Глава 1. ОСНОВНЫЕ МЕТОДЫ СТАТИСТИКИ

1. СРЕДНИЕ ВЕЛИЧИНЫ И ВАРИАЦИЯ

Средняя арифметическая величина

Имеется n объектов (единиц), которые обладают некоторым измеримым качеством, или признаком, x_i – значение признака i–го объекта.

Средняя арифметическая величина – сумма признаков всех объектов, деленная на количество объектов.

Простой средней называют число

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Пример. Вес трех студентов равен 55, 65 и 90 кг, тогда их средний вес равен (55 + 65 + 90)/3 = 70 кг.

Если совокупность объектов разделена на группы с одинаковым значением признака, тогда используют другую формулу средней величины.

Взвешенной средней называют число

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i f_i}{\sum_{i=1}^{m} f_i},$$

где f_i – количество объектов со значением признака x_i, m - количество групп объектов ($m<n$). Знаменатель дроби равен общему числу объектов.

Взвешенная средняя может быть записана в виде:

$$\bar{x} = \sum a_i x_i,$$

где a_i – удельный вес объектов i–й группы в общей численности объектов:

$$a_i = \frac{f_i}{\sum f_i}, \quad \sum a_i = 1.$$

Пример. Оценку «5» получили 20 студентов, оценку «4» – 80. Тогда a_1 = 20/(20 + 80) = 0,2, a_2 = 0,8. Средняя оценка равна 0,2×5 + 0,8×4 = 4,2.

Свойства арифметической средней.

1. Сумма отклонений индивидуальных значений признака от его среднего значения равна нулю: $\sum(x_i - \bar{x}) = 0.$

Доказательство: $\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0.$

2. Если каждое индивидуальное значение признака умножить или разделить на постоянное число, то средняя изменится во столько же раз.

3. Если к каждому индивидуальному значению признака прибавить или вычесть постоянное число, то средняя изменится на то же число.

4. Если веса средней взвешенной умножить или разделить на постоянное число, то средняя не изменится.

5. Сумма квадратов отклонений индивидуальных значений признака от средней меньше, чем для любого другого числа.

Доказательство. Найдем минимум функции $f(a)=\sum(x_i-a)^2$, для этого приравняем ее производную нулю, получим

$$\sum 2(-1)(x_i-a)=0, \quad \textit{или} \quad \sum x_i-\sum a=0, \quad \textit{отсюда} \quad a=\sum x_i/n=\bar{x}.$$

Другие формы средних величин

1. *Средняя квадратическая величина*:

$$\bar{x}_{кв}=\sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2}}{n}}.$$

Пример. Квадраты со сторонами 5 м и 3 м заменяют на два одинаковых квадрата с той же суммарной площадью. Искомую сторону квадрата обозначим через x, тогда $x^2 + x^2 = 5^2 + 3^2$, отсюда $x=\sqrt{(5^2+3^2)/2}=4$ м, т.е. искомая сторона равна средней квадратической величине сторон квадратов.

Замечание. При замене двух кубов с разными сторонами на два одинаковых куба с тем же суммарным объемом используют *среднюю кубическую.* Ее формула получается из заданной заменой показателя 2 на 3.

2. *Средняя геометрическая величина:*

$$\bar{x}_{геом}=\sqrt[n]{x_1\times x_2\times\ldots\times x_n}.$$

Пример. Цена выросла за первый год в 1,2 раза, за второй – в 1,5 раза. Определим, какой постоянный годовой темп роста цены (x) обеспечит такое же ее увеличение за два года. Имеем: $x\times x = 1{,}2\times 1{,}5$, отсюда $x=\sqrt{1{,}2\times 1{,}5}=1{,}34$ раза, т.е. искомая величина равна средней геометрической величине годовых темпов роста цены.

3. *Средняя гармоническая величина*:

$$\bar{x}_{гарм}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}.$$

Пример. Один рабочий выполняет задание за 2 ч, другой – за 3 ч. Их заменяют двумя рабочими равной производительности, которые вместе выполнят задание за то же время, что и данные рабочие. Обозначим время выполнения задания одним новым рабочим через x, тогда суммарная производительность двух новых (и старых) рабочих равна

$1/x + 1/x = 1/2 + 1/3$, отсюда $x = 2/(1/2 + 1/3) = 2{,}8$ ч.

Итак, время выполнения задания одним новым рабочим равно гармонической средней времени выполнения работы старыми рабочими.

4. *Степенная средняя величина* служит обобщением рассмотренных ранее средних величин, она равна

$$\bar{x}=\sqrt[k]{\frac{\sum_{i=1}^{n} x_i^{\,k}}{n}}.$$

При k=-1 получаем *гармоническую* среднюю, при k=1 – *арифметическую* среднюю, при k=2 – *квадратическую* среднюю, при k=3 – *кубическую* среднюю. Выполняется неравенство:

$$\bar{x}_{гарм} \le \bar{x}_{геом} \le \bar{x}_{арифм} \le \bar{x}_{квадр} \le \bar{x}_{куб},$$

причем равенство достигается в случае равенства всех значений признаков.

Вариация массовых явлений

Вариация признака – это различие его значений у разных объектов (единиц совокупности) в один и тот же момент времени.

Вариационный ряд – объекты расположены по возрастанию (убыванию) значений признака с указанием числа объектов с равными значениями признака. Существуют три формы вариационного ряда.

Ранжированный ряд – объекты расположены по возрастанию (убыванию) значений признака.

Пример: рост Ивана – 165 см, Петра – 170 см., Антона – 175 см.

Дискретный ряд – соответствие между значением признака (x_i) и числом объектов с этим значением (f_i). Число f_i называют *частотой* появления i –го значения признака.

Пример: 3 атлета имеют вес 65 кг, 5 – вес 70 кг, 2 – вес 75 кг.

Полигон – изображение дискретного ряда в виде ломаной, соединяющей точки (x_1, f_1), …, (x_k, f_k), где k – число значений признака.

Пример (см. таблицу): вершины полигона: (65,3), (70,8), (75,2).

Накопленная частота – число объектов со значением признака не меньше, чем x_i, она равна:

$$f_i' = \sum_{j=1}^{i} f_j .$$

Кумулятивный вариационный ряд – соответствие между значением признака (x_i) и накопленной частотой (f_i').

