переменного показательного процента в каждый момент времени вычисляют по формуле $\alpha_t = |E_t / A_t|$. Величину β выбирают в пределах от 0.1 до 0.2.

11.9. Дисперсия скользящих средних.

Рассмотрим на качественном уровне вопрос о том, как соотносится дисперсия значений исходного динамического ряда с дисперсией скользящей средней этого ряда. Для простоты будем предполагать, что исходный динамический ряд состоит из случайных величин, имеющих одинаковую дисперсию σ^2 , причем в пределах интервала сглаживания средняя величина коэффици

ента корреляции между значениями исходного ряда в различные моменты времени равна ρ .

В общем виде формула для вычисления любой скользящей средней имеет вид:

$$Y = \sum_k w_k y_k$$

Дисперсия случайной величины, являющейся линейной комбинацией коррелированных случайных величин равна:

$$\sigma_Y^2 = \sum_k w_k^2 \sigma_k^2 + 2 \sum_k \sum_{i>k} w_i w_k \rho_{ik} \sigma_i \sigma_k$$

Используя допущения о постоянстве дисперсий и коэффициентов корреляций, эту формулу можно упростить:

$$\sigma_Y^2 = \sigma^2 \sum_k w_k^2 + 2\rho\sigma^2 \sum_k \sum_{i>k} w_i w_k$$

Следовательно

$$\frac{\sigma_Y^2}{\sigma^2} = \sum_k w_k^2 + 2\rho \sum_k \sum_{i>k} w_i w_k$$

Согласно правилу нормирования весов справедливо равенство

$$\sum_k w_k^2 + 2 \sum_k \sum_{i>k} w_i w_k = 1$$

Отсюда можно сделать вывод, что так как $\rho \leq 1$, то $\sigma_Y^2 \leq \sigma^2$.

Дисперсия простой скользящей средней

Формула для простой скользящей средней имеет вид:

$$Y = \frac{1}{T} \sum_{k=t-T+1}^t y_k$$

Найдем суммы весов, входящие формулу для вычисления отношения дисперсии скользящей средней к дисперсии исходного ряда:

$$\begin{aligned} \sum_{k=t-T+1}^t w_k^2 &= \sum_{k=t-T+1}^t (1/T)^2 = \frac{1}{T} \\ 2 \sum_{k=t-T+1}^t \sum_{i=k+1}^t w_i w_k &= 2 \sum_{k=t-T+1}^t \sum_{i=k+1}^t (1/T)^2 = \frac{T-1}{T} \end{aligned}$$

В итоге получаем: $\frac{\sigma_Y^2}{\sigma^2} = \frac{1}{T} + \rho \cdot \frac{T-1}{T}$

Дисперсия экспоненциальной скользящей средней

Формула для экспоненциальной скользящей средней имеет вид:

$$Y = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i \cdot y_{t-i} + (1-\alpha)^t \cdot y_0$$

Приведем выражения для сумм весов, входящие в формулу для вычисления отношения дисперсии скользящей средней к дисперсии исходного ряда:

$$\sum_{k=0}^t w_k^2 = (1-\alpha)^{2t} + \alpha^2 \sum_{k=1}^t (1-\alpha)^{2k} = \frac{\alpha}{2-\alpha} + \frac{2-2\alpha}{2-\alpha} (1-\alpha)^{2t}$$

$$2 \sum_{k=0}^t \sum_{i=k+1}^t w_i w_k = \frac{2-2\alpha}{2-\alpha} - \frac{2-2\alpha}{2-\alpha} (1-\alpha)^{2t}$$

При достаточно большом t , так как $(1-\alpha) < 1$, то $(1-\alpha)^{2t} \approx 0$.

Следовательно $\frac{\sigma_Y^2}{\sigma^2} = \frac{\alpha}{2-\alpha} + \rho \frac{2-2\alpha}{2-\alpha} = \frac{1}{T} + \rho \cdot \frac{T-1}{T}$

12. АДАПТИВНОЕ МОДЕЛИРОВАНИЕ ДИНАМИЧЕСКИХ РЯДОВ

12.1. Введение.

Аналитическая аппроксимация динамического ряда какой-либо моделью с помощью МНК имеет ряд особенностей, которые накладывают ограничения на ее применение:

- динамический ряд, к которому применяется аппроксимация, должен быть достаточно длинным,
- применение аналитической аппроксимации эффективно только в случае, если уровни динамического ряда меняются достаточно плавно и медленно, то есть ряд должен быть неволатильным,
- аналитическая аппроксимация не адаптируется к появлению новых данных, то есть при появлении новых данных необходимо пересчитать параметры модели, а иногда возможно пересмотреть саму модель,
- при расчете параметров модели все эмпирические данные входят с одинаковым весом, хотя понятно, что более поздние данные имеют большую ценность.

Однако ряды цен активов как правило подвержены значительным колебаниям, которые аппроксимация не может предвидеть. Поэтому на практике применительно к таким рядам используют методы адаптивного моделирования, которые базируются на экспоненциальном сглаживании динамического ряда (экспоненциальной скользящей средней).

Основным преимуществом методов, основанных на экспоненциальном сглаживании, является учет временной ценности данных и, следовательно, постоянное адаптирование к изменяющимся уровням динамического ряда, что имеет решающее значение при моделировании и прогнозировании волатильных рядов.

12.2. Адаптивное моделирование линейного тренда с помощью экспоненциальных скользящих средних.

Пусть есть основания полагать, что исходный динамический ряд $\{y_t\}$ можно описать линейной функцией

$f(t) = a^{(0)} + a^{(1)} \cdot t$. Наличие случайных отклонений приведет к тому, что связь между рассчитанными по модели значениями f_t и реальными уровнями динамического ряда y_t будет выражаться в виде:

$$y_t = f_t + e_t = a^{(0)} + a^{(1)} \cdot t + e_t$$

где e_t - это расхождения между моделью и реальными уровнями. Используя экспоненциальные скользящие средние вычислим неизвестные параметры $(a^{(0)}, a^{(1)})$.

Обозначения

Введем следующие обозначения:

- $Y_t^{(1)}$ - ЕМА 1-го порядка исходного динамического ряда,
- $Y_t^{(2)}$ - ЕМА 2-го порядка исходного динамического ряда,
- $E_t^{(1)}$ - ЕМА 1-го порядка ошибок модели,
- $E_t^{(2)}$ - ЕМА 2-го порядка ошибок модели,
- α - показательный процент ЕМА.

Вычисление $Y_t^{(1)}$

$$\begin{aligned} Y_t^{(1)} &= \alpha \sum_{i=0}^{t-1} (1-\alpha)^i y_{t-i} + (1-\alpha)^t y_0 = \\ &= \alpha \sum_{i=0}^{t-1} (1-\alpha)^i (a^{(0)} + a^{(1)}(t-i) + e_{t-i}) + (1-\alpha)^t (a^{(0)} + e_0) = \\ &= \alpha (a^{(0)} + a^{(1)}t) \sum_{i=0}^{t-1} (1-\alpha)^i - \alpha a^{(1)} \sum_{i=0}^{t-1} (1-\alpha)^i i + (1-\alpha)^t a^{(0)} + \\ &+ \left[\alpha \sum_{i=0}^{t-1} (1-\alpha)^i e_{t-i} + (1-\alpha)^t e_0 \right] \end{aligned}$$

Суммы в последней формуле вычисляются как

$$\sum_{i=0}^{t-1} (1-\alpha)^i = \frac{1 - (1-\alpha)^t}{\alpha}$$

$$\sum_{i=0}^{t-1} (1-\alpha)^i i = \frac{(1-\alpha) - (1+\alpha t - \alpha)(1-\alpha)^t}{\alpha^2}$$

При достаточно большом t , так как $(1-\alpha) < 1$, то $(1-\alpha)^t \approx 0$ и можно написать приближенные выражения:

$$\sum_{i=0}^{t-1} (1-\alpha)^i \approx \frac{1}{\alpha}$$

$$\sum_{i=0}^{t-1} (1-\alpha)^i i \approx \frac{(1-\alpha)}{\alpha^2}$$

Выражение в квадратных скобках равно $E_t^{(1)}$.

С учетом всего вышесказанного формула для $Y_t^{(1)}$ примет вид:

$$Y_t^{(1)} = a^{(0)} + a^{(1)}t - a^{(1)} \frac{(1-\alpha)}{\alpha} + E_t^{(1)}$$

или

$$Y_t^{(1)} = f_t - a^{(1)} \frac{(1-\alpha)}{\alpha} + E_t^{(1)}$$

Очевидно, что между ЕМА 1-го порядка $Y_t^{(1)}$ и моделью f_t существует постоянный сдвиг, равный $-a^{(1)} \cdot (1-\alpha)/\alpha$. Величина этого сдвига пока неизвестна, так как она выражается через неизвестный параметр $a^{(1)}$.

Вычисление $Y_t^{(2)}$

$$\begin{aligned} Y_t^{(2)} &= \alpha \sum_{i=0}^{t-1} (1-\alpha)^i Y_{t-i}^{(1)} + (1-\alpha)^t Y_0^{(1)} = \\ &= \alpha \sum_{i=0}^{t-1} (1-\alpha)^i \left(a^{(0)} + a^{(1)}(t-i) - a^{(1)} \frac{(1-\alpha)}{\alpha} + E_{t-i}^{(1)} \right) + \\ &+ (1-\alpha)^t \left(a^{(0)} - a^{(1)} \frac{(1-\alpha)}{\alpha} + E_0^{(1)} \right) \end{aligned}$$

Дальнейшие выкладки полностью аналогичны тем, которые были сделаны при вычислении $Y_t^{(1)}$. Приведем сразу конечный результат:

$$Y_t^{(2)} = a^{(0)} + a^{(1)}t - 2a^{(1)} \frac{(1-\alpha)}{\alpha} + E_t^{(2)}$$

или

$$Y_t^{(2)} = f_t - 2a^{(1)} \frac{(1-\alpha)}{\alpha} + E_t^{(2)}$$

Вычисление параметров линейного тренда

Имеем систему уравнений с двумя неизвестными $(a^{(0)}, a^{(1)})$:

$$\begin{cases} Y_t^{(1)} = a^{(0)} + a^{(1)}t - a^{(1)} \frac{(1-\alpha)}{\alpha} + E_t^{(1)} \\ Y_t^{(2)} = a^{(0)} + a^{(1)}t - 2a^{(1)} \frac{(1-\alpha)}{\alpha} + E_t^{(2)} \end{cases}$$

Решая эту систему находим неизвестные параметры $(a^{(0)}, a^{(1)})$

$$a^{(0)} = Y_t^{(1)} + \left[(Y_t^{(1)} - Y_t^{(2)}) - (E_t^{(1)} - E_t^{(2)}) \right] \cdot \left(1 - \frac{\alpha}{1-\alpha}t \right) - E_t^{(1)}$$

$$a^{(1)} = \frac{\alpha}{1-\alpha} \cdot \left[(Y_t^{(1)} - Y_t^{(2)}) - (E_t^{(1)} - E_t^{(2)}) \right]$$

При переносе начала отсчета в точку t получим

$$a_t^{(0)} = (2Y_t^{(1)} - Y_t^{(2)}) - (2E_t^{(1)} - E_t^{(2)})$$

$$a_t^{(1)} = \frac{\alpha}{1-\alpha} \cdot \left[(Y_t^{(1)} - Y_t^{(2)}) - (E_t^{(1)} - E_t^{(2)}) \right]$$

В разные моменты времени t значения коэффициентов будут различны. Поэтому в формулах они отмечены соответствующими моменту времени индексами.

Прогноз уровней динамического ряда

Прогнозное значение динамического ряда в момент времени $t + \tau$ равно $f_{t+\tau} = a_t^{(0)} + a_t^{(1)} \cdot \tau$.

Замечание

В формулах для вычисления параметров линейной регрессии $(a_t^{(0)}, a_t^{(1)})$ присутствуют величины $E_t^{(1)}$ и $E_t^{(2)}$, которые являют

ся ЕМА от ошибок уравнения регрессии $e_t = y_t - f_t$, то есть при вычислении $(a_t^{(0)}, a_t^{(1)})$ возникает перекрестная ссылка. Поэтому на первом этапе нужно использовать упрощенные формулы, не учитывающие скользящих средних ошибок.

Алгоритм вычисления параметров линейного тренда

- 1) Рассчитать ЕМА 1-го и 2-го порядка исходного ряда:

$$Y_t^{(1)} \text{ и } Y_t^{(2)}$$

- 2) Вычислить в первом приближении параметры линейного тренда:

$$a_t^{(0)} = 2Y_t^{(1)} - Y_t^{(2)}$$

$$a_t^{(1)} = \frac{\alpha}{1-\alpha} \cdot (Y_t^{(1)} - Y_t^{(2)})$$

- 3) Для каждого момента времени t найти прогнозное значение на τ шагов вперед ($\tau \geq 1$) согласно уравнению регрессии:

$$f_{t+\tau} = a_t^{(0)} + a_t^{(1)} \cdot \tau$$

- 4) Рассчитать ошибки прогноза:

$$e_t = y_t - f_t$$

- 5) Вычислить ЕМА 1-го и 2-го порядка ошибок прогноза:

$$E_t^{(1)} \text{ и } E_t^{(2)}$$

- 6) Определить окончательные значения параметров линейного тренда:

$$a_t^{(0)} = (2Y_t^{(1)} - Y_t^{(2)}) - (2E_t^{(1)} - E_t^{(2)})$$

$$a_t^{(1)} = \frac{\alpha}{1-\alpha} \cdot [(Y_t^{(1)} - Y_t^{(2)}) - (E_t^{(1)} - E_t^{(2)})]$$

ЕМА ошибок могут ухудшить качество прогноза. В этом случае при расчете параметров линейного тренда нужно остановиться на шаге 2 этого алгоритма.

12.3. Адаптивное моделирование параболического тренда с помощью экспоненциальных скользящих средних.

Пусть исходный динамический ряд $\{y_t\}$ можно описать параболой $f(t) = a^{(0)} + a^{(1)}t + a^{(2)}t^2$. Наличие случайных отклонений

приведет к тому, что связь между рассчитанными по модели значениями f_t и реальными уровнями динамического ряда y_t будет выражаться в виде:

$$y_t = f_t + e_t = a^{(0)} + a^{(1)}t + a^{(2)}t^2 + e_t$$

где e_t - это расхождения между моделью и реальными уровнями. Используя экспоненциальные скользящие средние вычислим неизвестные параметры $(a^{(0)}, a^{(1)}, a^{(2)})$.

Обозначения

Введем обозначения:

- $Y_t^{(1)}$ - ЕМА 1-го порядка исходного динамического ряда,
- $Y_t^{(2)}$ - ЕМА 2-го порядка исходного динамического ряда,
- $Y_t^{(3)}$ - ЕМА 3-го порядка исходного динамического ряда,
- $E_t^{(1)}$ - ЕМА 1-го порядка ошибок модели,
- $E_t^{(2)}$ - ЕМА 2-го порядка ошибок модели,
- $E_t^{(3)}$ - ЕМА 3-го порядка ошибок модели,
- α - показательный процент ЕМА.

Вычисление $Y_t^{(1)}$, $Y_t^{(2)}$ и $Y_t^{(3)}$

$$\begin{aligned} Y_t^{(1)} &= \alpha \sum_{i=0}^{t-1} (1-\alpha)^i y_{t-i} + (1-\alpha)^t y_0 = \\ &= \alpha \sum_{i=0}^{t-1} (1-\alpha)^i (a^{(0)} + a^{(1)}(t-i) + a^{(2)}(t-i)^2 + e_{t-i}) + \\ &+ (1-\alpha)^t (a^{(0)} + e_0) \\ Y_t^{(1)} &= \alpha (a^{(0)} + a^{(1)}t + a^{(2)}t^2) \sum_{i=0}^{t-1} (1-\alpha)^i - \alpha (a^{(1)} + 2a^{(2)}t) \sum_{i=0}^{t-1} (1-\alpha)^i i + \\ &+ \alpha a^{(2)} \sum_{i=0}^{t-1} (1-\alpha)^i i^2 + (1-\alpha)^t a^{(0)} + \left[\alpha \sum_{i=0}^{t-1} (1-\alpha)^i e_{t-i} + (1-\alpha)^t e_0 \right] \end{aligned}$$

При достаточно большом t , так как $(1-\alpha) < 1$, то $(1-\alpha)^t \approx 0$. Следовательно:

$$\sum_{i=0}^{t-1} (1-\alpha)^i \approx \frac{1}{\alpha}$$

$$\sum_{i=0}^{t-1} (1-\alpha)^i i \approx \frac{(1-\alpha)}{\alpha^2}$$

$$\sum_{i=0}^{t-1} (1-\alpha)^i i^2 \approx \frac{(1-\alpha)(2-\alpha)}{\alpha^3}$$

Выражение в квадратных скобках равно $E_t^{(1)}$.