Кумулята – это полигон кумулятивного вариационного ряда, его график – восходящая ломаная.

Пример (см. таблицу): вершины кумуляты: (65,3), (70,8), (75,10).

Огива – это нисходящая ломаная с вершинами (x_i, f_i''), где $f_i'' = n - f_i'$ – «дополняющая» кумулятивная частота.

Значение признака, x_i	Число объектов, f_i	Накопленная частота, f_i'
65	3	3
70	5	8
75	2	10

Интервальный ряд – соответствие между интервалом изменения признака и числом объектов, попадающих в данный интервал.

Оптимальное *число интервалов* (групп) равно $k = 1 + 1{,}44\ln n$, где k – число групп, n – число объектов. Пример: при 20 объектах оптимальное число интервалов равно $1 + 1{,}44\ln 20 = 5{,}3 \approx 5$.

Длина интервала равна $(x_{max} - x_{min})/k$, где x_{max} и x_{min} – максимальное и минимальное значения признака, k – число интервалов. Формула используется для ряда с равными интервалами. Пример: максимальный рост – 175 см, минимальный – 165 см, число групп – 5, тогда длина интервала равна (175-165)/5 = 2 см.

Гистограмма – изображение интервального ряда в виде прямоугольников, основания которых равны длинам интервалов, а высоты – частотам.

Среднее значение признака для ряда с равными интервалами равно

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i' f_i,$$

где x_i' – середина i –го интервала, f_i – частота i–го значения признака, n - общее число объектов.

Пример. 2 атлета имеют вес 60-64 кг, 8 атлетов – вес 65-69 кг. Тогда середина 1-го интервала равна (60+64)/2=62 кг, 2-го – (65+69)/2=67 кг, а средний вес атлетов равен (62×2+67×8)/(2+8)= 66 кг.

Медиана и мода

Медиана – значение признака, делящее ряд на две равные части – со значением признака меньше медианы и со значением признака больше ее.

Для *ранжированного ряда* с нечетным числом объектов медиана равна значению признака среднего объекта.

Пример. Возраст рабочих 18, 20, 28, 57 и 64 лет. Тогда медианный рабочий – 3-й, а медиана равна 28 лет. Здесь медиана существенно отличается от среднего возраста, который составляет 37,4 лет.

Для *интервального ряда* используют алгоритм расчета медианы:

1. Найти *медианный объект* с номером $(n+1)/2$, где n – число объектов.

2. Найти *медианный интервал* с номером me, в котором находится медианный объект. Накопленные частоты в медианном интервале (f_{me}') и предыдущем (f_{me-1}') удовлетворяют неравенству: $f_{me-1}' < (n+1)/2 < f_{me-1}'$.

Нижняя граница медианного интервала равна x_0, его длина - i.

3. Найти *медиану*:

$$Me = x_0 + \frac{\frac{n}{2} - f_{me-1}'}{f_{me}} i.$$

Пример. Интервальный ряд роста студентов задан в таблице.

Группа, i	Рост, см	Частота, f_i	Накопленная частота, f_i'
1	160-165	2	2
2	165-170	4	6
3	170-175	9	15

4	175-180	6	21

1. Номер медианного студента равен (21+1)/2 = 11.

2. Медианная группа – 3-я, т.к. для нее выполняется: 6 <11<15. Нижняя граница медианного интервала равна 170 см, его длина – 5 см.

3. Медиана равна: $Me = 170 + \frac{21/2 - 6}{9} \times 5 = 172{,}5.$

Аналогично медиане вычисляют значения признака, делящие совокупность на четыре равные части. Эти значения называют *квартилями* и обозначают Q_1, Q_2, Q_3. Очевидно, $Q_2 = Me$. Алгоритмы расчета медианы и квартилей аналогичны, формула расчета 1-го квартиля:

$$Q_1 = x_1 + \frac{\frac{n}{4} - f_{q1-1}'}{f_{q1}} i,$$

где x_1 – нижняя граница 1-го квартильного интервала, f_{q1} – частота в этом интервале $f_{q1\text{-}1}'$ – накопленная частота в предыдущем интервале, i – длина 1-го квартильного интервала.

Пример (см. табл.). Здесь 1-й квартильный интервал – это 2-й интервал, т.к. для него выполняется: 2 < 21/4 < 6. Первый квартиль равен:

$$Q_1 = 165 + \frac{5{,}25 - 2}{4} \times 5 = 169{,}1.$$

Значения признака, делящие ряда на пять равных частей, называют *квинтилями*, на десять частей – *децилями*, на сто частей – *перцентилями*.

Мода – значение признака, которое встречается наиболее часто.

Пример: 3 атлета имеют вес 65 кг, 5 – вес 70 кг, 2 – вес 75 кг, тогда мода равна 70 кг.

Вариационный ряд может иметь несколько модальных значений, при двух значениях он называется *бимодальным*.

Для *интервального ряда* используют алгоритм расчета моды:

1. Найти *модальный интервал* с номером *mo*, в котором частота f_{mo} максимальна. Нижняя граница модального интервала равна x_0, его длина - i.

2. Найти *моду*:

$$Mo = x_0 + \frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})} i,$$

где f_{mo}, $f_{mo\text{-}1}$ и f_{mo+1} – частота в модальном, предыдущем и последующем интервалах соответственно. Из формулы следует:

- при равенстве частот в предыдущем и последующем интервалах мода совпадает с серединой модального интервала:

- мода ближе к нижней границе модального интервала, если частота в предыдущем интервале больше, чем в последующем интервале, и наоборот.

Пример (см. табл.). Модальный интервал – 3-й. Мода равна

$$Mo = 170 + \frac{9 - 4}{(9 - 4) + (9 - 6)} \times 5 = 173{,}12.$$

Размах и интенсивность вариации

Амплитуда (размах) вариации – разность между максимальным и минимальным значениями признака в совокупности объектов:

$$R = x_{max} - x_{min}.$$

Средний модуль отклонений (среднее линейное отклонение) равен

$$a = \frac{1}{n}\sum_{i=1}^{n} |\, x_i - \bar{x}\,|, \quad \textit{или} \quad a = \frac{1}{n}\sum_{j=1}^{k} |\, x_j' - \bar{x}\,|\, f_j,$$

где n – число объектов совокупности, k – число интервалов, x_j' – середина j – го интервала, f_j – число объектов в j –м интервале.