С учетом всего вышесказанного формула для $Y_t^{(1)}$ примет вид:

$$Y_t^{(1)} = b^{(0)} + b^{(1)}t + a^{(2)}t^2 + E_t^{(1)}$$

где

$$b^{(0)} = a^{(0)} - a^{(1)} \frac{1-\alpha}{\alpha} + a^{(2)} \frac{(1-\alpha)(2-\alpha)}{\alpha^2}$$

$$b^{(1)} = a^{(1)} - 2a^{(2)} \frac{1-\alpha}{\alpha}$$

Расчет $Y_t^{(2)}$ и $Y_t^{(3)}$ проводятся по той же схеме. Приведем сразу конечный результат.

$$Y_t^{(2)} = c^{(0)} + c^{(1)}t + a^{(2)}t^2 + E_t^{(2)}$$

где

$$c^{(0)} = b^{(0)} - b^{(1)} \frac{1-\alpha}{\alpha} + a^{(2)} \frac{(1-\alpha)(2-\alpha)}{\alpha^2}$$

$$c^{(1)} = b^{(1)} - 2a^{(2)} \frac{1-\alpha}{\alpha}$$

$$Y_t^{(3)} = d^{(0)} + d^{(1)}t + a^{(2)}t^2 + E_t^{(3)}$$

где

$$d^{(0)} = c^{(0)} - c^{(1)} \frac{1-\alpha}{\alpha} + a^{(2)} \frac{(1-\alpha)(2-\alpha)}{\alpha^2}$$

$$d^{(1)} = c^{(1)} - 2a^{(2)} \frac{1-\alpha}{\alpha}$$

Вычисление параметров параболического тренда

Используя эти результаты найдем неизвестные параметры параболического тренда $(a^{(0)}, a^{(1)}, a^{(2)})$. Перенеся начало системы отсчета в точку t после довольно громоздких преобразований можно получить:

$$\begin{aligned} a_t^{(0)} &= (3Y_t^{(1)} - 3Y_t^{(2)} + Y_t^{(3)}) - (3E_t^{(1)} - 3E_t^{(2)} + E_t^{(3)}) \\ a_t^{(1)} &= \frac{\alpha}{2(1-\alpha)^2} (Y_t^{(1)}(6-5\alpha) - Y_t^{(2)}(10-8\alpha) + Y_t^{(3)}(4-3\alpha)) - \\ &\quad - \frac{\alpha}{2(1-\alpha)^2} (E_t^{(1)}(6-5\alpha) - E_t^{(2)}(10-8\alpha) + E_t^{(3)}(4-3\alpha)) \\ a_t^{(2)} &= \frac{\alpha^2}{2(1-\alpha)^2} (Y_t^{(1)} - 2Y_t^{(2)} + Y_t^{(3)}) - \\ &\quad - \frac{\alpha^2}{2(1-\alpha)^2} (E_t^{(1)} - 2E_t^{(2)} + E_t^{(3)}) \end{aligned}$$

Прогноз уровней динамического ряда

Прогнозное значение динамического ряда в момент времени $t + \tau$ равно $f_{t+\tau} = a_t^{(0)} + a_t^{(1)} \cdot \tau + a_t^{(2)} \cdot \tau^2$.

Замечание

В формулах для вычисления параметров параболической регрессии $(a_t^{(0)}, a_t^{(1)}, a_t^{(2)})$ присутствуют величины $E_t^{(1)}, E_t^{(2)}$ и $E_t^{(3)}$, которые являются ЕМА от ошибок уравнения регрессии $e_t = y_t - f_t$, то есть при вычислении $(a_t^{(0)}, a_t^{(1)}, a_t^{(2)})$ возникает перекрестная ссылка. Поэтому на первом этапе нужно использовать упрощенные формулы, не учитывающие скользящих средних ошибок.

Алгоритм вычисления параметров параболического тренда

- 1) Рассчитать ЕМА 1-го, 2-го и 3-го порядков исходного ряда:

$$Y_t^{(1)}, Y_t^{(2)} \text{ и } Y_t^{(3)}$$

- 2) Вычислить в первом приближении параметры параболического тренда:

$$a_t^{(0)} = 3Y_t^{(1)} - 3Y_t^{(2)} + Y_t^{(3)}$$

$$a_t^{(1)} = \frac{\alpha}{2(1-\alpha)^2} (Y_t^{(1)}(6-5\alpha) - Y_t^{(2)}(10-8\alpha) + Y_t^{(3)}(4-3\alpha))$$

$$a_t^{(2)} = \frac{\alpha^2}{2(1-\alpha)^2} (Y_t^{(1)} - 2Y_t^{(2)} + Y_t^{(3)})$$

- 3) Для каждого момента времени t найти прогнозное значение на τ шагов вперед ($\tau \geq 1$) согласно уравнению регрессии:

$$f_{t+\tau} = a_t^{(0)} + a_t^{(1)} \cdot \tau + a_t^{(2)} \cdot \tau^2$$

- 4) Рассчитать ошибки прогноза:

$$e_t = y_t - f_t$$

- 5) Вычислить ЕМА 1-го, 2-го и 3-го порядков ошибок прогноза:

$$E_t^{(1)}, E_t^{(2)} \text{ и } E_t^{(3)}$$

- 6) Определить окончательные значения параметров параболического тренда:

$$a_t^{(0)} = (3Y_t^{(1)} - 3Y_t^{(2)} + Y_t^{(3)}) - (3E_t^{(1)} - 3E_t^{(2)} + E_t^{(3)})$$

$$a_t^{(1)} = \frac{\alpha}{2(1-\alpha)^2} (Y_t^{(1)}(6-5\alpha) - Y_t^{(2)}(10-8\alpha) + Y_t^{(3)}(4-3\alpha)) -$$

$$- \frac{\alpha}{2(1-\alpha)^2} (E_t^{(1)}(6-5\alpha) - E_t^{(2)}(10-8\alpha) + E_t^{(3)}(4-3\alpha))$$

$$a_t^{(2)} = \frac{\alpha^2}{2(1-\alpha)^2} (Y_t^{(1)} - 2Y_t^{(2)} + Y_t^{(3)}) - \frac{\alpha^2}{2(1-\alpha)^2} (E_t^{(1)} - 2E_t^{(2)} + E_t^{(3)})$$

ЕМА ошибок могут ухудшить качество прогноза. В этом случае при расчете параметров параболического тренда нужно остановиться на шаге 2 этого алгоритма.

12.4. Выбор величины показательного процента при адаптивном моделировании.

Для того, чтобы оценить, насколько хорошо подобрана величина показательного процента α , необходимо рассмотреть

ошибки, возникающие при прогнозировании уровня цены в момент времени $t + \tau$ моделью:

$$f_{t+\tau} = a_t^{(0)} + a_t^{(1)} \cdot \tau \quad \text{или} \quad f_{t+\tau} = a_t^{(0)} + a_t^{(1)} \cdot \tau + a_t^{(2)} \cdot \tau^2$$

Введем обозначения:

- ε_t - ошибка прогноза ($\varepsilon_t = y_t - f_t$):
 для линейной модели $\varepsilon_t = y_t - (a_{t-\tau}^{(0)} + a_{t-\tau}^{(1)} \cdot \tau)$,
 для параболы $\varepsilon_t = y_t - (a_{t-\tau}^{(0)} + a_{t-\tau}^{(1)} \cdot \tau + a_{t-\tau}^{(2)} \cdot \tau^2)$.
 Заметим, что ошибки прогноза зависят не только от α , но и от интервала прогнозирования τ .
- β - показательный процент сглаживания ряда квадратов ошибок прогноза,
- Q_t - ЕМА для ряда квадратов ошибок прогноза:

$$Q_t = \beta \cdot \varepsilon_t^2 + (1 - \beta) \cdot Q_{t-1}.$$

Оптимизация величины показательного процента α - это подбор такого его значения, чтобы при фиксированном β добиться того, чтобы $Q_t \rightarrow \min$. Обычно величину β выбирают в пределах от 0.1 до 0.2, что приблизительно соответствует периоду сглаживания в пределах от 10 до 20.

12.5. Адаптивное моделирование с переменным показательным процентом.

На нестабильных рынках имеет смысл использовать адаптивное моделирование с переменным показательным процентом α_t , который по мере получения новых данных постоянно подстраивается к текущей рыночной ситуации.

Введем обозначения:

- ε_t - ошибка прогноза ($\varepsilon_t = y_t - f_t$):
 для линейной модели $\varepsilon_t = y_t - (a_{t-\tau}^{(0)} + a_{t-\tau}^{(1)} \cdot \tau)$,
 для параболы $\varepsilon_t = y_t - (a_{t-\tau}^{(0)} + a_{t-\tau}^{(1)} \cdot \tau + a_{t-\tau}^{(2)} \cdot \tau^2)$.
- β - показательный процент сглаживания ошибок прогноза и модулей ошибок прогноза,
- E_t - ЕМА ошибок прогноза:

$$E_t = \beta \cdot \varepsilon_t + (1 - \beta) \cdot E_{t-1},$$

- A_t - ЕМА модулей ошибок прогноза,

$$A_t = \beta \cdot |\varepsilon_t| + (1 - \beta) \cdot A_{t-1}.$$

Значение переменного показательного процента в каждый момент времени вычисляют по формуле $\alpha_t = |E_t / A_t|$. Величину β выбирают в пределах от 0.1 до 0.2.

13. МЕХАНИЧЕСКИЕ ТОРГОВЫЕ СИСТЕМЫ

13.1. Введение.

Определим инвестирование как вложение свободных денежных средств в различные виды финансовых активов с целью получения прибыли. При формировании инвестиционного портфеля выбирается такой набор активов и такой способ управления ими, которые бы обеспечивали ожидаемый доход не ниже заранее заданного минимального значения при риске получения дохода не выше заранее заданного максимального значения. Существуют два основных способа управления портфелем ценных бумаг: активный и пассивный.

Суть пассивного управления состоит в создании хорошо диверсифицированного, состоящего из большого количества активов портфеля, и продолжительного удерживания его в неизменном состоянии. Пассивный портфель характеризуется низким оборотом и малым уровнем накладных расходов.

Мы сконцентрируем внимание на рассмотрении активного управления портфелем, которое нацелено на получение дохода выше среднерыночного уровня. Активное управление подразумевает:

- выбор небольшого количества высоколиквидных активов для формирования портфеля,
- определение правил открытия и закрытия позиций по каждому из активов,
- определение объема открываемых позиций,
- оптимизацию портфеля, то есть методы снижения рисков.

Определим некоторые понятия, которые будем использовать в дальнейшем.

Механическая торговая система (МТС) - набор правил, однозначно определяющих моменты открытия и закрытия позиций, то есть МТС задает правила входа в позицию, правила выхода из выигрывающей позиции, правила выхода из проигрывающей позиции.

Управление капиталом - набор правил, определяющих объем открываемых позиций в момент поступления соответствующих сигналов от МТС.

Оптимизация портфеля - методы, позволяющие выбрать такой состав портфеля активов и торгующих эти активы механических систем, которые бы в наибольшей степени соответствовали инвестиционным предпочтениям конкретного трейдера.

13.2. Механический и интуитивный подход к торговле.

Динамику биржевых цен активов можно представить как стохастический и нестационарный процесс. Однако, это не исключает возможность нахождения такого набора эмпирических правил, что проведение в соответствии с ними торговых операций позволяет увеличить доходность и/или уменьшить риск вложения в данный актив по сравнению с пассивной стратегией "купил и держи".

Создание механической торговой системы - это полная формализация таких правил. При этом нужно понимать, что так как правила открытия и закрытия позиций разрабатываются на основе прошлой истории цен, то не существует гарантии того, что МТС на их основе будет успешно работать и в будущем. Но чем более качественно проведено тестирование МТС, тем больше оснований надеяться на то, что ее результаты при реальной торговле будут находиться в приемлемых для трейдера пределах.

У механического подхода есть два преимущества по сравнению с часто практикуемым интуитивным подходом к торговле:

- при принятии торговых решений исключается эмоциональный фактор,
- принятые торговые решения не являются субъективными, следовательно оправдано формальное статистическое исследование результатов работы МТС, позволяющее найти и скорректировать ее слабые места.

Следует еще раз подчеркнуть, что для любой механической системы существует вероятность того, что по истечении любого интервала времени в результате проведения торговых операций по ее сигналам будет получен убыток. Однако, для достаточно хорошей системы эта вероятность тем меньше, чем больше время торговли. Поэтому считается, что для того, чтобы МТС смогла реализовать свое статистическое преимущество, необходимо не менее 2-3 лет.

13.3. Свойства МТС.

Показатели механической торговой системы характеризуют результаты ее работы. Величина, разброс и устойчивость показателей определяют качество системы. Подробно показатели МТС и методы их статистического исследования будут рассмотрены ниже.

Параметрами механической торговой системы называются переменные, присутствующие в правилах открытия и закрытия позиций. В результате тестирования и оптимизации МТС определяется такой набор параметров, при котором величина, разброс и устойчивость важнейших показателей системы находятся в оптимальных для конкретного трейдера пределах. Хорошая МТС должна обладать следующими свойствами:

- иметь небольшое количество оптимизируемых параметров,
- иметь удовлетворительные показатели работы при реальных рыночных комиссиях,
- обладать устойчивостью показателей работы в области оптимальности параметров,
- должно существовать по крайней мере несколько активов, на которых система имеет приемлемые результаты без повторной оптимизации.

Число оптимизируемых параметров

Чем больше оптимизируемых параметров имеет МТС, тем меньше вероятность того, что она будет удовлетворительно работать при реальной торговле. Это связано с тем, что при большом числе параметров система подгоняется под исторический ряд цен, на котором происходило тестирование. Но ряды цен активов в большой степени носят случайный характер. Задачей же хорошей системы является не учет всех особенностей конкретной случайной выборки, а выявление более или менее постоянно действующих на рынке закономерностей. Здесь уместна аналогия с регрессионным анализом, где также нужно с осторожностью относиться к излишнему переусложнению модели.

Величина комиссии

Важным моментом при тестировании МТС является величина комиссии, которая должна соответствовать реальным на

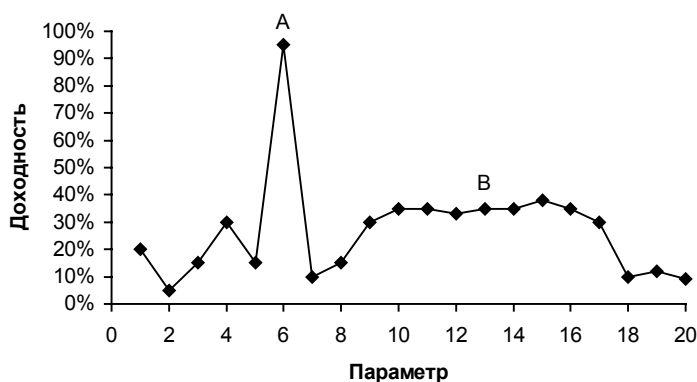
кладным расходам при совершении сделки, то есть она должна учитывать следующие величины:

- комиссию с оборота (брокерскую и биржевую),
- спрэд (разницу между котировками на покупку и продажу),
- проскальзывание (разницу между величиной сигнала МТС и реальной ценой исполнения сделки).

Если биржевая и брокерская комиссии не зависят от вида актива, то на величины спрэдов и проскальзываний существенное влияние оказывает ликвидность конкретного актива. В целом можно сказать, что при тестировании МТС суммарную комиссию следует выбирать не менее чем 0.5% от суммы сделки даже для наиболее ликвидных активов.

Устойчивость показателей в области оптимальности

Устойчивость показателей МТС в области оптимальности параметров проще всего проиллюстрировать графически для системы, зависящей от единственного параметра. Рассмотрим график зависимости одного из показателей системы (доходности в % годовых) от величины оптимизируемого параметра.



В данном случае выбор в качестве оптимальной величины параметра значение, соответствующее точке А (параметр=6, доходность≈95%год.), не является правильным, так как в районе этой точки величина доходности является неустойчивой и столь резкий пик вероятнее всего объясняется особенностями конкретной случайной выборки, на которой происходило тестирование.

Более разумным представляется выбор в качестве оптимальной величины параметра значение, соответствующее точке В (параметр=13, доходность≈35%год.), так как в области этой точки доходность устойчива к изменению параметра в достаточно широких пределах.

На устойчивость следует проверять все важные для трейдера показатели работы системы. У разных показателей области устойчивости будут вообще говоря разными. В качестве оптимального следует выбрать такое значение параметра, которое находится в области устойчивости большинства наиболее важных для трейдера показателей системы.

Удовлетворительные результаты на различных активах

Механическая система как правило создается, тестируется и оптимизируется на историческом ряде цен одного актива. Одной из главных задач создания хорошей МТС является предотвращение ее подгонки под конкретные исторические данные, которые наверняка не повторятся в будущем. Эффективным способом проверки МТС на излишнюю подгонку является переход на другие активы без изменения параметров системы. Если при переходе на исторические ряды цен других активов без повторной оптимизации параметров система продолжает показывать удовлетворительные результаты, то это повышает вероятность того, что она окажется прибыльной в реальной торговле. При этом надо иметь в виду, что если активы высококоррелированы, то это существенно снижает ценность такой проверки. Если же при переходе с ряда цен исходного актива (на котором МТС создавалась и оптимизировалась) на ряд цен высококоррелированного с ним актива система разваливается, то это служит веским основанием для отказа от нее.

13.4. Минимальное число сделок.

Для достоверной оценки величины и разброса показателей механической торговой системы количество сделок на периоде тестирования не должно быть меньше некоторого минимального значения. Считая, что результат отдельной сделки (например размер прибыли) является случайной величиной, оценим минимальный объем выборки для идентификации закона распределения этой величины. Для идентификации закона распределения необходимо построить гистограмму эмпирических частот и провести сравнение эмпирических и теоретических частот по критерию хи-квадрат.

Напомним, что минимальное количество столбцов гистограммы должно равняться пяти.

Так как приводимые ниже выкладки будут носить оценочный характер, предположим заранее, что исследуемая случайная величина подчиняется нормальному закону распределения.

Примем в качестве интервала возможных значений этой величины промежуток $(\mu - 2\sigma, \mu + 2\sigma)$. Вероятность оказаться внутри этого интервала 95.45%. Ширину столбца гистограммы можно найти, разделив ширину интервала возможных значений, то есть 4σ , на минимальное количество столбцов гистограммы, то есть число 5. В итоге получим, что ширина столбца равна 0.8σ . В таблице приведены значения нормированных теоретических частот попадания в соответствующие столбцы гистограммы для стандартной нормальной величины ($\mu = 0, \sigma = 1$).

Номер интервала	Левая граница	Правая граница	Нормированная частота
1	-2.0	-1.2	0.0923
2	-1.2	-0.4	0.2295
3	-0.4	0.4	0.3108
4	0.4	1.2	0.2295
5	1.2	2.0	0.0923
ИТОГО			0.9545

Для проверки закона распределения по критерию хи-квадрат в таблице должны присутствовать ненормированные частоты, причем минимальное значение частоты не должно быть меньше пяти. Следовательно, переход к ненормированной частоте можно сделать путем умножения на коэффициент $5 / 0.0923 = 54.16$ и округлив результат до целого. В итоге таблица частот примет вид:

Номер интервала	Левая граница	Правая граница	Частота
1	-2.0	-1.2	5
2	-1.2	-0.4	12
3	-0.4	0.4	17
4	0.4	1.2	12
5	1.2	2.0	5
ИТОГО			51

Таким образом мы получили, что для оценки показателей механической торговой системы объем выборки (количество сделок на периоде тестирования) не должен быть меньше 51. Заметим, что при вычислении этого значения мы использовали минимально возможное число столбцов гистограммы (5 столбцов) и не очень широкий доверительный интервал (95%). С увеличением количества столбцов (уменьшением ширины столбца) и увеличением доверительного интервала минимально необходимый объем выборки может существенно вырасти.

13.5. Тестирование МТС.

Целью тестирования механической торговой системы является проверка ее работы на историческом ряде цен. Тестирование как правило проводят, используя пакеты технического анализа. По результатам тестирования программа формирует отчеты, на основании которых делаются выводы о качестве МТС.

При оптимизации системы, в правилах, описывающих открытие и закрытие позиций, постоянные параметры заменяются на оптимизируемые переменные (ОРТ-переменные), для которых задается диапазон и шаг изменения. Затем программа проводит ряд тестов для всех возможных сочетаний ОРТ-переменных и формирует соответствующие отчеты. По результатам анализа этих отчетов выбирается такой набор параметров, при котором величина, разброс и устойчивость показателей системы являются оптимальными для трейдера.

Тестирование и оптимизацию системы рекомендуется проводить отдельно для длинных и коротких сделок.

Мы будем рассматривать следующие отчеты о тестировании системы: отчет о величине торгового счета (equity report), сгруппированный отчет о величине торгового счета, отчет о сделках (trades report), сводный отчет (results report).

ПРИМЕЧАНИЕ: далее в тексте содержание отчетов, а также обозначения показателей МТС и формулы для их вычисления могут отличаться от принятых в пакетах технического анализа.

13.6. Отчет о величине торгового счета.

Отчет о величине торгового счета показывает изменение стоимости портфеля на каждом ценовом баре. Отчет содержит следующие поля:

bar number	Порядковый номер текущего временного периода (бара).
date	Дата/время текущего бара.
position	Торговая позиция на конец текущего бара. Возможны следующие типы позиций: "OUT" - вне рынка, "LONG" - длинная позиция, "SHORT" - короткая позиция.
price	Текущая цена актива.
net change price	Изменение цены актива за текущий период.
% change price	Процентное изменение цены актива за текущий период.
equity	Текущая величина торгового счета.
net change equity	Изменение величины торгового счета за текущий период.
% change equity	Процентное изменение величины торгового счета за текущий период.

13.7. Сгруппированный отчет о величине торгового счета.

Сгруппированный отчет о величине торгового счета формируется на основе отчета о торговом счете и показывает изменение стоимости портфеля по укрупненным периодам времени. Отчет содержит следующие поля:

period grp	Период группировки: "D" - день, "W" - неделя, "M" - месяц, "Q" - квартал, "Y" - год.
period number	Порядковый номер текущего временного периода.
first date	Первая дата/время текущего периода.

last date	Последняя дата/время текущего периода.
price	Цена актива на конец текущего периода.
net change price	Изменение цены актива за текущий период.
% change price	Процентное изменение цены актива за текущий период.
equity	Величина торгового счета на конец текущего периода.
net change equity	Изменение величины торгового счета за текущий период.
% change equity	Процентное изменение величины торгового счета за текущий период.

13.8. Отчет о сделках.

Отчет о сделках формируется на основе отчета о торговом счете и описывает каждую торговую операцию, сгенерированную системой. Приведем список полей отчета о сделках и некоторые формулы для вычисления показателей системы:

Общая информация о сделке

trade number	Порядковый номер сгенерированной при тестировании сделки.
trade type	Тип сделки: "LONG" - длинная сделка, "SHORT" - короткая сделка.
bars in trade	Количество баров, в течение которых была открыта данная позиция.
days in trade	Число календарных дней, в течение которых была открыта данная позиция.
max price in trade	Максимальная цена актива от момента открытия позиции до момента ее закрытия.
min price in trade	Минимальная цена актива от момента открытия позиции до момента ее закрытия.

Показатели, характеризующие вход в позицию

enter bar	Номер бара, на котором была открыта позиция.
enter date	Дата входа в позицию.
enter price	Цена актива, по которой была открыта позиция.
enter equity	Величина торгового счета до входа в позицию.
enter comission	Сумма уплаченной комиссии за открытие позиции.
enter efficiency	Эффективность входа в позицию.

Показатели, характеризующие выход из позиции

exit bar	Номер бара, на котором была закрыта позиция.
exit date	Дата выхода из позиции.
exit price	Цена актива, по которой была закрыта позиция.
exit equity	Величина торгового счета после выхода из позиции.
exit comission	Сумма уплаченной комиссии за закрытие позиции.
exit efficiency	Эффективность выхода из позиции.

Показатели, характеризующие сделку

net profit	Величина дохода по данной сделке (в деньгах).
% profit	Величина дохода по данной сделке (в %).
net drawdown	Наибольшее снижение торгового счета в течение данной операции относительно входа в позицию (в деньгах).
% drawdown	Наибольшее снижение торгового счета в течение данной операции относительно входа в позицию (в %).

comission in trade	Сумма уплаченной комиссии за открытие и закрытие позиции.
trade efficiency	Эффективность сделки.

Эффективность входа в позицию показывает, насколько хорошо МТС в ходе конкретной сделки реализует потенциальную прибыль относительно цены входа в позицию и вычисляется по формулам:

- для длинных позиций

$$enter\ efficiency = \frac{max\ price\ in\ trade - enter\ price}{max\ price\ in\ trade - min\ price\ in\ trade}$$

- для коротких позиций

$$enter\ efficiency = \frac{enter\ price - min\ price\ in\ trade}{max\ price\ in\ trade - min\ price\ in\ trade}$$

Эффективность входа может принимать значения от 0 до 1.

Эффективность выхода из позиции показывает, насколько хорошо МТС в ходе конкретной сделки реализует потенциальную прибыль относительно цены выхода из позиции и вычисляется по формулам:

- для длинных позиций

$$exit\ efficiency = \frac{exit\ price - min\ price\ in\ trade}{max\ price\ in\ trade - min\ price\ in\ trade}$$

- для коротких позиций

$$exit\ efficiency = \frac{max\ price\ in\ trade - exit\ price}{max\ price\ in\ trade - min\ price\ in\ trade}$$

Эффективность выхода может принимать значения от 0 до 1.

Эффективность сделки показывает, насколько хорошо МТС в ходе конкретной сделки реализует общую потенциальную прибыль и вычисляется по формулам:

- для длинных позиций

$$trade\ efficiency = \frac{exit\ price - enter\ price}{max\ price\ in\ trade - min\ price\ in\ trade}$$

- для коротких позиций

$$trade\ efficiency = \frac{enter\ price - exit\ price}{max\ price\ in\ trade - min\ price\ in\ trade}$$

- общая формула

$$\text{trade efficiency} = \text{enter efficiency} + \text{exit efficiency} - 1$$

Эффективность сделки может принимать значения от -1 до 1.

Считается что у МТС средняя эффективность входов и средняя эффективность выходов должна быть больше 0.6, то есть средняя эффективность сделки должна превышать 0.2. Анализ эффективности наглядно показывает направления усовершенствования системы, так как позволяет раздельно оценить качество сигналов на вход в позицию и сигналов на выход из нее.

13.9. Сводный отчет.

Сводный отчет формируется на основе отчета о торговом счете и отчета о сделках и дает общую информацию о результатах тестирования системы. Приведем список полей сводного отчета и некоторые формулы для вычисления показателей системы:

Показатели, характеризующие линию торгового счета на периоде тестирования.

start date	Дата начала тестирования.
start equity	Величина торгового счета на начало тестирования (начальные инвестиции).
finish date	Дата окончания тестирования.
finish equity	Величина торгового счета после окончания тестирования.
total bars	Количество баров, в течение которых происходило тестирование.
total days	Число календарных дней, в течение которых происходило тестирование.
net system drawdown	Наибольшее снижение торгового счета относительно начальных инвестиций (в деньгах).
% system drawdown	Наибольшее снижение торгового счета относительно начальных инвестиций (%).

Показатели, характеризующие доходность стратегии "купил и держи" на периоде тестирования.

buy&hold net profit	Доход стратегии "купил и держи" на периоде тестирования (в деньгах).
buy&hold % profit	Доход стратегии "купил и держи" на периоде тестирования (%).
buy&hold % profit in year	Доходность стратегии "купил и держи" на периоде тестирования (в % годовых по формуле сложного процента).

Показатели, характеризующие доходность МТС на периоде тестирования.

total net profit	Доход на периоде тестирования (в деньгах).
total % profit	Доход на периоде тестирования (%).
total % profit in year	Доходность на периоде тестирования (в % годовых по формуле сложного процента).

Показатели, характеризующие сделки.

total trades	Общее число сделок.
% in trade	Доля времени на периоде тестирования, в течение которого система имела открытые позиции.
% out trade	Доля времени на периоде тестирования, в течение которого система была вне рынка.
avg net profit	Средний доход сделок (в деньгах).
stdev net profit	Среднеквадратичное отклонение дохода сделок (в деньгах).
avg % profit	Средний доход сделок (%).
stdev % profit	Среднеквадратичное отклонение дохода сделок (%).
avg net drawdown	Среднее наибольших снижений торгового счета (в деньгах).
stdev net drawdown	Среднеквадратичное отклонение наибольших снижений торгового счета (в деньгах).

avg % drawdown	Среднее наибольших снижений торгового счета (%).
stdev % drawdown	Среднеквадратичное отклонение наибольших снижений торгового счета (%).
max net drawdown	Наибольшее снижение торгового счета за отдельную сделку (в деньгах).
max % drawdown	Наибольшее снижение торгового счета за отдельную сделку (%).
avg net win / avg net loss	Отношение средней прибыли выигрышных сделок к среднему убытку проигрышных сделок.
total comission	Общая сумма уплаченной комиссии.

Показатели, характеризующие выигрышные сделки.

win trades	Количество выигрышных сделок.
win trades %	Процент выигрышных сделок.
win amount	Общая прибыль всех выигрышных сделок.
avg net win	Средняя прибыль выигрышных сделок (в деньгах).
stdev net win	Среднеквадратичное отклонение прибыли выигрышных сделок (в деньгах).
max net win	Максимальная прибыль выигрышной сделки (в деньгах).
avg % win	Средняя прибыль выигрышных сделок (%).
stdev % win	Среднеквадратичное отклонение прибыли выигрышных сделок (%).
max % win	Максимальная прибыль выигрышной сделки (%).
avg bars win	Средняя продолжительность выигрышных сделок (в барах).
stdev bars win	Среднеквадратичное отклонение продолжительности выигрышных сделок (в барах).

max bars win	Максимальная продолжительность выигрышной сделки (в барах).
max consecutive wins	Наибольшее количество выигрышных сделок, следовавших одна за другой.

Показатели, характеризующие проигрышные сделки.

loss trades	Количество проигрышных сделок.
loss trades %	Процент проигрышных сделок.
loss amount	Общий убыток всех проигрышных сделок.
avg net loss	Средний убыток проигрышных сделок (в деньгах).
stdev net loss	Среднеквадратичное отклонение убытка проигрышных сделок (в деньгах).
max net loss	Максимальный убыток проигрышной сделки (в деньгах).
avg % loss	Средний убыток проигрышных сделок (%).
stdev % loss	Среднеквадратичное отклонение убытка проигрышных сделок (%).
max % loss	Максимальный убыток проигрышной сделки (%).
avg bars loss	Средняя продолжительность проигрышных сделок (в барах).
stdev bars loss	Среднеквадратичное отклонение продолжительности проигрышных сделок (в барах).
max bars loss	Максимальная продолжительность проигрышной сделки (в барах).
max consecutive losses	Наибольшее количество проигрышных сделок, следовавших одна за другой.

Показатели, характеризующие эффективность сделок.

avg enter efficiency	Средняя эффективность открытия позиции.
stdev enter efficiency	Среднеквадратичное отклонение эффективности открытия позиции.

avg exit efficiency	Средняя эффективность закрытия позиции.
stdev exit efficiency	Среднеквадратичное отклонение эффективности закрытия позиции.
avg trade efficiency	Средняя эффективность сделок.
stdev trade efficiency	Среднеквадратичное отклонение эффективности сделок.

Приведем некоторые формулы для вычисления показателей механической торговой системы:

- показатели доходности системы:

$$total\ net\ profit = \sum_{i=1}^{total\ trades} net\ profit(i)$$

$$total\ \% \ profit = \prod_{i=1}^{total\ trades} (1 + \% \ profit(i)) - 1$$

$$total\ \% \ profit\ in\ year = (1 + total\ \% \ profit)^{365 / total\ days} - 1$$

- среднее значение и с.к.о. дохода сделок (в деньгах)

$$avg\ net\ profit = \frac{1}{total\ trades} \sum_{i=1}^{total\ trades} net\ profit(i)$$

$$stdev\ net\ profit =$$

$$= \sqrt{\frac{1}{total\ trades - 1} \sum_{i=1}^{total\ trades} (net\ profit(i) - avg\ net\ profit)^2}$$

- среднее значение и с.к.о. дохода сделок (в %)

$$avg\ \% \ profit = \left(\prod_{i=1}^{total\ trades} (1 + \% \ profit(i)) \right)^{1 / total\ trades} - 1 =$$

$$= \exp \left[\frac{1}{total\ trades} \sum_{i=1}^{total\ trades} \ln(1 + \% \ profit(i)) \right] - 1$$

$$stdev\ \% \ profit = (1 + avg\ \% \ profit) \times$$

$$\times \sqrt{\frac{1}{total\ trades - 1} \sum_{i=1}^{total\ trades} \left(\ln(1 + \% \ profit(i)) - \ln(1 + avg\ \% \ profit) \right)^2}$$

- среднее значение и с.к.о. наибольших снижений торгового счета (в деньгах)

$$avg\ net\ drawdown = \frac{1}{total\ trades} \sum_{i=1}^{total\ trades} net\ drawdown(i)$$

$$stdev\ net\ drawdown =$$

$$= \sqrt{\frac{1}{total\ trades - 1} \sum_{i=1}^{total\ trades} (net\ drawdown(i) - avg\ net\ drawdown)^2}$$

- среднее значение и с.к.о. наибольших снижений торгового счета (в %)

$$avg\ \% drawdown = \left(\prod_{i=1}^{total\ trades} (1 + \% drawdown(i)) \right)^{1/total\ trades} - 1 =$$

$$= \exp \left[\frac{1}{total\ trades} \sum_{i=1}^{total\ trades} \ln(1 + \% drawdown(i)) \right] - 1$$

$$stdev\ \% drawdown = (1 + avg\ \% drawdown) \times$$

$$\times \sqrt{\frac{1}{total\ trades - 1} \sum_{i=1}^{total\ trades} \left(\ln(1 + \% drawdown(i)) - \ln(1 + avg\ \% drawdown) \right)^2}$$

Показатели МТС, характеризующие только прибыльные и только убыточные сделки, вычисляются аналогичным образом.

13.10. Математическое ожидание дохода сделки.

Важнейшим показателем, характеризующим качество МТС, является математическое ожидание дохода отдельной сделки. У прибыльной системы эта величина больше нуля. Задача состоит в том, чтобы по выборке сделок оценить математическое ожидание дохода и убедиться в том, что полученная оценка положительна и значимо отличается от нуля. Выборками случайных величин, на основе которых можно рассчитать выборочную среднюю и выборочное с.к.о. являются:

- в денежном выражении *net profit*,
- в процентах *% profit*.

Будем считать, что величина торгового счета не может упасть ниже нуля. Следовательно убыток по сделке не может быть меньше

торгового счета перед проведением сделки. С другой стороны, прибыль по сделке может быть неограничено большой. Значит плотность вероятности дохода и в денежном и в процентном выражении имеет положительную асимметрию. Для проверки гипотезы о величине математического ожидания дохода отдельной сделки правильное будет перейти к случайной величине $x = \ln(1 + \% \text{ profit})$.

Пусть случайная величина x имеет математическое ожидание μ и генеральную дисперсию σ^2 . Оценками математического ожидания и дисперсии по выборке (x_1, x_2, \dots, x_N) будут выборочная средняя и выборочная дисперсия:

$$\bar{X} = \frac{1}{N} \sum_{k=1}^N x_k \quad \bar{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{X})^2$$

где $N = \text{total trades}$.

При достаточно большом числе сделок доверительный интервал для μ , характеризующийся доверительной вероятностью P , задается в виде:

$$\bar{X} - t_{1-q/2, \nu} \frac{\bar{\sigma}}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{1-q/2, \nu} \frac{\bar{\sigma}}{\sqrt{N}}$$

где $\nu = N - 1$, $q = 1 - P$.

Гипотеза о том, что оценка математического ожидания дохода отдельной сделки больше нуля формулируется в виде:

$$H_0: \bar{X} = 0$$

$$H_1: \bar{X} > 0$$

Проверка подобных гипотез подробно рассмотрена ранее в этой книге. Для более жесткой проверки гипотезы выборку (x_1, x_2, \dots, x_N) усекают справа, то есть из нее исключают значения, превосходящие правую границу доверительного интервала, после чего пересчитывают величины \bar{X} и $\bar{\sigma}$.