Среднее квадратическое отклонение равно:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \textit{или} \quad s = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(x_j' - \bar{x})^2\, f_j}.$$

Дисперсия (s^2) – это квадрат среднего квадратического отклонения.

Свойство. Дисперсия равна среднему квадрату значений признака минус квадрат среднего значения: $s^2 = \overline{x^2} - \bar{x}^2$.

Доказательство.

$$s^2 = \frac{1}{n}\sum(x_i^{\,2} - 2x_i\bar{x} + \bar{x}^2) = \frac{\sum x_i^{\,2}}{n} - \frac{2\bar{x}\sum x_i}{n} + \frac{\sum \bar{x}^2}{n} = \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2.$$

Среднее квартильное расстояние – характеризует силу вариации в центральной части совокупности, оно равно

$$q = \frac{Q_3 - Q_1}{2}.$$

Относительные показатели вариации:

- относительный размах вариации: $p = R : \bar{x}$,
- относительное отклонение по модулю: $m = a : \bar{x}$,
- коэффициент вариации: $v = s : \bar{x}$,
- относительное квартильное расстояние: $d = q : \bar{x}$.

Пример. Задан вариационный ряд: 2, 7, 3, 4 (см. таблицу).

i	x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$
1	2	-2	2	4
2	7	3	3	9
3	3	-1	1	1
4	4	0	0	0
Сумма	16	0	6	14
Среднее	4	0	1,5	3,5
Показатель	$\bar{x}$	–	a	s^2

Амплитуда вариации: R = 7-2=5, дисперсия: s^2 = 3,5 (см. таблицу), среднее квадратическое отклонение: $s = \sqrt{3{,}5} = 1{,}87$, относительные показатели вариации: p = 5:4=1,25, m = 1,5:4=0,37, v = 1,87:4=0,47.

Таблица. Максимальные значения показателей вариации

R	a	s	p	m	v
$\bar{x}n$	$2\bar{x}(1-\frac{1}{n})$	$\bar{x}\sqrt{n-1}$	n	$2(1-\frac{1}{n})$	$\sqrt{n-1}$

Моменты распределения

Момент распределения t – го порядка равен

$$M_t = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^t.$$

Момент распределения 1-го порядка (M_1) равен нулю, т.к. сумма отклонений значений признака от среднего значения равна нулю.

Момент распределения 2-го порядка (M_2) равен дисперсии.

Момент распределения 3-го порядка (M_3) характеризует степень асимметричности распределения. Он равен нулю при строго симметричном распределении, для которого сумма кубов отрицательных отклонений равна сумме кубов положительных отклонений.

Коэффициент асимметрии равен

$$As = \frac{M_3}{s^3},$$

где s – среднее квадратическое отклонение, он характеризует асимметрию для крайних значений признака.

Показатель Пирсона характеризует асимметрию в средней части распределения, он равен

$$As_{\text{П}} = \frac{\bar{x} - Mo}{s}.$$

Асимметрия *правосторонняя* при $As_{\text{п}} > 0$ и *левосторонняя* при $As_{\text{п}} < 0$.

Пример. Возраст рабочих: 22, 48, 28, 22 лет. Средний возраст – 30, мода – 22, среднее квадратическое отклонение – 10,68, момент распределения 3-го порядка – $(-8^3 + 18^3 - 2^3 - 8^3)/4 = 1200$, коэффициент асимметрии – $1200/10{,}68^3 = 0{,}98$, показатель Пирсона – $(30-22)/10{,}68 = 0{,}75$ – он положителен, т.е. имеется правосторонняя асимметрия. Поскольку коэффициент асимметрии больше показателя Пирсона, степень асимметрии более значительна для крайних значений признака, чем в средней части распределения.

Эксцессом называют показатель

$$Ex = \frac{M_4}{s^4} - 3,$$

где M_4 – это момент распределения 4-го порядка. Показатель используют для сравнения симметричных рядов с нормальным распределением, для которого отношение M_4/s^4 равно 3:

- если эксцесс близок к *нулю*, то распределение близко к нормальному;

- если эксцесс *отрицателен*, то распределение характеризуется наличием слабо варьирующего «ядра», причем оно «плотнее», чем у нормального распределения;

- если эксцесс *положителен*, то «ядро» более слабое по сравнению с нормальным распределением или оно вообще отсутствует.

Пример. 2 рабочих получают зарплату 10 тыс. руб. и т.д. (см. таблицу).

i	x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$f_i (x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^4$
1	10	2	20	-20	800	320000
2	20	3	60	-10	300	30000
3	30	10	300	0	0	0
4	40	3	120	10	300	30000
5	50	2	100	20	800	320000
Сумма	-	20	600	0	2200	700000
Среднее	-	-	30	-	110	35000
Показатель	-	n	$\bar{x}$	-	s^2	M_4

Эксцесс равен $35000/110^2 - 3 = -0{,}11$. Следовательно, данное распределение близко к нормальному, однако его ядро более плотно по сравнению с нормальным распределением, т.е. значения признака в меньшей степени тяготеют к среднему значению.

Задачи

1. Интервальный ряд задан в таблице. Найти (с точн. до 0,01 года): а) средний возраст рабочих; б) медиану; в) первый квартиль; г) моду.

Группа	Возраст, лет	Число рабочих
1	15-25	8
2	25-35	15
3	35-45	24
4	45-55	31
5	55-65	19

2. Вариационный ряд: 8, 6, 5, 1. Найти (с точн. до 0,01): а) размах вариации; б) среднее линейное отклонение; в) дисперсию, г) среднее квадратическое отклонение; д) коэффициент вариации.

3. В бригаде 6 рабочих в возрасте 18-20 лет и 14 рабочих в возрасте 20-22 лет. Найти (с точн. до 0,01): а) средний квадрат возраста; б) средний возраст; в) дисперсию значений возраста; г) коэффициент вариации.

4. Возраст рабочих: 50, 80, 70, 80 лет. Найти (с точн. до 0,01): а) коэффициент асимметрии; б) показатель Пирсона.

5. Зарплату 10 тыс.руб. получают 2 чел., 20 тыс.руб. – 1 чел., 30 тыс.руб. – 4 чел., 40 тыс.руб. – 1 чел., 50 тыс.руб. – 2 чел.. Найти: а) дисперсию, б) момент распределения 4-го порядка; в) эксцесс.

2. ГРУППИРОВКА

Группировка – это распределение объектов совокупности таким образом, что различия между объектами одной группы меньше, чем различия между объектами разных групп.

Группировочный признак – характеристика, используемая при разбиении объектов на группы.

Интервал группировки – значения группировочного признака, относящиеся к одной группе.

Виды группировок: структурная, аналитическая, многомерная.

Структурная группировка

Структурная группировка – характеризует структуру совокупности объектов по одному признаку, она описывается вектором (a_1, …, a_n), где a_i – удельный вес объектов i-й группы, n – число групп, $\sum a_i = 1$.

Неравномерность структуры измеряется дисперсией (s^2) и *индексом Герфиндаля* $H = \sum_{i=1}^{n} a_i^{\,2}$, которые связаны соотношением $s^2 = \frac{H}{n} - \frac{1}{n^2}$.

Пример. Население в возрасте 0-30 лет составляет 20%, 30-60 лет – 50%, старше 60 лет – 30%, тогда структура задается вектором (0,2; 0,5; 0,3). Индекс Герфиндаля равен 0,04+0,25+0,09=0,38, дисперсия 0,38/3-1/9=0,015.

Показатели структурных сдвигов – характеризуют изменение структуры. Пусть начальная структура ($a_1^{\,0}$, …, $a_n^{\,0}$), новая – ($a_1^{\,1}$, …, $a_n^{\,1}$). Обозначим изменение i–й компоненты структуры через $d_i = a_i^{\,1} - a_i^{\,0}$ и определим показатели.

Индекс различий: $K_1 = \frac{\sum |d_i|}{2}$.

Линейный коэффициент структурных сдвигов: $K_2 = \frac{\sum |d_i|}{n}$.

Квадратический коэффициент структурных сдвигов: $K_3 = \sqrt{\frac{\sum d_i^{\,2}}{n}}$.

Коэффициент Гатева: $K_4 = \sqrt{\frac{\sum d_i^{\,2}}{\sum (a_i^{\,0})^2 + \sum (a_i^{\,1})^2}} = \sqrt{\frac{\sum d_i^{\,2}}{H_0 + H_1}}$,

где H_0 и H_1 – индексы Герфиндаля для старой и новой структуры.

Свойства показателей структурных сдвигов:

1. *Минимальное значение* (ноль) достигается при равенстве структур.

2. *Максимальное значение* достигается при равенстве нулю скалярного произведения структур (старое или новое значение каждой доли равно нулю), оно равно:

- индекс различий – 1;
- линейный коэффициент структурных сдвигов – 2/n;
- квадратический коэффициент структурных сдвигов – $\sqrt{(H_0 + H_1)/n}$;
- коэффициент Гатева – 1.

Пример. Старая структура – (0,2; 0,5; 0,3), новая – (0,4; 0,4; 0,2), тогда

$K_1 = (0{,}2+0{,}1+0{,}1)/2=0{,}2$; $K_2 = 0{,}4/3=0{,}13$;

$K_3 = ((0{,}04+0{,}01+0{,}01)/3)^{0{,}5} = 0{,}14$; $K_4 = (0{,}06/(0{,}38+0{,}36))^{0{,}5} = 0{,}28$.

Аналитическая группировка

Аналитическая группировка – характеризует взаимосвязь между результативным признаком и факторными признаками.

Пусть имеется один факторный признак. Обозначим:
- число интервалов изменения факторного признака (групп) – n;
- середина i–го интервала факторного признака – x_i';
- среднее значение результативного признака в i–й группе – $\bar{y}_i$.

Сила связи для i–го интервала равна отношению прироста среднего значения результативного признака и прироста среднего интервального значения факторного признака (аналог производной функции):

$$b_i = \frac{\bar{y}_i - \bar{y}_{i-1}}{x_i' - x_{i-1}'}.$$

Данный показатель не определен для первой группы.

Средняя сила связи равна отношению общего прироста среднего значения результативного признака и общего прироста среднего интервального значения факторного признака:

$$b = \frac{\bar{y}_n - \bar{y}_1}{x_n' - x_1'}.$$

Пример. Возраст ребенка – факторный признак, его рост – результативный признак. Средний рост равен: дети 4-6 лет – 120 см, дети 6-10 лет – 135 см, дети 10-14 лет – 146 см. Середины интервалов факторного признака равны 5, 8 и 12 лет, а показатели силы связи равны:
- для 2-го интервала: (135-120)/(8-5) = 5 см/год;
- для 3-го интервала: (146-135)/(12-8) = 2,75 см/год;
- в целом: (146-120)/(12-5) = 3,71 см/год.

Правило сложения дисперсий

Пусть объекты совокупности разделены на группы, обозначим:

N – число объектов совокупности;

m – число групп;

n_i – число объектов j–й группы, $\sum n_j = N$;

y_{ij} – значение признака i–го объекта в j–й группе;

$\bar{y}_j$ – среднее значение признака в j–й группе;

$\bar{y}$ – среднее значение признака по всей совокупности.

Общая дисперсия равна:

$$s_{общ}^{\ 2} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2.$$

В соответствии с принципом разбиения объектов на группы дисперсия признака внутри каждой *j*–й группы должна быть относительно небольшой, она называется *остаточной (внутригрупповой) дисперсией* и равна:

$$s_j^{\ 2} = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2.$$

Средняя внутригрупповая (остаточная) дисперсия – средневзвешенная величина остаточных дисперсий всех групп:

$$s_{ост}^{\ 2} = \frac{1}{N}\sum_{j=1}^{m} n_j s_j^{\ 2} = \frac{1}{N}\sum_{j=1}^{m} n_j \frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2.$$

Межгрупповая (факторная) дисперсия – средневзвешенная величина квадратов отклонений групповой средней от общей средней:

$$s_{факт}^{\ 2} = \frac{1}{N}\sum_{j=1}^{m}(\bar{y}_j - \bar{y})^2 n_j.$$

Правило сложения дисперсий: общая дисперсия равна сумме средней остаточной дисперсии и факторной дисперсии:

$$s_{общ}^{\ 2} = s_{ост}^{\ 2} + s_{факт}^{\ 2}.$$

Правило сложения дисперсий

$$\sum\sum(y_{ij} - \bar{y})^2 = \sum\sum(y_{ij} - \bar{y}_j)^2 + \sum(\bar{y}_j - \bar{y})^2 n_j$$