При изучении математического ожидания дохода отдельной сделки полезно рассмотреть вопрос о том, что определяет эту величину. Вернемся к формуле для вычисления среднего значения дохода сделок:

$$avg\ net\ profit = \frac{1}{total\ trades} \sum_{i=1}^{total\ trades} net\ profit\ (i)$$

Эту формулу можно записать в другом виде, разделив сделки на прибыльные и убыточные:

$$avg\ net\ profit = \frac{win\ trades \times avg\ net\ win - loss\ trades \times |avg\ net\ loss|}{total\ trades}$$

Так как

$$\frac{win\ trades}{total\ trades} = win\ trades\ \%$$

$$\frac{loss\ trades}{total\ trades} = loss\ trades\ \% = 1 - win\ trades\ \%$$

то

$$avg\ net\ profit = win\ trades\ \% \times avg\ net\ win - (1 - win\ trades\ \%) \times |avg\ net\ loss|$$

Из последней формулы следует, что система может быть прибыльной либо за счет увеличения процента прибыльных сделок, либо за счет увеличения отношения средней прибыли выигрышных сделок к среднему убытку проигрышных сделок. У прибыльной системы среднее значение дохода больше нуля, то есть отношение средней прибыли выигрышных сделок к среднему убытку проигрышных сделок соотносится с процентом выигрышных сделок следующим образом:

$$win\ trades\ \% > \frac{|avg\ net\ loss|}{avg\ net\ win + |avg\ net\ loss|}$$

или

$$\frac{avg\ net\ win}{|avg\ net\ loss|} > \frac{1 - win\ trades\ \%}{win\ trades\ \%}$$

Аналогичное соотношение для процентных показателей можно получить, используя формулу для вычисления среднего значения дохода сделок в %:

$$(1 + avg\ \% profit)^{total\ trades} = (1 + avg\ \% win)^{win\ trades} \times (1 + avg\ \% loss)^{loss\ trades}$$

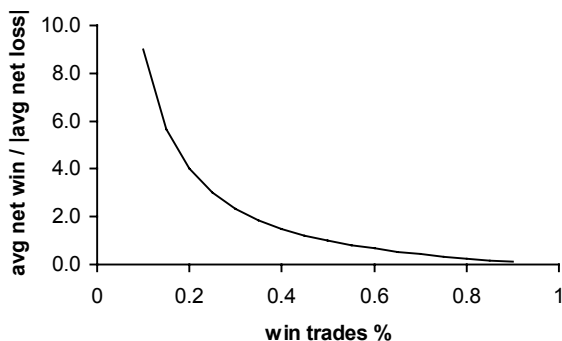
Из этого выражения следует, что

$$\ln(1 + \text{avg \% profit}) = \text{win trades \%} \times \ln(1 + \text{avg \% win}) - (1 - \text{win trades \%}) \times |\ln(1 + \text{avg \% loss})|$$

Для прибыльной системы выполняется условие

$$\frac{\ln(1 + \text{avg \% win})}{|\ln(1 + \text{avg \% loss})|} > \frac{1 - \text{win trades \%}}{\text{win trades \%}}$$

Связь между процентом выигрышных сделок и отношением средней прибыли выигрышных сделок к среднему убытку проигрышных сделок можно представить графически.



Линия на графике соответствует системам с нулевой доходностью. Прибыльные системы находятся выше этой линии, причем чем выше, тем больше у них запас прочности, то есть больше вероятность того, что МТС будет продолжать оставаться прибыльной в реальной торговле.

Соотношение между процентом выигрышных сделок и отношением средней прибыли выигрышных сделок к среднему убытку проигрышных сделок можно ужесточить, используя показатели рассеяния:

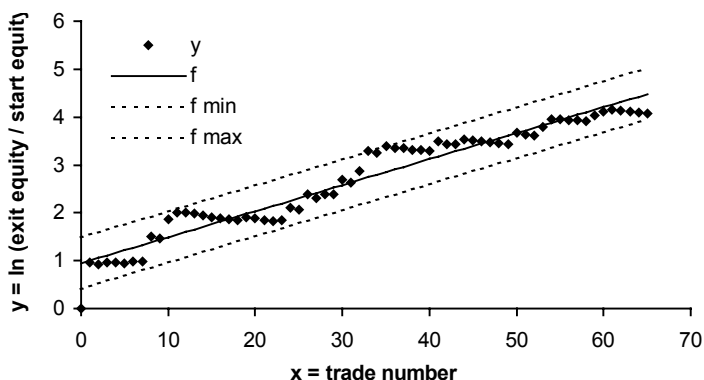
$$\frac{\text{avg net win}}{|\text{avg net loss}|} - \delta \times \left(\frac{\text{stdev net win}}{|\text{avg net loss}|} + \frac{\text{avg net win} \times \text{stdev net loss}}{|\text{avg net loss}|^2} \right) > \frac{1 - \text{win trades \%}}{\text{win trades \%}}$$

где δ - неотрицательное число, характеризующее запас прочности МТС - чем больше δ , тем выше запас прочности. У хороших систем последнее неравенство справедливо при $\delta \geq 0.5$.

13.11. Кумулятивная кривая дохода сделок.

Кумулятивная кривая дохода сделок показывает изменение торгового счета от сделки к сделке. При оценке качества торговой системы полезно изучить эту кривую в полулогарифмическом масштабе. Для исключения влияния величины начальных инвестиций кривую можно нормировать. После этого полученная зависимость исследуется с применением регрессионного анализа. Зависимость величины торгового счета от номера сделки можно получить непосредственно из отчета о сделках.

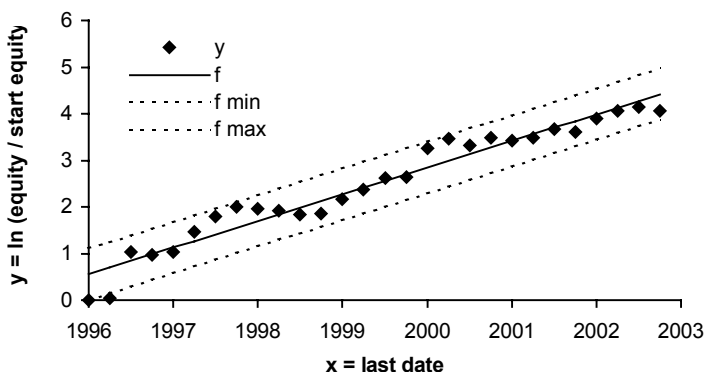
Рассмотрим результаты работы МТС, тестирование которой проводилось на временном ряде цен закрытия по индексу РТС в период с января 1996 г. по сентябрь 2002 г. За это время система совершила 65 сделок, то есть объем выборки равен 66: результаты после 65 сделок + результат до первой сделки (начальные инвестиции). На рисунке изображен логарифм эмпирической нормированной кумулятивной кривой дохода сделок (y), линейная аппроксимация (f) и 95%-ный доверительный интервал линии регрессии.



Показателем, характеризующим доходность МТС, является угол наклона линии регрессии к оси абсцисс. Чем выше угол наклона, тем более доходна МТС. Риск МТС характеризует необъ

ясненная дисперсия рассеяния эмпирических данных вокруг линии регрессии. Чем выше необъясненная дисперсия, тем больше разброс эмпирических точек, то есть выше риск системы. Отношение тангенса угла наклона линии регрессии к величине необъясненного с.к.о. является сводным показателем, характеризующим и доходность и риск системы.

Кумулятивная кривая дохода сделок показывает изменение торгового счета от сделки к сделке. Для анализа поведения счета *во времени* используют сгруппированный отчет о величине торгового счета. Изучение поведения счета по укрупненным периодам времени полностью аналогично изучению кумулятивной кривой дохода сделок. Для той же механической системы на рисунке изображен логарифм эмпирической нормированной величины торгового счета на конец каждого квартала на периоде тестирования (y), линейная аппроксимация (f) и 95%-ный доверительный интервал линии регрессии. Как правило, для анализа линии торгового счета выбираются месячные или квартальные данные.



13.12. Вероятность получения убытка в серии последовательных сделок.

В этом параграфе будет показано, как на основании показателей МТС оценить вероятность получения убытка в серии последовательных сделок.

Для упрощенного расчета вероятности убытка используем три показателя МТС, которые приведены в сводном отчете:

win trades % - процент прибыльных сделок системы,

avg %win - средняя величина выигрыша (%),

avg %loss - средняя величина проигрыша (%).

Введем обозначения:

N - заданная длина серии сделок,

n - количество выигрышных сделок в серии,

$(N - n)$ - количество проигрышных сделок в серии,

p - вероятность выигрыша ($p \equiv \text{win trades \%}$),

total %profit - доход по итогам серии сделок.

Будем приближенно считать, что все выигрышные сделки будут приносить одинаковый доход *avg %win*, а все проигрышные сделки будут приносить одинаковый убыток *avg %loss*. Тогда, если задано количество сделок и вероятность выигрыша, то доход является функцией от числа выигрышных сделок и равен

$$\text{total \%profit}(n) = (1 + \text{avg \%win})^n \times (1 - |\text{avg \%loss}|)^{N-n} - 1$$

Вероятность появления в серии определенного числа выигрышных сделок описывается биномиальным распределением:

$$\text{Prob}(n) = \frac{N!}{n! (N - n)!} p^n (1 - p)^{N-n} \quad n = 0, 1, \dots, N$$

Количество всех возможных комбинаций числа выигрышных и числа проигрышных сделок в серии длиной N будет равно $N + 1$. Для всех этих комбинаций необходимо рассчитать величину дохода *total %profit(n)* и соответствующую ей вероятность *Prob(n)*. Тогда вероятность убытка можно найти как:

$$\text{Prob loss} = \sum_{n=0}^N \text{Prob}(n)$$

где соответствующее слагаемое входит в сумму при условии, что *total %profit(n) ≤ 0*.

Приведем пример такого расчета для торговой системы при длине серии последовательных сделок равной $N = 20$. Пусть величина показателей системы составляет

$$p \equiv \text{win trades \%} = 45\%$$

$$\text{avg \%win} = 8\% \quad \text{avg \%loss} = -5\%$$

Эта система имеет положительное математическое ожидание дохода в расчете на одну сделку

$$\text{avg \%profit} = (1 + \text{avg \%win})^p \times (1 - |\text{avg \%loss}|)^{1-p} - 1$$

$$\text{avg \%profit} = 0.64\%$$

В приведенной ниже таблице содержатся все возможные комбинации числа выигрышных и числа проигрышных сделок, а также соответствующие этим комбинациям величины дохода по итогам серии сделок и вероятности.

<i>n</i>	<i>(N - n)</i>	<i>total \%profit</i>	<i>Prob</i>	<i>Prob loss</i>
0	20	-64.15%	0.0006%	0.0006%
1	19	-59.25%	0.0105%	0.0105%
2	18	-53.67%	0.0816%	0.0816%
3	17	-47.33%	0.4006%	0.4006%
4	16	-40.12%	1.3930%	1.3930%
5	15	-31.93%	3.6471%	3.6471%
6	14	-22.61%	7.4600%	7.4600%
7	13	-12.02%	12.2072%	12.2072%
8	12	0.02%	16.2300%	
9	11	13.70%	17.7055%	
10	10	29.26%	15.9349%	
11	9	46.95%	11.8524%	
12	8	67.06%	7.2731%	
13	7	89.92%	3.6620%	
14	6	115.91%	1.4981%	
15	5	145.46%	0.4903%	
16	4	179.05%	0.1254%	
17	3	217.23%	0.0241%	
18	2	260.64%	0.0033%	
19	1	309.99%	0.0003%	
20	0	366.10%	0.00001%	
<i>ИТОГО</i>			<i>100%</i>	<i>25%</i>

В результате получаем, что вероятность убытка в серии сделок $\text{Prob loss} = 25\%$.

В данном случае система с положительным математическим ожиданием дохода после достаточно длинной серии сделок с вероятностью 25% принесет убыток. Можно привести другие примеры, где МТС с отрицательным математическим ожиданием после серии сделок с достаточно высокой вероятностью приносит прибыль. То есть при биржевой торговле в силу естественных законов статистики правильные решения не всегда сопровождаются прибылью, а неправильные - убытком.

Следует помнить, что приведенная выше оценка вероятности убытка после серии сделок строилась на том, что все выигрышные сделки приносят одинаковый доход $avg \%win$, а все проигрышные сделки приносят одинаковый убыток $avg \%loss$. Это является достаточно грубым приближением, которое не учитывает разброс результатов конкретной сделки. Более точная оценка вероятности убытка основана на многократном численном моделировании результатов серии сделок по методу Монте-Карло. Принципиальная схема такого алгоритма имеет вид:

- 1) Задание входных данных
 - 1.1) Из отчета о сделках массив значений доходов сделок

$$\{\%profit(k)\}$$

$$k = 1, \dots, total\ trades$$
 - 1.2) Количество розыгрышей M (чем больше розыгрышей, тем достовернее результат).
 - 1.3) Длина серии сделок N .
- 2) Вычисление вспомогательного массива $\{x(k)\}$

$$x(k) = \ln(1 + \%profit(k))$$

$$k = 1, \dots, total\ trades$$
- 3) Вычисление в табличном виде гистограммы плотности вероятности значений величины x (методика подробно изложена в главе 6).
- 4) Вычисление в табличном виде интегральной функции распределения значений величины x на основании полученной в предыдущем пункте гистограммы.
- 5) Задаем стартовое значение номера текущего розыгрыша $m = 0$.

- 6) Задаем стартовые значения количества розыгрышей, приводящих к убытку $N\ loss = 0$
- 7) Номер текущего розыгрыша $m = m + 1$
- 8) Проводим отдельный розыгрыш результатов сделок:
 - 8.1) Генерируем набор случайных чисел в количестве N , равномерно распределенных в интервале от 0 до 1.
 - 8.2) Из равномерно распределенного набора случайных чисел с помощью полученной на шаге 4 интегральной функции получаем набор случайных чисел, распределенных как величина x . Обозначим этот набор как $\{y(i)\}$, $i = 1, \dots, N$.
 - 8.3) Находим сумму массива случайных чисел y

$$S = \sum_{i=1}^N y(i)$$
 - 8.4) Находим текущее значение количества розыгрышей, приводящих к убытку:
если $S \leq 0$, то $N\ loss = N\ loss + 1$
- 9) Если номер текущего розыгрыша m меньше, чем общее число розыгрышей M , то переходим на шаг 7.
- 10) После того, как сделаны все розыгрыши (то есть $m = M$), вычисляем вероятность убытка

$$Prob\ loss = \frac{N\ loss}{M}$$

13.13. Вероятность разорения в серии последовательных сделок.

Вероятность разорения - это вероятность того, что в серии последовательных сделок по сигналам МТС величина убытков в силу естественных законов статистики превысит заранее заданное критическое значение. Это может привести к остановке торговли и ошибочному отказу от на самом деле прибыльной МТС.

Аналитический расчет вероятности разорения связан со значительными трудностями ввиду того, что получение критического убытка зависит не только от показателей системы (процента прибыльных сделок, средней величины выигрыша, средней величины проигрыша), но и от очередности прибыльных и убыточных сделок. В результате неблагоприятной последова-

тельности сделок при работе с прибыльной системой разорение может наступить раньше, чем система докажет свое статистическое преимущество.

Так как критический убыток может быть получен до окончания серии сделок, то аналитический расчет вероятности разорения приводит к необходимости учета не только всех возможных сочетаний числа прибыльных и числа убыточных сделок в серии, но и анализу для каждого такого сочетания всех возможных последовательностей сделок. Это весьма трудоемкая задача, особенно для длинных серий.

В данном случае вычисление вероятности разорения гораздо проще можно провести численно путем многократного моделирования результатов серии сделок по методу Монте-Карло. Изложим принципиальную схему алгоритма:

- 1) Задание входных данных
 - 1.1) Из отчета о сделках массив значений доходов сделок

$$\{\%profit(k)\}$$

$$k = 1, \dots, total\ trades$$
 - 1.2) Количество розыгрышей M (чем больше розыгрышей, тем достовернее результат).
 - 1.3) Длина серии сделок N .
 - 1.4) Критическое значение убытка $\% max\ loss$.
- 2) Вычисление вспомогательного массива $\{x(k)\}$

$$x(k) = \ln(1 + \%profit(k))$$

$$k = 1, \dots, total\ trades$$
- 3) Вычисление в табличном виде гистограммы плотности вероятности значений величины x (методика подробно изложена в главе 6).
- 4) Вычисление в табличном виде интегральной функции распределения значений величины x на основании полученной в предыдущем пункте гистограммы.
- 5) Задаем стартовое значение номера текущего розыгрыша $m = 0$.
- 6) Задаем стартовые значения количества розыгрышей, приводящих к разорению $N\ max\ loss = 0$
- 7) Номер текущего розыгрыша $m = m + 1$

- 8) Проводим отдельный розыгрыш результатов сделок:
 - 8.1) Генерируем набор случайных чисел в количестве N , равномерно распределенных в интервале от 0 до 1.
 - 8.2) Из равномерно распределенного набора случайных чисел с помощью полученной на шаге 4 интегральной функции получаем набор случайных чисел, распределенных как величина x . Обозначим этот набор как $\{y(i)\}$, $i = 1, \dots, N$.
 - 8.3) Задаем стартовое значение суммы (нарастающим итогом) массива случайных чисел y : $S = 0$
 - 8.4) Задаем стартовое значение номера случайного числа из массива y : $i = 0$
 - 8.5) Номер текущего случайного числа из массива y : $i = i + 1$
 - 8.6) Находим текущее значение суммы массива случайных чисел y : $S = S + y(i)$
 - 8.7) Проверяем, наступило ли разорение:
если $S \leq \ln(1 - \% \text{ max loss } |)$
то разорение достигнуто, поэтому находим текущее значение количества розыгрышей, приводящих к разорению $N \text{ max loss} = N \text{ max loss} + 1$
и переходим на шаг 7.
 - 8.8) Если номер текущего случайного числа i меньше, чем длина серии N , то переходим на шаг 8.5. В противном случае переходим на шаг 9.
- 9) Если номер текущего розыгрыша m меньше, чем общее число розыгрышей M , то переходим на шаг 7.
- 10) После того, как сделаны все розыгрыши (то есть $m = M$), вычисляем вероятность разорения

$$\text{Prob max loss} = \frac{N \text{ max loss}}{M}$$

14. УПРАВЛЕНИЕ КАПИТАЛОМ

14.1. Введение.

Управление капиталом - это набор правил, определяющих объем открываемых позиций в момент поступления соответствующих сигналов от механической торговой системы. Так как метод управления капиталом непосредственно влияет на динамику торгового счета, то при его выборе нужно четко определить свои инвестиционные цели.

Технически задача сводится к тому, чтобы задать такой алгоритм вычисления доли участвующего в конкретной сделке капитала, чтобы максимизировать один из показателей динамики торгового счета. Этими показателями могут быть средний доход на одну сделку, соотношение дохода и риска сделок, средний прирост торгового счета по фиксированным промежуткам времени (например по месяцам) и т.д.

Как и торговая система, метод управления капиталом должен быть тщательным образом протестирован. При этом требования к нему схожи с требованиями к торговой системе:

- небольшое количество оптимизируемых параметров,
- устойчивость в области оптимальности параметров,
- должно существовать по крайней мере несколько активов, на которых совокупность из торговой системы и метода управления капиталом имеет удовлетворительные результаты без повторной оптимизации.

14.2. Ограничение суммы убытка в сделке.

Пусть торговая система дала сигнал о покупке актива по цене *enter price*, причем величина риска по сделке составляет *%risk*, то есть в случае движения против открытой позиции сделка должна быть закрыта по цене

$$exit\ price = enter\ price \times (1 - \%risk).$$

Рассмотрим вопрос о том, как выбрать долю участвующего в сделке капитала таким образом, чтобы в случае неблагоприятного развития ситуации убыток не превысил заранее заданного значения. Размер капитала до и после проигрышной сделки и величина максимального убытка (в деньгах) связаны соотношением

$$\text{max net loss} = \text{exit equity} - \text{enter equity}$$

Обозначим как α долю участвующего в сделке капитала, причем $0 < \alpha < 1$. Тогда количество покупаемых ценных бумаг (лотов) равно

$$\text{volume} = \frac{\alpha \times \text{enter equity}}{\text{enter price} \times (1 + \% \text{comission})}$$

После закрытия сделки величина капитала будет равна

$$\text{exit equity} = (1 - \alpha) \times \text{enter equity} +$$

$$+ \text{exit price} \times (1 - \% \text{comission}) \times \text{volume}$$

Подставляя в последнюю формулу написанные ранее соотношения после несложных преобразований получаем выражение для доли участвующего в сделке капитала

$$\alpha = \frac{|\text{max net loss}|}{\text{enter equity}} \times \frac{1 + \% \text{comission}}{2 \times \% \text{comission} + \% \text{risk} \times (1 - \% \text{comission})}$$

При определенном соотношении между *enter equity*, *max net loss* и *%risk* вычисленная по этой формуле величина α может оказаться больше 1. В этом случае в данной сделке участвует весь капитал. Выражение для количества покупаемых бумаг (лотов) имеет вид

$$\text{volume} = \text{ЦЕЛОЕ} \left(\frac{\alpha \times \text{enter equity}}{\text{enter price} \times (1 + \% \text{comission})} \right)$$

Данный метод управления капиталом имеет очень простой смысл - ограничить предельно допустимый убыток по конкретной сделке не ограничивая при этом возможную прибыль. Оптимизация метода проводится по величине предельно допустимого убытка *max net loss*.

Основным недостатком этого метода является отсутствие адаптации к текущей величине торгового счета. При существенном изменении капитала относительно начальных инвестиций величину предельно допустимого убытка нужно пересматривать и заново оптимизировать.

14.3. Ограничение процента убытка в сделке.

В отличие от рассмотренного в предыдущем параграфе, метод риска фиксированным процентом капитала в каждой сделке

автоматически адаптируется к текущей величине торгового счета.

Размер капитала до и после проигрышной сделки и величина максимального убытка (в %) связаны соотношением

$$\max \% loss = \frac{\text{exit equity}}{\text{enter equity}} - 1$$

Проведя выкладки, аналогичные сделанным в предыдущем параграфе, можно получить выражение для доли участвующего в сделке капитала

$$\alpha = |\max \% loss| \times \frac{1 + \% \text{comission}}{2 \times \% \text{comission} + \% \text{risk} \times (1 - \% \text{comission})}$$

При определенном соотношении между $\max \% loss$ и $\% \text{risk}$ вычисленная по этой формуле величина α может оказаться больше 1. В этом случае в данной сделке участвует весь капитал. Выражение для количества покупаемых бумаг (лотов) имеет вид

$$\text{volume} = \text{ЦЕЛОЕ} \left(\frac{\alpha \times \text{enter equity}}{\text{enter price} \times (1 + \% \text{comission})} \right)$$

Этот метод позволяет автоматически реинвестировать прибыль выигрышной МТС. С другой стороны, в случае попадания в полосу убыточности, величина предельно допустимых потерь в денежном выражении постоянно уменьшается после каждой проигрышной сделки.

14.4. Максимизация средней величины дохода МТС.

Рассмотрим формулу для вычисления среднего значения дохода сделок (в %) для случая, когда в каждой сделке участвует весь капитал:

$$\text{avg } \% \text{profit} = \left(\prod_{i=1}^{\text{total trades}} (1 + \% \text{profit}(i)) \right)^{1/\text{total trades}} - 1$$

Если в каждой сделке участвует только доля капитала α , причем $0 < \alpha < 1$, и эта доля одинакова для всех сделок, то

$$\text{avg } \% \text{profit}(\alpha) = \left(\prod_{i=1}^{\text{total trades}} (1 + \alpha \times \% \text{profit}(i)) \right)^{1/\text{total trades}} - 1$$

В отдельных случаях выбором соответствующего значения α можно максимизировать величину $avg \%profit(\alpha)$ так, чтобы $avg \%profit(\alpha) > avg \%profit$.

В общем случае для произвольной последовательности сделок задача решается численно путем перебора всех значений α с небольшим шагом изменения (например 0.01). Однако, так как эта операция достаточно трудоемка, то хотелось бы найти простые характеристики торговых систем, которые бы определяли возможность такой оптимизации.

Вернемся к формуле среднего значения дохода сделок. Ее можно записать, используя средние значения выигрышных и проигрышных сделок, а также процент выигрышных сделок для случая, когда в каждой сделке участвует весь капитал:

$$1 + avg \%profit =$$

$$= (1 + avg \%win)^{win\ trades\ \%} \times (1 - |avg \%loss|)^{1 - win\ trades\ \%}$$

Будем приближенно считать, что

$$1 + avg \%profit(\alpha) \approx$$

$$\approx (1 + \alpha \times avg \%win)^{win\ trades\ \%} \times (1 - \alpha \times |avg \%loss|)^{1 - win\ trades\ \%}$$

Последнее равенство не является строгим, однако для оно вполне пригодно для оценочных вычислений оптимальной величины α .

Введем функцию $S(\alpha) = \ln(1 + avg \%profit(\alpha))$. Так как натуральный логарифм является монотонно возрастающей функцией, то максимум функции $S(\alpha)$ соответствует максимуму функции $avg \%profit(\alpha)$.

Запишем $S(\alpha)$ в явном виде:

$$S(\alpha) = win\ trades\ \% \times \ln(1 + \alpha \times avg \%win) + \\ + (1 - win\ trades\ \%) \times \ln(1 - \alpha \times |avg \%loss|)$$

Возьмем производную $S(\alpha)$ по α :

$$\frac{dS(\alpha)}{d\alpha} = \frac{win\ trades\ \% \times avg \%win}{1 + \alpha \times avg \%win} - \frac{(1 - win\ trades\ \%) \times |avg \%loss|}{1 - \alpha \times |avg \%loss|}$$

Приравняв производную к нулю, получим формулу для вычисления оптимального значения α (в том случае, если экстремум существует):

$$\alpha_{opt} = \frac{win\ trades\ \% \times avg\ \%win - (1 - win\ trades\ \%) \times |avg\ \%loss|}{avg\ \%win \times |avg\ \%loss|}$$

Так как $\alpha_{opt} > 0$ то $win\ trades\ \% > win\ trades\ \% min$

где

$$win\ trades\ \% min = \frac{|avg\ \%loss|}{avg\ \%win + |avg\ \%loss|}$$

Если последнее неравенство не соблюдается, то экстремума не существует и $\alpha_{opt} = 0$, то есть по такой МТС торговать нельзя.

Так как $\alpha_{opt} < 1$, то $win\ trades\ \% < win\ trades\ \% max$

где

$$win\ trades\ \% max = \frac{|avg\ \%loss|}{avg\ \%win + |avg\ \%loss|} \cdot (1 + avg\ \%win)$$

Если последнее неравенство не соблюдается, то экстремума не существует и $\alpha_{opt} = 1$, то есть по такой МТС нужно торговать всем капиталом в каждой сделке.

Приведем несколько примеров решения задачи оптимизации доходности МТС.

Неоптимизируемая система с положительным математическим ожиданием дохода

Результаты тестирования МТС:

$$win\ trades\ \% = 50\%$$

$$avg\ \%win = 15\% \quad avg\ \%loss = -10\%$$

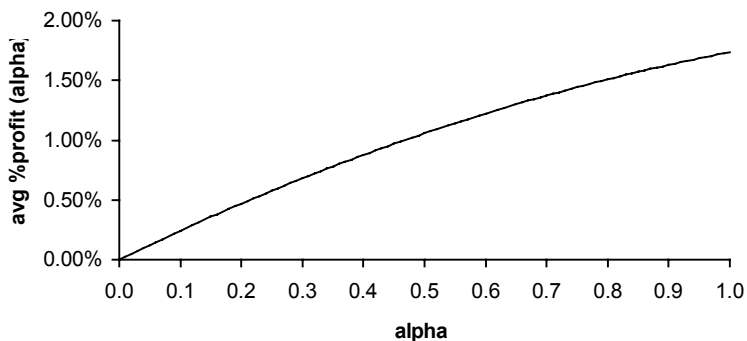
Неоптимизированная прибыль на сделку:

$$avg\ \%profit = 1.73\%$$

Границы области оптимизации:

$$win\ trades\ \% min = 40\% \quad win\ trades\ \% max = 46\%$$

Так как $win\ trades\ \% > win\ trades\ \% max$, то функция $avg\ \%profit(\alpha)$ не имеет экстремума и монотонно возрастает при увеличении α .



По такой МТС нужно торговать всем капиталом в каждой сделке.

Неоптимизируемая система с отрицательным математическим ожиданием дохода

Результаты тестирования МТС:

$win\ trades\ \% = 35\%$

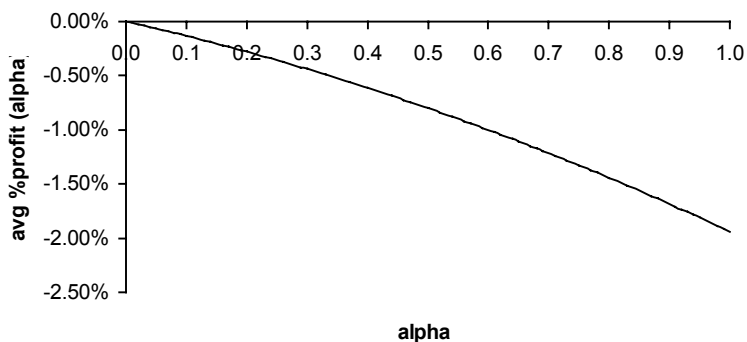
$avg\ \%win = 15\%$ $avg\ \%loss = -10\%$

Неоптимизированная прибыль на сделку:

$avg\ \%profit = -1.94\%$

Границы области оптимизации:

$win\ trades\ \%min = 40\%$ $win\ trades\ \%max = 46\%$



Так как $win\ trades\ \% < win\ trades\ \%min$, то функция $avg\ \%profit(\alpha)$ не имеет экстремума и монотонно убывает при увеличении α . По такой МТС торговать нельзя.

Оптимизируемая система с положительным математическим ожиданием дохода

Результаты тестирования МТС:

$win\ trades\ \% = 44\%$

$avg\ \%win = 15\%$ $avg\ \%loss = -10\%$

Неоптимизированная прибыль на сделку:

$avg\ \%profit = 0.25\%$

Границы области оптимизации:

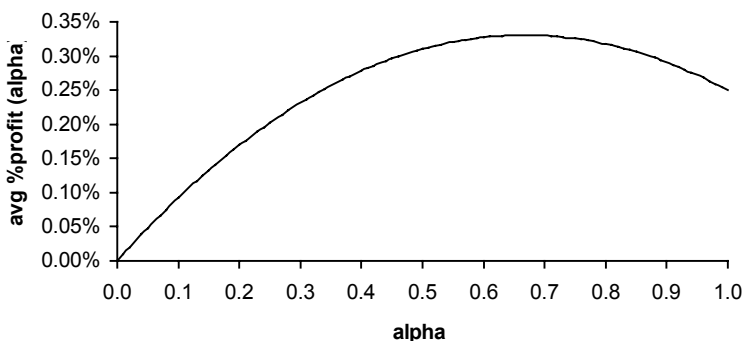
$win\ trades\ \%min = 40\%$ $win\ trades\ \%max = 46\%$

В данном случае возможна оптимизация системы, так как

$win\ trades\ \%min < win\ trades\ \% < win\ trades\ \%max$.

Функция $avg\ \%profit(\alpha)$ имеет экстремум в точке

$\alpha_{opt} = 0.67$, при этом $avg\ \%profit(\alpha_{opt}) = 0.33\%$.



Мы получили, что в результате оптимизации повышается доходность прибыльной системы.

Оптимизируемая система с отрицательным математическим ожиданием дохода

Результаты тестирования МТС:

$win\ trades\ \% = 42.5\%$

$avg\ \%win = 15\%$ $avg\ \%loss = -10\%$

Неоптимизированная прибыль на сделку:

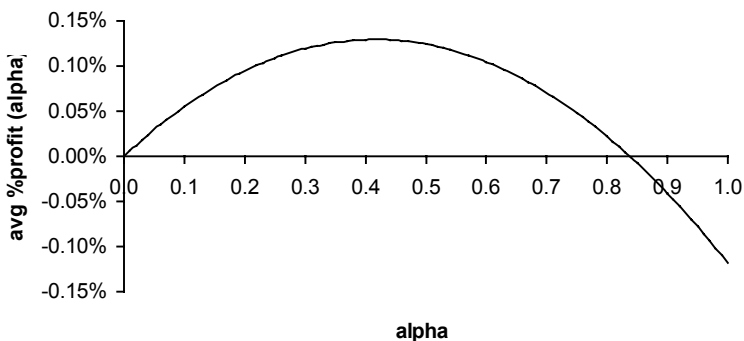
$avg\ \%profit = -0.12\%$

Границы области оптимизации:

$win\ trades\ \%min = 40\%$ $win\ trades\ \%max = 46\%$

В данном случае возможна оптимизация системы, так как $win\ trades\ \%min < win\ trades\ \% < win\ trades\ \%max$.

Функция $avg\ \%profit(\alpha)$ имеет экстремум в точке $\alpha_{opt} = 0.42$, при этом $avg\ \%profit(\alpha_{opt}) = 0.13\%$.



Мы получили, что в результате оптимизации убыточная система становится прибыльной.

14.5. Оптимизация соотношения дохода и риска МТС.

Для оптимизации соотношения дохода и риска МТС нужно найти такую долю α участвующего в сделке капитала, которая максимизировала бы следующую функцию:

$$Q(\alpha) = (1 + \alpha \times avg\ \%win)^{win\ trades\ \%} \times (1 - \alpha \times |avg\ \%loss|)^{1 - win\ trades\ \%} \times (1 - \alpha \times \%risk) - 1$$

В данной формуле $\%risk$ - это один из показателей риска МТС, который в зависимости от предпочтений конкретного трейдера может принимать, например, следующие значения:

$$\%risk = stdev\ \%profit$$

$$\%risk = |avg\ \%drawdown|$$

$$\%risk = |max\ \%drawdown|$$

Перейдем к функции $S(\alpha) = \ln(1 + Q(\alpha))$, экстремум которой, если он существует, совпадает с экстремумом $Q(\alpha)$.

$$S(\alpha) = \text{win trades \%} \times \ln(1 + \alpha \times \text{avg \%win}) + \\ + (1 - \text{win trades \%}) \times \ln(1 - \alpha \times |\text{avg \%loss}|) + (1 - \alpha \times \%risk)$$

Взяв производную $S(\alpha)$ по α и приравняв ее к нулю получим:

$$\alpha_{opt} = \frac{a_1 - \sqrt{a_1^2 - 4a_0a_2}}{2a_2}$$

где

$$a_0 = \text{win trades \%} \times \text{avg \%win} - (1 - \text{win trades \%}) \times |\text{avg \%loss}| - \\ - \%risk$$

$$a_1 = \text{avg \%win} \times (\%risk + |\text{avg \%loss}|) + \\ + \text{win trades \%} \times \%risk \times (\text{avg \%win} + |\text{avg \%loss}|) - \\ - 2 \times \%risk \times |\text{avg \%loss}|$$

$$a_2 = 2 \times \text{avg \%win} \times |\text{avg \%loss}| \times \%risk$$

Так как $\alpha_{opt} > 0$ то $\text{win trades \%} > \text{win trades \% min}$

где

$$\text{win trades \% min} = \frac{\%risk + |\text{avg \%loss}|}{\text{avg \%win} + |\text{avg \%loss}|}$$

Так как $\alpha_{opt} < 1$, то $\text{win trades \%} < \text{win trades \% max}$

где

$$\text{win trades \% max} = \frac{\%risk + |\text{avg \%loss}| - 2 \times \%risk \times |\text{avg \%loss}|}{\text{avg \%win} + |\text{avg \%loss}|} \times \\ \times \frac{1 + \text{avg \%win}}{1 - \%risk}$$

Если не соблюдается двойное неравенство

$$\text{win trades \% min} < \text{win trades \%} < \text{win trades \% max},$$

то экстремума не существует, то есть по данной методике нельзя формальным образом выбрать наилучшее соотношение дохода и риска.

Приведем пример решения задачи оптимизации соотношения доходности и риска МТС.