Доказательство. Получим вспомогательное выражение для суммы квадратов разностей значений признака для *j*–й группы и среднего значения для всех объектов. Из каждой такой разности вычтем и прибавим среднее значение признака в группе и представим названную сумму в виде трех слагаемых, среднее из которых равно нулю, т.к. сумма отклонений значений признака в группе от среднего значения в группе равно нулю. Имеем:

$$\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j + \bar{y}_j - \bar{y})^2 = \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 + 2(\bar{y}_j - \bar{y})\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j) +$$

$$+ \sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 = n_j s_j^{\ 2} + 0 + n_j(\bar{y}_j - \bar{y})^2.$$

Подставим полученное выражение в формулу общей дисперсии:

$$s_{общ}^{\ 2} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \frac{1}{N}\sum_{j=1}^{m}[n_j s_j^{\ 2} + n_j(\bar{y}_j - \bar{y})^2] = s_{ост}^{\ 2} + s_{факт}^{\ 2}.$$

Коэффициент детерминации – отношение факторной дисперсии и общей дисперсии:

$$\eta^2 = \frac{s_{факт}^{\ 2}}{s_{общ}^{\ 2}},$$

где *эмпирическое корреляционное отношение* η показывает, какую часть общей колеблемости результативного признака вызывает изучаемый фактор. Чем оно больше, тем теснее статистическая связь между факторным и результативным признаком.

Пример. Имеются 3 группы рабочих: с начальным, средним и высшим образованием. В каждой группе – 2 рабочих. Зарплата в 1-й группе – 4 и 6 тыс. руб., во 2-й – 6 и 8 тыс. руб., в 3-й – 8 и 10 тыс. руб.

Группа	Признак	Средний признак в группе	Квадрат отклонения от общей средней	Квадрат отклонения от групповой средней	Квадрат разности средних ×объем группы
j	y_{ij}	$\bar{y}_j$	$(y_{ij}-\bar{y})^2$	$(y_{ij}-\bar{y}_j)^2$	$(\bar{y}_j-\bar{y})^2 n_j$
1	4 и 6	5	9 и 1	1 и 1	4×2=8
2	6 и 8	7	1 и 1	1 и 1	0×2=0
3	8 и 10	9	1 и 9	1 и 1	4×2=8
Сумма	42	–	22	6	16
Среднее	7	–	3,67	1	2,67
Показатель	$\bar{y}$	–	$s_{общ}^2$	$s_{ост}^2$	$s_{факт}^2$

Из таблицы следует, что общая дисперсия зарплаты равна 3,67, остаточная дисперсия – 1, факторная дисперсия – 2,67. Выполняется равенство: 3,67=1+2,67. Эмпирическое корреляционное отношение равно

$$\eta = \sqrt{\frac{2{,}67}{3{,}67}} = 0{,}85,$$

т.е. выявлена тесная связь (85%) между уровнем образования и зарплатой.

Многомерная средняя

Многомерная группировка использует произвольное число группировочных признаков.

Имеется n объектов и исследуется некоторое их общее качество, которое характеризуется набором k положительных частных показателей (признаков), выраженных в различных единицах измерения. Требуется упорядочить объекты по интегральному измерителю данного качества.

Частные показатели относятся к двум *типам.* Для показателя 1-го типа увеличение его значения говорит об увеличении интегрального измерителя, а для показателя 2-го типа – о его снижении. Например, уровень здоровья населения оценивают с помощью средней продолжительности жизни (показатель 1-го типа) и коэффициента детской смертности (2-го типа).

Рассмотрим два случая.

Случай 1. Все частные показатели относятся к первому типу. Тогда *многомерной средней* для i–го объекта называют показатель

$$\bar{p}_i = \frac{1}{k}\sum_{j=1}^{k}\frac{x_{ij}}{\bar{x}_j},$$

где x_{ij} – значение j–го признака для i–го объекта, $\bar{x}_j$ – среднее значение j–го признака.

Свойство 1. В первом случае многомерная средняя может быть больше или меньше 1. Если все индивидуальные значения признаков меньше средних значений, то она меньше 1, если они больше средних значений, то она больше 1, если они равны средним значениям, то она равна 1.

Свойство 2. В первом случае среднее значение многомерных средних равно 1:

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n}\bar{p}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{j=1}^{k}\frac{x_{ij}}{\bar{x}_j} = \frac{1}{nk}\sum_{j=1}^{k}\frac{1}{\bar{x}_j}\sum_{i=1}^{k}x_{ij} = \frac{1}{nk}\sum_{j=1}^{k}\frac{1}{\bar{x}_j}\bar{x}_j n = \frac{kn}{nk} = 1.$$

Пример. Первый студент получил 12 баллов по истории и 30 баллов по физике, второй – 15 и 40, третий – 9 и 20. Оценим их успеваемость в целом.

Студент, i	Сумма баллов, j		Многомерная средняя, $\vec{p}_i$
	история	физика	
1	12	30	1,000
2	15	40	1,292
3	9	20	0,708
Среднее	12	30	1

Многомерная средняя равна: для 1-го студента: (12/12+30/30)/2=1, для 2-го студента: (15/12+40/30)/2=1,292, для 3-го студента: (9/12+20/30)/2=0,708. Итак, 2-й студент – лучший по успеваемости, 3-й студент – худший.

Случай 2. Имеются показатели обеих типов. Тогда *многомерной средней* для i–го объекта называют показатель

$$\bar{p}_i = \frac{1}{k}\sum_{j=1}^{k}s_j\frac{x_{ij}}{\bar{x}_j},$$

где s_j равен 1, если j–й показатель относится к первому типу и равен -1, если он относится ко второму типу.

Свойство 1′. Во втором случае многомерная средняя может иметь любой знак. Если все индивидуальные значения признаков равны средним значениям, то она: положительна, – если преобладают признаки 1-то типа; отрицательна, – если преобладают признаки 2-го типа; равна нулю, – если число признаков обеих типов одинаково.

Свойство 2′. Во втором случае среднее значение многомерных средних равно $2a$–1, где a – удельный вес признаков 1-го типа. При a=1 мы получаем первый случай. Если число признаков обеих типов одинаково, то среднее значение многомерных средних равно нулю.

Пример. Уровень жизни оценивают с помощью показателя ВВП на душу населения (Россия – 100%) и степени неравенства доходов

(коэффициент Джини). В 2009 г. данные показатели составили: Россия – 100% и 41, Болгария – 92% и 29,2, Румыния – 87% и 31 (Россия и страны-члены ЕС. 2007. С. 79). Первый признак относится к 1-му типу, а второй – ко 2-му, их средние значения – 93% и 33,73. Многомерная средняя равна: Россия – (100/93-41/33,73)= -0,14, Болгария +0,123, Румыния +0,016. Итак, Болгария лидировала по уровню жизни.

Кластерный анализ

Имеется n объектов, каждый характеризуется k признаками. Требуется разбить объекты на группы, объединяющие сходные (близкие по значениям признаков) объекты.

Представим i–й объект как точку k–мерного пространства (x_{i1}, …, x_{ik}). Эвклидову метрику здесь использовать нельзя, поскольку признаки измеряют в разных единицах, и сумма квадратов их значений лишена смысла. Поэтому определим безразмерный измеритель отклонений значений признаков.

Нормированная разность значений j–го признака для p–го и q–го объектов равна:

$$d_{pq,j} = \frac{x_{pj} - x_{qj}}{s_j},$$

где x_{pj} и x_{qj} – значения j-го признака для p–го и q–го объектов, s_j – среднее квадратическое отклонение значений j-го признака.

Рассмотрим случаи, когда признаки равноправны и не равноправны.

Случай 1. Признаки *равноправны,* тогда расстояние между p–м и q–м объектами в признаковом пространстве равно обычному эвклидовому расстоянию («теорема Пифагора»):

$$r_{pq} = \sqrt{\sum_{j=1}^{k} d_{pq,j}^{\,2}}.$$

Пример. Успеваемость оценивается числом прогулов и оценкой. Эти показатели равны: Петр – 2 и 5, Иван – 10 и 3, Глеб – 3 и 4. Дисперсия равна 12,67 и 0,67, тогда расстояние между Петром и Иваном равно

$$r_{ПИ} = \sqrt{\frac{(2-10)^2}{12{,}67} + \frac{(5-3)^2}{0{,}67}} = 3{,}33.$$

Расстояние между Петром и Глебом – 1,25, между Иваном и Глебом – 2,31. Если требуется образовать группу из двух студентов, расположенных наиболее близко друг к другу, тогда в нее войдут Петр и Глеб.

Случай 2. Признаки *не равноправны*, тогда расстояние между p–м и q–м объектами в признаковом пространстве равно взвешенному эвклидовому расстоянию:

$$r_{pq} = \sqrt{\sum_{j=1}^{k} d_{pq,j}^{\,2} w_j},$$

где w_j – вес (значимость) j–го признака, сумма всех w_j равна k. Значения весов субъективны, их обычно определяют методом экспертных оценок.

Пример. Значения признаков для объекта А равны 2 и 8, В – 3 и 4, С – 4 и 9. Дисперсии значений признаков равны 0,67 и 4,67. Если признаки равноправны, то расстояния от А до В и С равны 2,22 и 2,49. Если же вес первого признака равен 0,4, а второго – 1,6, то расстояние от А до В равно

$$r_{AB} = \sqrt{\frac{(2-3)^2}{0{,}67}0{,}4 + \frac{(8-4)^2}{4{,}67}1{,}6} = 2{,}47.$$

Аналогично, расстояние от А до С станет равным 1,66. Таким образом, в первом случае В ближе к А, чем С, а втором случае – наоборот.

Кластер – группа объектов, расположенных близко друг другу. Это локальное скопление точек в заданном признаковом пространстве.

Задачи

1. Работники, обучавшиеся 4-10 лет, имеют средний доход 25 тыс. руб., 10-15 лет – 28 тыс. руб., 15-18 лет – 30 тыс. руб. Найти силу связи между образованием и доходом (с точн. до 0,01): а) для второй группы работников; б) для третьей группы, в) в среднем.

2. Имеются три группы рабочих: молодые (16-30 лет) – 2 чел., среднего возраста (30-60 лет) – 3 чел. и пенсионеры (старше 60 лет) – 2 чел. Зарплата равна: в 1-й группе – 13 и 11 тыс. руб., во 2-й – 16, 12 и 17 тыс. руб., в 3-й – 20 и 23 тыс. руб. Найти (с точн. до 0,01):

а) общую дисперсию зарплаты;

б) остаточную дисперсию;

в) факторную дисперсию;

г) эмпирическое корреляционное отношение.

3. Студенты получили следующие суммы баллов по истории, экономике и статистике: Антон – 40, 24 и 76, Борис – 50, 12 и 80, Семен – 30, 15 и 84. Найти многомерные средние для лучшего и худшего студентов.

4. Уровень здоровья населения оценивают с помощью ожидаемой продолжительности жизни при рождении и коэффициента материнской смертности (число умерших рожениц на 100 000 родившихся живых детей). В 2009 г. данные показатели составили: Россия – 68,7 и 16,5, Венгрия – 74 и 18,7, Румыния – 73,5 и 21,1 (РСЕ. 2011. С. 745, 747). Найти: а) многомерную среднюю для России и Венгрии; б) среднюю многомерную среднюю.

5. Уровень жизни населения оценивают с помощью индекса развития человеческого потенциала (ИРЧП) и среднесуточного потребления калорий на душу населения (см. табл.). Найти (с точн. до 0,0001):

а) дисперсию значений ИРЧП и потребления калорий;

б) расстояние от России до других стран.

Страна	ИРЧП	Потребление калорий
Россия	0,797	3118
Литва	0,857	3372
Польша	0,862	3366

Швеция	0,951	3208

Источник: Россия и страны-члены ЕС. 2007. С. 77.

3. ВЫБОРКА

Общие понятия

Из совокупности N объектов, которые характеризуются значениями признака x, выбраны n объектов ($n \leq N$). Полученная совокупность называется *выборкой*, а исходная совокупность – *генеральной совокупностью*.