Результаты тестирования МТС:

$$win\ trades\ \% = 64\%$$

$$avg\ \%win = 15\%$$

$$avg\ \%loss = -5\%$$

Неоптимизированная прибыль на сделку:

$$avg\ \%profit = 7.36\%$$

Принятая величина риска:

$$\%risk = 7\%$$

Найдем показатели, характеризующие соотношение дохода и риска до оптимизации:

$$(1 + avg\ \%profit) \times (1 - \%risk) - 1 =$$

$$= (1 + 0.0736) \times (1 - 0.07) - 1 = -0.0016$$

$$\frac{avg\ \%profit}{\%risk} = \frac{7.36}{7} = 1.0509$$

Границы области оптимизации:

$$win\ trades\ \%min = 60\%$$

$$win\ trades\ \%max = 70\%$$

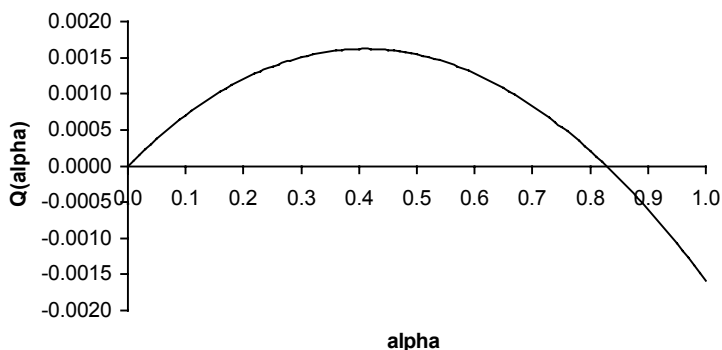
В данном случае возможна оптимизация системы, так как

$$win\ trades\ \%min < win\ trades\ \% < win\ trades\ \%max.$$

Функция $Q(\alpha)$ имеет экстремум в точке $\alpha_{opt} = 0.41$, при этом

$$avg\ \%profit(\alpha_{opt}) = 3.12\%,$$

$$\%risk(\alpha_{opt}) = \%risk \times \alpha_{opt} = 7\% \times 0.41 = 2.87\%$$



Соотношения дохода и риска после оптимизации:

$$Q(\alpha_{opt}) = (1 + avg \%profit(\alpha_{opt})) \times (1 - \%risk(\alpha_{opt})) - 1 = \\ = (1 + 0.0312) \times (1 - 0.0287) - 1 = 0.0016$$

$$\frac{avg \%profit(\alpha_{opt})}{\%risk(\alpha_{opt})} = \frac{3.12}{2.87} = 1.0878$$

Как видно из полученных результатов, после оптимизации довольно существенно снизился доход на одну сделку, однако улучшились соотношения дохода и риска. Кроме того, небольшая величина доли капитала, участвующего в сделках, позволяет чувствовать себя более комфортно на волатильных рынках.

14.6. Анализ соотношения скользящих средних от кумулятивной кривой дохода сделок.

В предыдущих параграфах рассматривались методы оптимизации доли участвующего в сделках капитала в зависимости от показателей МТС (процента прибыльных сделок, средней величины выигрыша, средней величины проигрыша). Однако, не была учтена возможность корреляции между результатами последовательных сделок, то есть возможность возникновения серий из прибыльных или убыточных сделок.

Формально, наличие корреляции в доходах сделок можно выявить, найдя коэффициент автокорреляции динамического ряда $\%profit(trade\ number)$, где $trade\ number$ - номер сделки, $\%profit$ - величина дохода по данной сделке (%). После этого нужно проверить гипотезу о том, значительно ли отличается вычисленная по выборке величина коэффициента автокорреляции от нуля.

Высокая положительная корреляция свидетельствует о том, что за выигрышем чаще следует очередной выигрыш и реже следует проигрыш и наоборот. При высокой отрицательной корреляции за выигрышем чаще следует проигрыш и реже - очередной выигрыш и наоборот.

Наличие корреляции между результатами сделок можно использовать в торговле двумя способами. Во-первых, модификацией решающих правил торговой системы, однако это может привести к ее излишнему переусложнению. Во-вторых, использованием методов управления капиталом, что представляется более предпочтительным. Как правило, если корреляция присутствует, то она по

ложительна, следовательно возможны достаточно длинные серии из прибыльных или убыточных сделок. Очевидно, что в полосе прибыльности нужно увеличивать долю участвующего в сделке капитала, а в полосе убыточности - уменьшать.

При наличии высокой положительной корреляции доходов сделок очень простым и достаточно эффективным способом управления величиной участвующего в сделке капитала является метод, основанный на соотношении скользящих средних различных порядков от размера торгового счета после каждой сделки *exit equity* (*trade number*).

Обозначим как *short line* и *long line* короткую и длинную простую скользящую среднюю от величины *exit equity*.

Если *short line* > *long line*, то в среднем в настоящее время торговая система показывает лучшие результаты, чем в среднем в прошлом. Следовательно, система находится в полосе прибыльности и нужно увеличивать объемы сделок. Самое простое решение - участвовать всем капиталом в каждой сделке.

Если *short line* < *long line*, то в среднем в настоящее время торговая система показывает худшие результаты, чем в среднем в прошлом. Следовательно, система находится в полосе убыточности и нужно уменьшать объемы сделок. Самое простое решение - вообще не открывать позиции.

Однако, возможны ситуации, когда единственная не проведенная прибыльная сделка по упущенной выгоде перекрывает экономию от нескольких не проведенных убыточных сделок. Для исключения таких случаев можно предложить следующую методику:

- 1) Рассчитаем набор случайных величин $\{x_k\}$
где $x = \ln(\text{short line} / \text{long line})$.
- 2) По полученной выборке найдем оценки среднего значения и с.к.о. величины x . Так как эти величины должны пересчитываться после каждой закрытой сделки, то обозначим их как \overline{X}_k и σ_k .
- 3) Если приближенно считать, что случайная величина x подчиняется нормальному распределению, то в качестве доли

участвующего в очередной сделке капитала можно выбрать значение интегральной функции распределения в последней перед планируемой сделкой точке выборки. Используя функции Microsoft Excel можно найти долю капитала по формуле:

$$\alpha_{k+1} = \text{НОРМСТРАСП} \left(\frac{x_k - \overline{X}_k}{\sigma_k} \right)$$

Возможны следующие варианты

$$x_k \ll \overline{X}_k \Rightarrow \alpha_{k+1} \approx 0$$

$$x_k < \overline{X}_k \Rightarrow 0 < \alpha_{k+1} < 0.5$$

$$x_k = \overline{X}_k \Rightarrow \alpha_{k+1} = 0.5$$

$$x_k > \overline{X}_k \Rightarrow 0.5 < \alpha_{k+1} < 1$$

$$x_k \gg \overline{X}_k \Rightarrow \alpha_{k+1} \approx 1$$

Таким образом, при *short line = long line* доля участвующего в сделке капитала будет равна 50%, при *short line > long line* доля будет возрастать от 50% вплоть до 100%, а при *short line < long line* доля будет уменьшаться от 50% вплоть до 0%.

Метод управления капиталом по соотношению скользящих средних может быть оптимизирован по двум параметрам: периоду короткой скользящей средней и периоду длинной скользящей средней.

14.7. Критерий серий.

Одним из важнейших показателей механической торговой системы является доля выигрышных сделок *win trades %*. Однако, эта величина никак не характеризует наличие или отсутствие корреляции между результатами сделок. Возможны следующие варианты:

- 1) Исходы последовательных сделок независимы друг от друга, то есть выигрыши и проигрыши чередуются случайным образом. В этом случае мы имеем нулевую корреляцию между результатами сделок.

- 2) За выигрышем чаще следует выигрыш, за проигрышем – проигрыш, то есть возникают полосы прибыльности и убыточности. В этом случае мы имеем положительную корреляцию между результатами сделок.
- 3) За выигрышем чаще следует проигрыш, за проигрышем – выигрыш. В этом случае мы имеем отрицательную корреляцию между результатами сделок

Рассмотрим результаты последовательных сделок торговой системы. Назовем *серией* несколько следующих подряд прибыльных сделок или несколько следующих подряд убыточных сделок. В случае положительной корреляции количество серий на периоде тестирования будет меньше, чем количество серий при независимом чередовании прибылей и убытков. При отрицательной корреляции ситуация будет обратной. Заметим, что при расчете серий учитывается только знак дохода по сделке, а не его абсолютная величина, при этом сделки с нулевым доходом учитываются как убыточные.

Критерий серий позволяет определить, насколько значимо отличие полученного по выборке сделок эмпирического значения числа серий от ожидаемого числа серий при независимом распределении прибылей и убытков. Введем обозначения:

N - количество сделок ($N = total\ trades$).

p - вероятность прибыльной сделки ($p = win\ trades\ \%$).

R - эмпирическое значение числа серий по выборке сделок.

При достаточно большом N (больше 50) и при отсутствии корреляции между результатами сделок случайная величина R будет подчиняться нормальному распределению с параметрами:

$2p(1-p)N$ - ожидаемое значение числа серий.

$2p(1-p)\sqrt{N}$ - с.к.о. числа серий.

Для приведения случайной величины R к стандартному виду (то есть к величине с нулевым математическим ожиданием и единичной дисперсией) нужно сделать преобразование

$$Z = \frac{R - 2p(1-p)N}{2p(1-p)\sqrt{N}}$$

Полученную таким образом нормированную случайную величину называют Z -счетом. Положительный Z -счет возникает

ет, если количество серий больше ожидаемого, то есть при отрицательной корреляции. Отрицательный Z -счет возникает, если количество серий меньше ожидаемого, то есть при положительной корреляции.

Вероятность того, что стандартная нормальная величина находится в интервале от $-|Z|$ до $|Z|$, или соответствующую величине Z -счета доверительную вероятность, можно вычислить с помощью таблиц Microsoft Excel: $P(Z) = 2 \cdot \text{НОРМСТРАСП}(|Z|) - 1$.

Как правило считают, что при $|Z| \leq 1.96$ ($P(Z) \leq 95\%$), отклонение эмпирического числа серий от ожидаемого незначимо и объясняется случайностью выборки. В этом случае нецелесообразно модифицировать торговую стратегию на основании Z -счета.

Если же мы имеем систему, у которой $|Z| > 1.96$ ($P(Z) > 95\%$), то можно использовать полученную закономерность следующим образом:

- 1) При большом по модулю и положительном по знаку Z -счете за прибыльной сделкой скорее всего последует убыточная, а за убыточной - прибыльная. Следовательно, после выигрышей нужно уменьшать объем следующей сделки, а после проигрышей - увеличивать.
- 2) При большом по модулю и отрицательном по знаку Z -счете за прибыльной сделкой скорее всего последует очередная прибыльная сделка, а за убыточной - очередная убыточная. Следовательно, после выигрышей нужно увеличивать объем следующей сделки, а после проигрышей - уменьшать.

Отметим, что критерий серий может служить фильтром для рассмотренного в предыдущем параграфе метода скользящих средних.

14.8. Увеличение объема выигрывающей позиции.

В этом параграфе будет изложен метод увеличения объема выигрывающей длинной позиции для трендовых систем. Как правило такие торговые системы имеют сравнительно небольшой процент выигрышных сделок и высокое отношение средней прибыли к среднему убытку.

Пусть торговая система дала сигнал о покупке актива в момент времени t_1 по цене P_1 . В этом случае проводится первая покупка, причем сумма сделки равна α_1 процентов от текущей величины свободных денежных средств. Если цена продолжает рост, то в момент времени t_2 делается вторая покупка по цене P_2 на сумму α_2 процентов от текущей величины свободных денежных средств. Нарастивание объема бумаг продолжается до тех пор, пока растет рыночная цена. Заметим, что начиная со второй сделки средняя цена покупки меньше текущей цены актива. При развороте рынка и достижении ценой некоторого критического уровня все позиции закрываются.

Изложим эту методику формальным образом. Введем обозначения:

X_0 - начальная величина свободных денежных средств,

N - общее число проведенных сделок,

P_k - цена, по которой проводится k -я сделка (рыночная цена),

c - комиссия за проведение сделки (%),

X_N - величина свободных денежных средств после N сделок,

$Y_N^{(B)}$ - общие затраты на покупку после N сделок (балансовая стоимость бумаг),

V_N - количество бумаг в портфеле после N сделок,

$P_N^{(B)}$ - средняя цена покупки актива после N сделок (балансовая цена).

После проведения N сделок, где сумма k -й сделки составляет α_k процентов от текущей величины свободных денежных средств, перечисленные выше переменные будут выражаться формулами:

$$X_N = X_0 \cdot \prod_{k=1}^N (1 - \alpha_k)$$

$$Y_N^{(B)} = X_0 - X_N = X_0 \cdot \left(1 - \prod_{k=1}^N (1 - \alpha_k) \right)$$

$$V_N = \frac{X_0}{1+c} \cdot \sum_{k=1}^N \left(\frac{\alpha_k}{P_k} \cdot \prod_{i=1}^{k-1} (1-\alpha_i) \right)$$

$$P_N^{(B)} = \frac{Y_N^{(B)}}{V_N} = (1+c) \cdot \frac{1 - \prod_{k=1}^N (1-\alpha_k)}{\sum_{k=1}^N \left(\frac{\alpha_k}{P_k} \cdot \prod_{i=1}^{k-1} (1-\alpha_i) \right)}$$

Для последующих рассуждений эти формулы нужно разумным образом упростить. Сделаем два допущения:

- 1) Все α_k одинаковы и равны α .
- 2) Все покупки, начиная со второй, происходят при определенном процентном приросте рыночной цены относительно цены предыдущей сделки, то есть

$$P_k = P_1 \cdot (1+\beta)^{k-1}$$

После этих упрощений формулы примут вид:

$$X_N = X_0 \cdot (1-\alpha)^N$$

$$Y_N^{(B)} = X_0 \cdot (1 - (1-\alpha)^N)$$

$$V_N = \frac{\alpha \cdot X_0}{P_1 \cdot (1+c)} \cdot \frac{1+\beta}{\alpha+\beta} \cdot \left(1 - \left(\frac{1-\alpha}{1+\beta} \right)^N \right)$$

$$P_N^{(B)} = P_1 \cdot (1+c) \cdot \frac{1}{\alpha} \cdot \frac{\alpha+\beta}{1+\beta} \cdot \frac{1 - (1-\alpha)^N}{1 - \left(\frac{1-\alpha}{1+\beta} \right)^N}$$

Рассмотрим зависимость этих переменных от количества сделок N на примере. Примем следующие значения входящих в формулы параметров:

$$\alpha = 30\% \quad \beta = 5\% \quad c = 0.2\%$$

$$P_1 = 1$$

$$X_0 = 100\,000$$

Результаты расчетов с этими значениями параметров приведены в таблице:

N	1	2	3	4	5
P_N	1.000	1.050	1.103	1.158	1.216
X_N	70 000	49 000	34 300	24 010	16 807
$Y_N^{(B)}$	30 000	51 000	65 700	75 990	83 193
V_N	29 940	49 900	63 207	72 078	77 992
$P_N^{(B)}$	1.002	1.022	1.039	1.054	1.067

При работе по такой методике закрытие позиций может осуществляться следующим образом:

- 1) Если после первой сделки рынок пошел против открытой позиции, то выход осуществляется по сигналу *stop loss*, причем так как объем сделки сравнительно невелик, то убыток будет небольшим.
- 2) После второй сделки цену выхода можно поставить на границу безубыточности (с учетом комиссии), то есть $P_2^{(exit)} = P_2^{(B)} / (1 - c)$.

Для данного примера $P_2^{(exit)} = 1.0022 / (1 - 0.002) = 1.024$.

После второй сделки, если не произойдут форс-мажорные события, портфель застрахован от убытков при неограниченном потенциале прибыли.

- 3) Начиная с третьей сделки цена выхода может равняться цене входа для предыдущей сделки. Например, цена выхода после пятой сделки равна цене входа по четвертой сделке, то есть 1.158, при этом средняя цена покупки с учетом комиссии равна 1.067.

Если открытие позиций осуществляется частями, то выход происходит всем объемом одновременно. Выход частями может быть целесообразен только в случае, когда разница между текущей рыночной ценой и средней ценой покупки достаточно велика, то есть если входы были сделаны на сильном и продолжительном восходящем тренде.

15. УПРАВЛЕНИЕ РИСКОМ ПОРТФЕЛЯ НА ОСНОВЕ АНАЛИЗА КОВАРИАЦИЙ АКТИВОВ

15.1. Введение.

В этой главе рассматриваются вопросы, связанные с оптимизацией портфеля активов. Изучается влияние корреляции между отдельными парами активов на общий риск портфеля, при этом в качестве меры риска принимается дисперсия (или среднеквадратичное отклонение). Рассказано о том, что такое эффективная диверсификация и как общий риск портфеля, составленного из произвольного количества активов, можно разделить на несистематический (диверсифицируемый) риск и рыночный (недиверсифицируемый) риск. Дано понятие границы эффективности на примере портфеля из двух активов и приведены формулы, которые позволяют выбрать на границе эффективности портфель с минимальным ожидаемым риском и портфель с максимальным отношением ожидаемого дохода к ожидаемому риску. Поставлена задача по оптимизации портфеля из произвольного количества активов с учетом ограничений на состав и веса активов в портфеле (лимитов), и приведен алгоритм поиска решений этой задачи методом Монте-Карло.

15.2. Корреляция активов и риск портфеля.

Результаты решения об инвестировании в тот или иной финансовый инструмент (актив) всегда имеют некоторую неопределенность. В большинстве случаев реально полученный доход от вложения в некий финансовый инструмент не совпадает с ожидаемым доходом по этому инструменту на момент принятия решения об инвестировании, то есть инвестирование - это сфера деятельности, связанная с риском.

Меру рассеяния реально полученных доходов относительно ожидаемого дохода, то есть риск актива, будем характеризовать дисперсией (или среднеквадратичным отклонением) доходов по данному активу.

Рассмотрим случай, когда инвестирование проводится в несколько активов (портфель). Портфель является линейной комбинацией активов, каждый из которых имеет собственное математическое ожидание дохода и дисперсию дохода.

Вспомним формулы для вычисления математического ожидания и дисперсии многомерной случайной величины y , являющейся линейной комбинацией коррелированных случайных величин:

$$y = \sum_{k=1}^N a_k x_k$$

$$\mu_y = \sum_{k=1}^N a_k \mu_k$$

$$\sigma_y^2 = \sum_{k=1}^N a_k^2 \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N a_i a_k \rho_{ik} \sigma_i \sigma_k$$

Поэтому для математического ожидания и дисперсии дохода портфеля активов можно использовать следующие формулы:

$$\mu_y = \sum_{k=1}^N w_k \mu_k$$

$$\sigma_y^2 = \sum_{k=1}^N w_k^2 \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N w_i w_k \rho_{ik} \sigma_i \sigma_k$$

где w - веса активов в портфеле.