Обозначим *среднюю величину* признака: в генеральной совокупности – μ, в выборке – $\bar{x}$. Обозначим *дисперсию* в генеральной совокупности – σ^2, в выборке – p. Характеристики генеральной совокупности называют *генеральными параметрами.*

Репрезентативная выборка – наиболее полно отражает свойства генеральной совокупности.

Повторная выборка – соответствует схеме возвратного шара. Вероятность попадания объекта в выборку равна $1/N$ и остается неизменной на протяжении процедуры отбора.

Бесповторная выборка – соответствует схеме безвозвратного шара. Поскольку в процессе отбора число элементов генеральной совокупности сокращается, вероятность попадания объекта в выборку изменяется от $1/N$ для первой отбираемой единицы до $1/(N-n+1)$ – для последней.

Пример. Телефонный опрос избирателей: в случае бесповторной выборки опрошенный избиратель исключается из дальнейшего опроса, а в случае повторной выборки ему могут позвонить несколько раз.

Недостатки метода повторного отбора:

- искажающий эффект – в выборке может повторяться один и тот же объект, в то время как в генеральной совокупности это недопустимо (в ней могут повторяться одинаковые значения признака, но не их носители). Из-за искажающего эффекта ошибки расчетов при повторной выборке больше, чем при бесповторной выборке;

- организация повторной выборки сопряжена с организационными сложностями, т.к. повторный опрос может вызвать негативную реакцию у опрашиваемых и они могут отказаться участвовать в дальнейшем опросе.

Преимущество повторного отбора – простота расчетных формул.

Ошибка выборки

Ошибка выборки (ошибка репрезентативности) – разница между характеристикой выборки и генеральным параметром. Она равна: для выборочной средней: $\varepsilon_{\bar{x}} = \bar{x} - \mu$; для дисперсии: $\varepsilon_{s^2} = s^2 - \sigma^2$.

Рассматриваются выборки одинакового объема с выборочными средними $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$.

Средняя ошибка выборочной средней – это среднее квадратическое отклонение выборочных средних от генеральной средней:

$$s_{\bar{x}} = \sqrt{\frac{\sum(\bar{x}_i - \mu)^2 f_i}{\sum f_i}},$$

где f_i – число выборок с одинаковым значением выборочной средней $\bar{x}_i$, $\sum f_i = m$ – общее число выборок данного объема. Если все f_i равны единице, то

$$s_{\bar{x}} = \sqrt{\frac{\sum \varepsilon_{\bar{x}}^2}{m}}.$$

Средняя ошибка выборочной средней в случае повторного отбора приближенно равна

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

где n – объем выборки.

Смысл средней ошибки выборочной средней: отклонение выборочной средней от генеральной средней в среднем равно $\pm s_{\bar{x}}$.

Пример. Задана генеральная совокупность: 2,4,12, ее генеральная средняя равна μ = (2+4+12)/3 = 6, генеральная дисперсия: σ^2 = (16+4+36)/3= 18,67. Составим все выборки объема 2 посредством повторного отбора.

Выборка 1-я: 2,4, средняя – 3, отклонение от генеральной средней – -3.

Выборка 2-я: 2,12, средняя – 7, отклонение от генеральной средней – 1.

Выборка 3-я: 4,12, средняя – 8, отклонение от генеральной средней – 2.

Средняя ошибка выборочной средней равна

$$s_{\bar{x}} = \sqrt{\frac{(-3)^2 + 1^2 + 2^2}{3}} = 2{,}16.$$

Приближенно: $$s_{\bar{x}} = \sqrt{\frac{18{,}67}{2}} = 3{,}05.$$

Из примера следует, что для выборок равного объема:

- среднее значение выборочных средних равно генеральной средней: (3+7+8)/3=6;

- сумма отклонений выборочной средней от генеральной средней равна нулю: (-3)+1+2=0.

Нормированное отклонение

Рассматривается выборка с выборочной средней $\bar{x}$.

Нормированное отклонение (t) – это отношение частной и средней ошибок выборочной средней:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}, \quad отсюда \quad \bar{x} = \mu + ts_{\bar{x}},$$

т.е. ошибка выборочной средней в среднем равна $\pm ts_{\bar{x}}$. Чем больше модуль нормированного отклонения, тем больше ошибка выборочной средней.

Пример. Генеральная совокупность: 2,4,12, генеральная средняя: – 6, средняя ошибка выборочной средней – 2,16. Нормированное отклонение

равно: для выборки «2,4» – (3-6)/2,16=-1,39, для выборки «2,12» – 0,46, для выборки «4,12» – 0,93.

Свойства нормированного отклонения.

1. Это *безразмерная* величина.

2. Поскольку для выборок равного объема среднее значение выборочных средних равно генеральной средней, *среднее значение* нормированного отклонения равно нулю: $\bar{t} = 0$.

3. *Дисперсия* значений нормированного отклонения равна 1.

Доказательство.

$$s_t^{\,2} = \frac{1}{m}\sum_{i=1}^{m}(t_i - \bar{t})^2 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\varepsilon_i}{s_{\bar{x}}} - 0\right)^2 = \frac{1}{s_{\bar{x}}^{\,2}} \times \frac{\sum_{i=1}^{m}\varepsilon_i^{\,2}}{m} = \frac{1}{s_{\bar{x}}^{\,2}} \times s_{\bar{x}}^{\,2} = 1,$$

где m – общее число выборок данного объема.

4. Распределение нормированного отклонения описывается *нормальной кривой*, т.е. вероятность его попадания в интервал $(-t,t)$ равна

$$F(t) = \frac{1}{\sqrt{2\pi}}\int_{-t}^{t} e^{-t^2/2} dt,$$

где $\pi = 3{,}14$, $e = 2{,}718$. При нулевом значении нормированного отклонения этот интеграл равен нулю, а при t увеличении интеграл увеличивается.

5. Вероятность попадания t в интервал (-3,3) равна 0,9973, т.е. с вероятностью почти 100% нормированное отклонение *меньше 3*.

Значения интеграла $F(t)$ вычислены и представлены в виде таблицы.