В отличие от произвольной линейной комбинации случайных величин, веса активов подчиняются правилу нормирования:

$$\sum_{k=1}^N w_k = 1$$

Следовательно:

- *математическое ожидание дохода портфеля* - это взвешенная сумма математических ожиданий доходов по отдельным активам,
- *риск дохода портфеля* - это взвешенная сумма ковариаций всех пар активов в портфеле, при этом вес каждой ковариации равен произведению весов соответствующей пары активов, а ковариация актива с самим собой является дисперсией данного актива.

15.3. Понижение риска портфеля. Диверсификация.

Из формулы для дисперсии портфеля активов очевидно, что риск портфеля можно разделить на две группы слагаемых.

Слагаемые вида $w_k^2 \sigma_k^2$, которые характеризуют риск отдельных активов, всегда положительны. Следовательно они могут только увеличивать риск портфеля.

Слагаемые вида $w_i w_k \rho_{ik} \sigma_i \sigma_k$, которые характеризуют ковариацию различных пар активов в портфеле, могут быть положительными, равными нулю и отрицательными. Все зависит от величины коэффициента корреляции между активами. Следовательно, эти слагаемые могут уменьшить риск портфеля в целом по сравнению с риском отдельных активов.

Проиллюстрируем этот эффект диверсификации на примере портфеля, состоящего из двух активов. Рассмотрим три случая, в каждом из которых математические ожидания и дисперсии активов одинаковы, а различными являются коэффициенты корреляции между ними:

1) *Активы полностью коррелированы*

В этом случае коэффициент корреляции между активами равен 1. Следовательно для дисперсии и с.к.о. портфеля получаем

$$\sigma_y^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \sigma_1 \sigma_2 = (w_1 \sigma_1 + w_2 \sigma_2)^2$$

$$\sigma_y = w_1 \sigma_1 + w_2 \sigma_2$$

Следовательно, с.к.о. портфеля равно средней взвешенной с.к.о. отдельных активов, то есть понижение риска портфеля от диверсификации отсутствует.

2) *Активы не коррелированы*

Коэффициент корреляции между активами равен 0.

$$\sigma_y^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2$$

$$\sigma_y = \sqrt{w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2}$$

В этом случае с.к.о. портфеля меньше средней взвешенной с.к.о. отдельных активов, то есть присутствует понижение риска портфеля от диверсификации, однако с.к.о. портфеля больше нуля при любых ненулевых дисперсиях отдельных активов.

3) *Активы антикоррелированы*

Коэффициент корреляции между активами равен -1.

$$\sigma_y^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 - 2w_1 w_2 \sigma_1 \sigma_2 = (w_1 \sigma_1 - w_2 \sigma_2)^2$$

$$\sigma_y = |w_1 \sigma_1 - w_2 \sigma_2|$$

В этом случае эффект диверсификации наиболее значителен, с.к.о. портфеля может быть сведено к нулю соответствующим выбором весов отдельных активов в портфеле, то есть может быть получен безрисковый портфель.

Заметим, что во всех трех случаях математическое ожидание дохода портфеля одинаково и равно $\mu_y = w_1 \mu_1 + w_2 \mu_2$.

Можно сказать, что эффективная *диверсификация* - это добавление таких активов в портфель, доходы по которым имеют наименьший коэффициент корреляции с активами, уже имеющимися в портфеле.

Рассмотрим теперь портфель, состоящий из большого количества активов N , каждый из которых входит в портфель с одинаковым весом $1/N$. В случае более 2-х активов невозможно добиться того, чтобы каждая пара активов имела бы коэффициент корреляции, равный -1. Пусть все активы имеют вообще говоря различные конечные дисперсии и каждая пара активов имеет вообще говоря разные коэффициенты корреляции. Тогда формула для дисперсии портфеля примет вид:

$$\sigma_y^2 = \sum_{k=1}^N \frac{1}{N^2} \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N \frac{1}{N^2} \rho_{ik} \sigma_i \sigma_k$$

Преобразуем эту формулу:

$$\sigma_y^2 = \frac{1}{N} \sum_{k=1}^N \frac{\sigma_k^2}{N} + \frac{N-1}{N} \sum_{k=1}^N \sum_{i=k+1}^N \frac{\sigma_{ik}}{N(N-1)/2}$$

Первая сумма - это средняя дисперсия активов, вторая (удвоенная) сумма - это средняя ковариация всех пар различных активов, следовательно:

$$\sigma_y^2 = \frac{1}{N} \cdot \overline{\sigma_k^2} + \frac{N-1}{N} \cdot \overline{\sigma_{ik}}$$

Поэтому, при достаточно большом количестве активов N первым слагаемым можно пренебречь и для дисперсии портфеля можно написать приближенное выражение $\sigma_y^2 \approx \overline{\sigma_{ik}}$, то есть дисперсия портфеля приближенно равна средней ковариации активов.

Таким образом, общий риск портфеля можно разделить на две части:

- *несистематический риск*, определяемый средней дисперсией активов, который может быть исключен путем формирования портфеля из большого количества активов (диверсификацией),
- *систематический (рыночный) риск*, определяемый средней ковариацией пар различных активов, который не может быть исключен путем формирования портфеля из большого количества активов (диверсификацией).

При этом математическое ожидание дохода портфеля равно среднему математическому ожиданию дохода входящих в портфель активов:

$$\mu_y = \sum_{k=1}^N \frac{\mu_k}{N} = \overline{\mu_k}$$

15.4. Граница эффективности.

В предыдущем параграфе было показано, что в случае, когда коэффициент корреляции между активами меньше 1, диверсификация портфеля может улучшить соотношение между ожидаемым доходом и ожидаемым риском. Это связано с тем, что ожидаемый доход портфеля является линейной комбинацией ожидаемых доходов по входящим в портфель активам, а дисперсия портфеля является квадратичной функцией от с.к.о. входящих в портфель активов.

При заданных математических ожиданиях и дисперсиях активов, а также коэффициентах корреляции между различными парами активов, путем оптимизации весов активов в портфеле можно добиться улучшения соотношения между доходом и риском портфеля.

Рассмотрим портфель, состоящий из двух активов, с коэффициентом корреляции между ними равным $\rho = 0.5$:

- 1-й актив

ожидаемый доход $\mu_1 = 10\%$

с.к.о. ожидаемого дохода (риск) $\sigma_1 = 15\%$

- 2-й актив

ожидаемый доход $\mu_2 = 13\%$

с.к.о. ожидаемого дохода (риск) $\sigma_2 = 16\%$

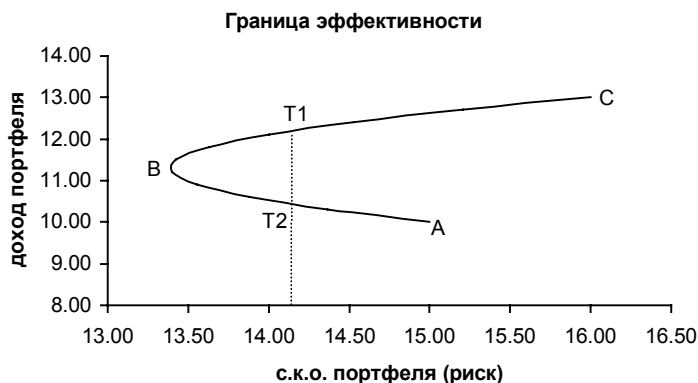
Ожидаемый доход и среднеквадратичное отклонение портфеля из двух активов вычисляются по формулам:

$$\mu_y = w_1\mu_1 + w_2\mu_2 \quad \sigma_y^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2\rho w_1w_2\sigma_1\sigma_2$$

Расчетные значения этих величин при различном соотношении весов активов представлены в таблице.

Вес 1-го актива	Вес 2-го актива	С.к.о. портфеля	Доход портфеля
w_1	w_2	σ_y	μ_y
0.0	1.0	16.00	13.00
0.1	0.9	15.21	12.70
0.2	0.8	14.54	12.40
0.3	0.7	14.01	12.10
0.4	0.6	13.64	11.80
0.5	0.5	13.44	11.50
0.6	0.4	13.41	11.20
0.7	0.3	13.56	10.90
0.8	0.2	13.89	10.60
0.9	0.1	14.37	10.30
1.0	0.0	15.00	10.00

Зависимость ожидаемого дохода портфеля от с.к.о. портфеля (риска) приведена на рисунке



На линии ABC лежат все возможные комбинации дохода и риска портфеля. Точка А соответствует портфелю, состоящему только из 1-го актива, точка С соответствует портфелю, состоящему только из 2-го актива, точка В соответствует портфелю с наименьшим риском.

Аналитически координаты точки В в случае портфеля из 2-х активов можно найти из следующих соображений. Так как веса активов связаны соотношением $w_2 = 1 - w_1$, то математическое ожидание и дисперсию портфеля можно представить как функцию только от w_1 :

$$\mu_y = w_1\mu_1 + (1 - w_1)\mu_2$$

$$\sigma_y^2 = w_1^2\sigma_1^2 + (1 - w_1)^2\sigma_2^2 + 2\rho w_1(1 - w_1)\sigma_1\sigma_2$$

Точку с минимальным значением дисперсии (с.к.о.) портфеля можно найти, взяв производную величины σ_y^2 по w_1 и приравняв ее к нулю. Решив уравнение относительно w_1 получим:

$$w_1 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

Подставив это выражение в формулы для μ_y и σ_y , можно найти ожидаемый доход и риск портфеля (с.к.о.) в точке В. Для рассмотренного здесь примера получим:

$$w_1 = 0.56 \quad w_2 = 0.44 \quad \mu_y = 11.32\% \quad \sigma_y = 13.39\%$$

Линия ABC:

- вогнута влево при коэффициенте корреляции $\rho < 1$, при этом вогнутость тем сильнее, чем меньше коэффициент корреляции,
- является отрезком прямой, соединяющим точки А и С, при коэффициенте корреляции $\rho = 1$.

Верхняя часть линии ABC (линия BC) является *границей эффективности*. Линия BC является эффективной в том смысле, что на ней невозможно повысить доход без повышения риска и снизить риск без снижения дохода (в отличие от линии AB). Выше линии BC находятся недостижимо привлекательные комбинации дохода и риска. Ниже линии BC находятся худшие комбинации дохода и риска, которые могут быть улучшены

переходом на линию ВС. Например, переход из точки Т2 в точку Т1 приводит к увеличению дохода портфеля при неизменном уровне риска.

На границе эффективности не существует наилучшего портфеля. Выбор точки на этой линии зависит от инвестиционных предпочтений портфельного менеджера, то есть какую плату, выражаемую в единицах риска, он готов нести за добавочную единицу дохода.

Распространенным методом выбора точки на границе эффективности является выбор такого портфеля, для которого отношение ожидаемого дохода к ожидаемому риску является максимальным. Аналитически в случае портфеля из 2-х активов эту точку можно найти, приравняв к нулю производную по w_1 от функции

$$\frac{\mu_y}{\sigma_y} = \frac{w_1\mu_1 + (1-w_1)\mu_2}{\sqrt{w_1^2\sigma_1^2 + (1-w_1)^2\sigma_2^2 + 2\rho w_1(1-w_1)\sigma_1\sigma_2}}$$

Решив полученное уравнение относительно w_1 найдем, что:

$$w_1 = \frac{\mu_2\rho\sigma_1\sigma_2 - \mu_1\sigma_2^2}{(\mu_1\rho\sigma_1\sigma_2 - \mu_2\sigma_1^2) + (\mu_2\rho\sigma_1\sigma_2 - \mu_1\sigma_2^2)}$$

Для рассмотренного здесь примера точка с максимальным отношением дохода к риску:

$$w_1 = 0.37 \quad w_2 = 0.63 \quad \mu_y = 11.89\% \quad \sigma_y = 13.72\%$$

15.5. Постановка задачи по оптимизации портфеля.

Под оптимизацией портфеля, состоящего из произвольного количества активов, мы будем понимать поиск таких наборов весов активов, которые обеспечивали бы:

- ожидаемый доход портфеля больший или равный наперед заданному минимальному значению дохода,
- ожидаемый риск портфеля меньший или равный наперед заданному максимальному значению риска.

Следовательно, если предполагаются известными ожидаемые доходы по каждому из активов, ожидаемые дисперсии (с.к.о.) дохода по каждому из активов, ковариации (коэффициенты кор

реляции) между каждой парой различных активов, то задача оптимизации портфеля сводится к тому, чтобы найти такие наборы весов активов, которые бы удовлетворяли системе

$$\mu_y = \sum_{k=1}^N w_k \mu_k \geq \mu_{\min}$$

$$\sigma_y^2 = \sum_{k=1}^N w_k^2 \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N w_i w_k \rho_{ik} \sigma_i \sigma_k \leq \sigma_{\max}^2$$

$$\sum_{k=1}^N w_k = 1$$

Будем предполагать, что в составе портфеля в качестве одного из активов могут находиться денежные средства, то есть безрисковый актив, имеющий нулевое ожидание дохода, нулевую дисперсию дохода и нулевой коэффициент корреляции с любым другим активом, поэтому равенство единице суммы весов всех активов является строгим.

Ожидаемый доход портфеля и ожидаемая дисперсия портфеля являются *целевыми функциями*. Целевые функции определяют задачу которая должна быть решена в процессе оптимизации. В данном случае задачей является максимизировать линейную функцию μ_y при одновременной минимизации квадратичной функции σ_y^2 с учетом заданных ограничений.

15.6. Введение ограничений на состав и веса активов в портфеле (лимитов).

В постановке задачи по оптимизации портфеля активов, сделанной в предыдущем параграфе, неявным образом предполагалось, что:

- портфельный менеджер имеет возможность инвестировать в любые активы, обращающиеся на рынке, то есть отсутствуют ограничения на состав активов в портфеле,
- отсутствуют ограничения на вес отдельного актива в портфеле.

Как правило присутствуют оба вида ограничений (лимиты). Дополним задачу оптимизации введением лимитов, то есть будем предполагать, что:

- в формулах для целевых функций присутствуют только разрешенные активы,
- вводятся ограничения на вес каждого конкретного актива в портфеле.

Ограничения на вес любого актива будем вводить как сверху, так и снизу. На практике низкое ограничение сверху вводят на потенциально очень доходные, но и очень рискованные активы. Самые низкие ограничения сверху вводят на веса активов, инвестиции в которые с высокой вероятностью могут быть потеряны полностью. В лучшем случае эти активы могут принести доход за пределами горизонта инвестирования. Ограничение сверху не может превышать 1. Ненулевые ограничения снизу вводятся как правило для создания в составе портфеля "подушки безопасности" из низкодоходных, но и низкорискованных активов. Ограничение снизу не может быть меньше 0. Кроме того, и набор ограничений сверху, и набор ограничений снизу имеют свои правила нормирования.

Итак, ограничения на веса активов в портфеле вводятся следующим образом:

$$w_{(\min)k} \leq w_k \leq w_{(\max)k}$$

$$0 \leq w_{(\min)k} < 1 \quad 0 < w_{(\max)k} \leq 1$$

$$\sum_{k=1}^N w_{(\min)k} \leq 1 \quad \sum_{k=1}^N w_{(\max)k} \geq 1$$

15.7. Численное решение задачи оптимизации портфеля с учетом лимитов методом Монте-Карло.

Итак, задачу об оптимизации портфеля активов с учетом лимитов можно сформулировать в виде:

$$\mu_y = \sum_{k=1}^N w_k \mu_k \geq \mu_{\min}$$

$$\sigma_y^2 = \sum_{k=1}^N w_k^2 \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N w_i w_k \rho_{ik} \sigma_i \sigma_k \leq \sigma_{\max}^2$$

$$S = \sum_{k=1}^N w_k = 1$$

$$w_{(\min)k} \leq w_k \leq w_{(\max)k}$$

$$0 \leq w_{(\min)k} < 1 \qquad 0 < w_{(\max)k} \leq 1$$

$$S_{(\min)} = \sum_{k=1}^N w_{(\min)k} \leq 1 \qquad S_{(\max)} = \sum_{k=1}^N w_{(\max)k} \geq 1$$

В этой задаче искомыми величинами являются наборы весов активов $\{w_k\}$, удовлетворяющие всем уравнениям и неравенствам. Решить такую задачу оптимизации аналитическими методами достаточно трудно, особенно в случае большого количества активов. Поэтому есть смысл попробовать найти ее решение численно методом Монте-Карло, генерируя случайным образом на компьютере удовлетворяющие ограничениям наборы весов активов и проверяя эти наборы на соответствие целевым функциям.

Разумеется, в конечной последовательности розыгрышей (генераций наборов весов) скорее всего не удастся найти все решения задачи оптимизации. Однако, каждое найденное решение будет удовлетворять всем условиям задачи, то есть портфель, построенный с помощью этого набора весов будет "точно оптимальным". Если решений будет несколько, из них можно выбрать то, при котором отношение ожидаемого дохода портфеля к ожидаемому риску будет максимальным.

Алгоритм решения задачи оптимизации

1) Задаем входные данные

1.1) Набор ожидаемых доходов по активам:
 $\{\mu_k\}, k = 1, \dots, N$

1.2) Набор среднеквадратичных отклонений активов:
 $\{\sigma_k\}, k = 1, \dots, N$

1.3) Набор коэффициентов корреляций между различными активами

$$\{\rho_{ik}\}$$

$$k = 1, \dots, N$$

$$i = k + 1, \dots, N$$

1.4) Минимально ожидаемый доход портфеля: μ_{\min}

- 1.5) Максимально ожидаемый риск портфеля: σ_{\max}
- 1.6) Набор ограничений снизу на веса активов:
 $\{w_{(\min)k}\}, k = 1, \dots, N$
- 1.7) Набор ограничений сверху на веса активов:
 $\{w_{(\max)k}\}, k = 1, \dots, N$
- 1.8) Количество розыгрышей (генераций наборов весов):
 M
- 1.9) Точность нормирования (малое положительное число): δ
- 2) Задаем стартовое значение номера текущего розыгрыша
 $m = 0$
- 3) Номер текущего розыгрыша $m = m + 1$
- 4) Разыгрываем случайным образом набор весов, соответствующий набору ограничений на веса активов
 - 4.1) Задаем начальные минимально возможные значения весов
 $w_k = w_{(\min)k}, k = 1, \dots, N$
 - 4.2) Вычисляем текущую сумму весов

$$S = \sum_{k=1}^N w_k$$
 - 4.3) $k = 0$
 - 4.4) $k = k + 1$
 - 4.5) Разыгрываем приращение веса k -го актива
 $dW = \text{СЛУЧАЙНОЕ_ЧИСЛО}(0, \text{МИНИМУМ}(1 - S, w_{(\max)k} - w_k))$
 - 4.6) Вычисляем текущий вес k -го актива
 $w_k = w_k + dW$
 - 4.7) Вычисляем текущую сумму весов
 $S = S + dW$
 - 4.8) Если $k < N$, то переходим на шаг 4.4
 - 4.9) Если не соблюдается точность нормирования, то есть $(1 - S) > \delta$, то переходим на шаг 4.3

- 5) Проверяем, соответствует ли портфель, составленный из полученных на шаге 4 весов, ограничениям на минимально ожидаемый доход и максимально ожидаемый риск портфеля

$$\mu_y = \sum_{k=1}^N w_k \mu_k \geq \mu_{\min}$$

$$\sigma_y^2 = \sum_{k=1}^N w_k^2 \sigma_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N w_i w_k \rho_{ik} \sigma_i \sigma_k \leq \sigma_{\max}^2$$

Если решение соответствует этим неравенствам, то фиксируем его, то есть запоминаем текущий набор весов и соответствующие ему ожидаемый доход, ожидаемый риск, отношение дохода к риску. Текущий розыгрыш за номером m завершен.

- 6) Если номер текущего розыгрыша m меньше, чем общее число розыгрышей M , то переходим на шаг 3
- 7) После того, как сделаны все розыгрыши (то есть $m=M$), среди всех найденных решений выбираем то, при котором отношение ожидаемого дохода портфеля к ожидаемому риску портфеля будет максимальным.

Если не было найдено ни одного решения, то необходимо:

- либо увеличить количество розыгрышей,
- либо ослабить ограничения по целевым функциям, то есть уменьшить ожидаемый доход и/или увеличить ожидаемый риск портфеля,
- либо пересмотреть ограничения на веса активов в портфеле в сторону расширения интервалов разрешенных значений весов.

Пример решения задачи оптимизации

Рассмотрим портфель, состоящий из трех активов:

- *1-й актив*
ожидаемый доход 10%
с.к.о. ожидаемого дохода (риск) 15%
отношение доход/риск = 0.67
минимальный вес в портфеле 0
максимальный вес в портфеле 0.50
- *2-й актив*
ожидаемый доход 13%
с.к.о. ожидаемого дохода (риск) 20%
отношение доход/риск = 0.65

минимальный вес в портфеле 0
максимальный вес в портфеле 0.50

- 3-й актив
ожидаемый доход 18%
с.к.о. ожидаемого дохода (риск) 30%
отношение доход/риск = 0.60
минимальный вес в портфеле 0
максимальный вес в портфеле 0.50

Коэффициенты корреляции между активами:

- 1-й и 2-й активы: $\rho_{12} = 0.5$
- 1-й и 3-й активы: $\rho_{13} = -0.5$
- 2-й и 3-й активы: $\rho_{23} = 0$

Зададим минимально ожидаемый доход портфеля 15%, максимальный ожидаемый риск портфеля 18%, количество розыгрышей весов активов 10000.

Тогда решение задачи оптимизации по приведенному в этом параграфе алгоритму с выбором среди полученных решений того, которое обеспечивает максимальное отношение ожидаемого дохода к ожидаемому риску следующее:

- вес 1-го актива $w_1 = 0.1141$
- вес 2-го актива $w_2 = 0.3859$
- вес 3-го актива $w_3 = 0.5000$
- ожидаемый доход портфеля $\mu_y = 15.16\%$
- ожидаемый риск портфеля $\sigma_y = 16.58\%$
- отношение доход/риск = 0.91

Мы получили доход портфеля больше среднего арифметического доходов активов (13.67%), риск портфеля меньше среднего арифметического рисков активов (21.67%), и существенно улучшили соотношение дохода и риска по сравнению с этим соотношением для любого отдельного актива. Следует еще раз подчеркнуть, что это решение наверняка не является оптимальным. Более длинные серии розыгрышей вероятно способны его улучшить. Однако, данное решение является "достаточно оптимальным", так как удовлетворяет всем поставленным условиям.

16. УПРАВЛЕНИЕ РИСКОМ ПОРТФЕЛЯ НА ОСНОВЕ АНАЛИЗА КВАНТИЛЬНЫХ МЕР РИСКА

16.1. Введение.

В предыдущей главе были рассмотрены вопросы, связанные с управлением риском портфеля, при этом в качестве меры рассеяния ожидаемого дохода по конкретному активу и по портфелю, то есть в качестве меры риска, была использована дисперсия (или среднеквадратичное отклонение).

Широкое использование дисперсии в качестве оценки рассеяния ожидаемого дохода портфеля связано с тем, что ее можно вычислить аналитически, если известны дисперсии каждого актива и коэффициенты корреляции между активами. Действительно, дисперсия ожидаемого дохода портфеля - это взвешенная сумма ковариаций всех пар активов в портфеле, причем вес каждой ковариации равен произведению весов соответствующей пары активов, а ковариация актива с самим собой является дисперсией данного актива. При этом суммирование проводится безотносительно к разнообразию законов распределений каждого из слагаемых и возможной деформации законов распределения при суммировании.

Ситуация существенным образом усложняется, если в качестве меры рассеяния ожидаемого дохода портфеля необходимо указать квантильное отклонение с заданной доверительной вероятностью, так как это невозможно сделать, если неизвестен закон распределения доходов портфеля. Теоретически, закон распределения доходов портфеля можно найти, если известны законы распределения доходов входящих в него активов. Однако на практике аналитическое решение такой задачи сопряжено со значительными трудностями даже для малого числа активов (за исключением некоторых частных случаев). Это происходит потому, что, во-первых, доходы по входящим в портфель активам могут быть коррелированы между собой, во-вторых, при суммировании доходов активов законы их распределения могут существенным образом деформироваться, поэтому распределение дохода портфеля может сильно отличаться от распределений доходов составляющих его активов.

В этой главе будут рассмотрены практические методы вычисления квантильных мер риска дохода портфеля из произвольного количества активов и управления риском портфеля на основе их анализа.

16.2. Понятие Value-at-risk и Shortfall-at-risk.

Допустим, что предполагается сформировать портфель, состоящий из некоторого набора активов. Введем обозначения:

N - количество активов в портфеле,

T - промежуток времени, в течение которого предполагается поддерживать портфель в неизменном состоянии,

$\{w_n\}$ - набор весов, с которыми активы входят в портфель на момент его формирования,

$\{x_n\}$ - набор случайных величин, равных доходам по каждому из активов, которые будут реально получены по истечении промежутка времени T ,

y - случайная величина, равная реально полученному доходу портфеля по истечении промежутка времени T ,

$\{\mu_n\}$ - набор ожидаемых значений доходов по каждому из активов,

μ_y - ожидаемое значение дохода портфеля.

Реально полученный доход портфеля и ожидаемое значение дохода портфеля за время T можно вычислить по формулам:

$$y = \sum_{n=1}^N w_n x_n \quad \mu_y = \sum_{n=1}^N w_n \mu_n$$

В большинстве случаев оказывается, что величины y и μ_y не равны друг другу. В частности, по истечении промежутка времени T результатом инвестирования может оказаться убыток, то есть $y < 0$.

В качестве квантильной меры риска портфеля используется величина *Value-at-risk* (*VAR*), то есть *VAR* - это мера риска портфеля с заданной доверительной вероятностью.

Утверждение о том, что портфель имеет определенное значение *VAR* фактически означает следующее: в течение проме

жутка времени T с доверительной вероятностью P абсолютная величина убытка по портфелю не может быть больше, чем VAR (доход по портфелю не может быть меньше $-VAR$), при этом абсолютная величина убытка, превосходящая VAR , также не исключена, однако такой убыток может случиться лишь с малой вероятностью $1 - P$.

Пусть случайная величина y (доход портфеля) имеет плотность распределения $p(y)$ и функцию распределения $F(y)$. Зададим доверительную вероятность P . Обозначим как y_{1-P} такую квантиль распределения переменной y , что

- вероятность того, что случайная величина y окажется меньше или равной y_{1-P} , равна $1 - P$, т.е.

$$Prob\{y \leq y_{1-P}\} = 1 - P$$

- вероятность того, что случайная величина y окажется больше y_{1-P} , равна P , т.е.

$$Prob\{y > y_{1-P}\} = P$$

Доверительную вероятность P как правило выбирают в пределах от 0.95 до 0.99. В этом случае квантиль y_{1-P} является отрицательным числом. Тогда величина VAR равна модулю этого числа:

$$VAR = |y_{1-P}|$$

Для того, чтобы найти VAR случайной величины y , нужно решить уравнение

$$1 - P = F(-VAR)$$

или (эквивалентная форма записи)

$$1 - P = \int_{-\infty}^{-VAR} p(y) dy$$

VAR является достаточно распространенной мерой риска портфеля, которая имеет однако ряд существенных недостатков:

- 1) В отличие от дисперсии, VAR не обладает свойством аддитивности. Это означает, что даже если известны величины VAR составляющих портфель активов и коэффициенты корреляции между активами, то в общем случае на основании

этих данных невозможно рассчитать VAR портфеля. Это связано с тем, что VAR является квантильной оценкой, поэтому ее нельзя вычислить без знания закона распределения дохода портфеля. Но при суммировании доходов активов законы их распределения могут деформироваться, поэтому распределение дохода портфеля может сильно отличаться от распределений доходов составляющих его активов.

- 2) VAR не учитывает возможных больших убытков, которые могут произойти с малыми вероятностями.
- 3) При указании в качестве меры риска только величины VAR , мы не имеем информации о виде распределения доходов портфеля. При этом, если распределение доходов портфеля является островершинным (эксцесс больше 3), то риск портфеля будет недооцениваться, а если распределение доходов является плосковершинным (эксцесс меньше 3), то риск будет переоцениваться.

Исходя из перечисленных выше недостатков VAR , хотелось бы иметь характеристику риска портфеля, которая описывает реализующиеся с малыми вероятностями аномально большие убытки. Такой мерой риска является *Shortfall-at-risk* (SAR). SAR - это ожидаемое значение убытка портфеля, при условии, что абсолютная величина убытка превосходит VAR . Исходя из данного определения, значение SAR может быть вычислено по формуле:

$$SAR = \int_{-\infty}^{-VAR} y \cdot p(y) dy$$

Совместное использование в качестве мер риска VAR и SAR позволяет иметь более полную информацию о хвостах распределения доходов портфеля. При этом представляется целесообразным рассчитывать эти величины одновременно для нескольких различных значений доверительной вероятности P (например 0.950, 0.975, 0.990, 0.999).

16.3. Вычисление Value-at-risk и Shortfall-at-risk.

Прежде всего рассмотрим, каким образом можно найти значения VAR и SAR для отдельного актива. Для расчета этих величин необходима информация о плотности распределения дохо

дов, которая может быть найдена по эмпирической выборке рыночных цен данного актива. Введем обозначения:

$0, \dots, t_{\max}$ - интервал времени, на котором производится выборка цен,

$Price_t$ - цена актива в момент времени t ($0 \leq t \leq t_{\max}$).

Тогда величиной, характеризующей однодневный доход от вложения в данный актив от момента $t-1$ до момента t будет

$x_t = \ln(Price_t / Price_{t-1})$, где $1 \leq t \leq t_{\max}$.

Значения VAR и SAR могут быть определены по выборке случайных величин $\{x_t\}$ следующими способами:

1) *Непосредственно по выборке $\{x_t\}$.*

При использовании этого способа выборка $\{x_t\}$ должна быть упорядочена по возрастанию. После этого нужно найти номер M в упорядоченной по возрастанию выборке, соответствующий доверительной вероятности P :

$$M = 1 + \text{ЦЕЛОЕ}((1 - P) \cdot (t_{\max} - 1))$$

Тогда

$$VAR = |x_M| \quad SAR = \frac{1}{t_{\max}} \sum_{t=1}^M x_t$$

Данный метод является самым простым и самым грубым. Он исключительно чувствителен к конкретной случайной реализации цен на интервале $0, \dots, t_{\max}$. Кроме того, этот способ сильно зависит от объема выборки. Действительно, при малой величине t_{\max} мы можем для различных значений P получать один и тот же номер M , и как следствие этого, одни и те же величины VAR и SAR .

2) *Путем идентификации закона распределения случайной величины x .*

Для этого по выборке $\{x_t\}$, используя изложенную в главе 6 методику, необходимо построить эмпирическую гистограмму распределения случайной величины x , после чего аппроксимировать эту гистограмму одним из аналитических

законов распределения, которые описаны в главе 2. Тем самым будет найдена формула для плотности распределения $p(x)$. Для заданной доверительной вероятности P величины VAR и SAR могут быть найдены из уравнений

$$1 - P = \int_{-\infty}^{-VAR} p(x) dx \qquad SAR = \int_{-\infty}^{-VAR} x \cdot p(x) dx$$

Решение этих уравнений проводится как правило численно, методика приведена в главе 2.

3) *Путем многократного моделирования методом Монте-Карло*

Как и в предыдущем случае, по эмпирической выборке $\{x_i\}$ необходимо найти аналитическую формулу для плотности распределения $p(x)$. Далее, используя генератор псевдослучайных чисел, нужно провести многократное моделирование случайной величины x (методика подробно изложена в главе 2). После этого величины VAR и SAR определяются уже по смоделированной выборке.

ПРИМЕЧАНИЕ. Таким образом мы нашли однодневные VAR и SAR . Если необходимо вычислить например пятидневные значения этих величин, то вместо дневных баров цен используются недельные бары и т.д.

После того как известно, каким образом можно найти значения VAR и SAR для отдельного актива, можно описать процедуру поиска этих величин для портфеля.

Проще всего VAR и SAR для портфеля можно найти, если на интервале времени $0, \dots, t_{\max}$ рассмотреть поведение виртуального портфеля, сформированного в момент $t = 0$ из интересующих нас активов, каждый из которых входит в портфель с заданным весом. Если считать, что $Price_t$ - это стоимость виртуального портфеля в момент времени t , то вся процедура поиска VAR и SAR для него уже описана выше. Такой подход обладает следующими достоинствами:

- отсутствует необходимость учета корреляции между входящими в портфель активами (это происходит автоматически),

- отсутствует необходимость учета законов распределения входящих в портфель активов, так как нас интересует только закон распределения результирующего виртуального портфеля, а его можно определить по выборке $\{Price_t\}$.

Существенным недостатком такого подхода является то, что текущие веса активов в портфеле равны заданному набору весов $\{w_n\}$ только в начальный момент времени $t = 0$. Это происходит потому, что динамика цен входящих в виртуальный портфель активов различна для разных активов, результатом чего является постоянное изменение соотношения текущих весов активов в виртуальном портфеле в различные моменты времени (каждый актив учитывается по его текущей рыночной цене). Действительно, если считать, что в момент $t = 0$ стоимость виртуального портфеля $Price_0 = 1$, то в последующие моменты времени его стоимость выражается формулой:

$$Price_t = \sum_{n=1}^N w_n \frac{Price_{n,t}}{Price_{n,0}} \quad t = 1, \dots, t_{\max}$$

где $Price_{n,t}$ - это цена n -го актива в момент времени t .

Из этой ситуации может быть два выхода. Во-первых, можно пренебречь изменением соотношения весов. Во-вторых, после каждого торгового дня приводить текущие веса активов в соответствие с заданным набором $\{w_n\}$. В этом случае стоимость виртуального портфеля в произвольный момент времени будет равна:

$$Price_0 = 1$$

$$Price_t = \prod_{i=1}^t \left(\sum_{n=1}^N w_n \frac{Price_{n,i}}{Price_{n,i-1}} \right) \quad t = 1, \dots, t_{\max}$$

Более удобно использовать рекуррентную формулу

$$Price_t = Price_{t-1} \cdot \sum_{n=1}^N w_n \frac{Price_{n,t}}{Price_{n,t-1}}$$

В некоторых частных случаях VAR портфеля можно найти аналитически, если известен набор $\{VAR_n\}$ входящих в портфель

активов и коэффициенты корреляции между активами. Это следующие случаи:

- 1) Если доходы всех активов распределены нормально, то их сумма (доход портфеля), также распределен нормально, то есть не происходит деформации закона распределения при суммировании.
- 2) Если при не обязательно нормальном распределении доходов активов выбран уровень доверительной вероятности $P = 0.95$. Дело в том, что для большого числа наиболее употребительных законов распределения 5%-ная квантиль распределения выражается через математическое ожидание и с.к.о. по единой формуле $x_{0.05} \approx \mu - 1.6\sigma$.

В обоих случаях VAR портфеля вычисляется по формуле, аналогичной формуле для дисперсии портфеля:

$$VAR_y^2 = \sum_{k=1}^N w_k^2 \cdot VAR_k^2 + 2 \sum_{k=1}^N \sum_{i=k+1}^N w_i w_k \rho_{ik} \cdot VAR_i \cdot VAR_k$$

16.4. Оптимизация портфеля с учетом Value-at-risk и Short-fall-at-risk.

Алгоритм численного решения задачи оптимизации портфеля по соотношению математического ожидания дохода и среднеквадратичного отклонения дохода приведен в главе 15 параграфе 15.7 этой книги. Введение в рассмотрение мер риска портфеля VAR и SAR лишь дополняет этот алгоритм. Выбор конкретного решения из множества решений задачи оптимизации портфеля (то есть оптимальное соотношение величин μ_y , σ_y , VAR_y и SAR_y), каждый портфельный менеджер осуществляет с учетом своих инвестиционных предпочтений.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Гнеденко Б.В., Курс теории вероятностей, "Наука", 1971.
2. Крамер Г., Математические методы статистики, "Мир", 1975.
3. Новицкий П.В., Зограф И.А., Оценка погрешностей результатов измерений, "Энергоатомиздат", 1985.
4. Корн Г., Корн Т., Справочник по математике для научных работников и инженеров, "Наука", 1984.
5. Прудников А.П., Брычков Ю.А., Маричев О.И., Интегралы и ряды, "Наука", 1981.