Табл.1. Значение интеграла вероятностей $F(t)$

t	Сотые доли				
	0	1	…	8	9
0,0	0000	0080	…	0638	0718
0,1	0797	0876	…	1428	1507
0,2	1585	1663	…	2205	2282
…	…	…	…	…	…
1,8	9281	9297	…	9399	9412
1,9	9425	9438	…	9523	9534
2,0	9545	9556	…	9625	9634
…	…	…	…	…	…

Используя таблицу 1, определим вероятность того, что модуль нормированного отклонения меньше 0,18. На пересечении строки «0,1» и столбца «8» стоит число 1428, т.е. искомая вероятность равна 0,1428 (14,3%).

Доверительная вероятность

Доверительная вероятность (P) – устанавливается нами при расчете выборочной средней или другой выборочной характеристики.

Доверительное отклонение (t) – нормированное отклонение, которое отвечает заданной доверительной вероятности $P=F(t)$ и определяется с помощью таблиц интеграла вероятностей.

Доверительное отклонение равно $t = F^{-1}(P)$, где F^{-1} – функция, обратная функции интеграла вероятностей, P – доверительная вероятность.

Доверительная ошибка (Δ) – произведение доверительного отклонения и средней ошибки выборочной средней: $\Delta = t \times s_{\bar{x}}$.

Доверительный интервал – интервал, в который попадает генеральная средняя μ с доверительной вероятностью $P=F(t)$, он имеет вид $(\bar{x} - \Delta, \bar{x} + \Delta)$, где $\bar{x}$ – некоторая выборочная средняя.

Алгоритм определения доверительного интервала.

1. Составить выборку объемом n, рассчитать выборочную среднюю $\bar{x}$ и выборочную дисперсию s^2.

2. Задать доверительную вероятность P.

3. Найти приближенно среднюю ошибку выборочной средней:

- для повторного отбора: $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$,

- для бесповторного отбора: $s_{\bar{x}} = \dfrac{s}{\sqrt{n}} \times \sqrt{1 - \dfrac{n}{N}}$.

Из приведенных формул следует, что у бесповторной выборки средняя ошибка выборочной средней меньше в $\sqrt{1-n/N}$ раз, чем у бесповторной выборки, что объясняется описанным выше искажающим эффектом. Однако чем меньше относительный объем выборки, тем меньше этот корректирующий коэффициент. Если выборка составляет 10% от генеральной совокупности, то он равен 0,95, при 5% – 0,97, при 2% – 0,99. При меньших объемах выборки можно пренебречь различием в способах отбора ее единиц из генеральной совокупности и использовать в расчетах более простую формулу для повторного отбора.

4. С помощью таблицы значений интеграла вероятностей по доверительной вероятности $P=F(t)$ найти доверительное отклонение t.

5. Рассчитать доверительную ошибку: $\Delta = t \times s_{\bar{x}}$.

6. Рассчитать границы доверительного интервала: $(\bar{x} - \Delta, \bar{x} + \Delta)$. Генеральная средняя попадет в него с вероятностью P.

Пример 1. Оценки по физике четырех студентов группы, отобранных по повторной схеме, равны 7,8,8,9. Используем описанный алгоритм.

1. Выборочная средняя – 8, выборочная дисперсия – (1+0+0+1)/4=0,5.

2. Зададим доверительную вероятность 97%.

3. Средняя ошибка выборочной средней для повторного отбора приближенно равна $\sqrt{0{,}5:4} = 0{,}35$.

4. По таблице определяем: если $F(t) = 0{,}97$, то $t = 2{,}17$.

5. Доверительная ошибка равна $2{,}17 \times 0{,}35 = 0{,}76$.

6. Границы доверительного интервала $8 \pm 0{,}76$.

Итак, с вероятностью 97% средняя оценка студентов группы находится в пределах (7,24; 8,76).

Пример 2. Из набора чисел 4, 8, 3, 5 получена выборка 4, 3 посредством бесповторного отбора. Ее выборочная средняя равна 3,5, дисперсия равна 0,25, тогда средняя ошибка выборочной средней приближенно равна

$$s_{\bar{x}} = \sqrt{\frac{0{,}25}{2}} \times \sqrt{1 - \frac{2}{4}} = 0{,}25.$$

При доверительной вероятности 97% доверительная ошибка равна 2,17×0,29=0,54, поэтому с этой вероятностью генеральная средняя больше 3,5-0,54=2,96 и меньше 3,5+0,54=4,04. На самом деле она равна 5 и лежит вне этого промежутка. Выборка нерепрезентативна, поскольку ее значения признака - наименьшие. Она относится к тем 3% выборок, на основе которых получаются ложные выводы о значении генеральной средней.

Если признак x является *булевым* (дихотомическим, альтернативным), т.е. принимает значения 0 или 1, тогда генеральная средняя равна удельному весу единиц в генеральной совокупности, а выборочная средняя (p) – их удельному весу в выборке. Пусть f_1 – удельный вес единиц, f_0 – удельный вес нулей, их сумма равна n – объему выборки. Тогда дисперсия равна

$$s^2 = \frac{(1-p)^2 f_1 + (0-p)^2 f_0}{f_1 + f_0} = \frac{p^2(f_1 + f_0)}{n} + \frac{(1-2p) f_1}{n} = p^2 + (1-2p)p = p - p^2.$$

Отсюда средняя ошибка выборочной средней приближенно равна:

- для повторного отбора: $s_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$,

- для бесповторного отбора: $s_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \times \sqrt{1 - \frac{n}{N}}$.

Доверительный интервал рассчитывают по описанному выше алгоритму с использованием одной из данных формул.

Пример 3. На таможне выборочно проверили 100 из 600 машин, при этом доля нарушений составила 10%. Поскольку каждую машину проверяют один раз, здесь реализован бесповторный отбор. Определим доверительный интервал при доверительной вероятности 95,4%. Средняя ошибка выборочной средней здесь равна:

$$s_{\bar{p}} = \sqrt{\frac{0{,}1(1-0{,}1)}{100}} \times \sqrt{1 - \frac{100}{600}} = 0{,}027.$$

Доверительное отклонение при P = 0,954 равно 2 (см. табл.), тогда доверительная ошибка равна $\Delta = 2 \times 0{,}027 = 0{,}054$. Итак, с вероятностью 95,4% удельный вес нарушителей среди водителей составит $0{,}1 \pm 0{,}05$, т.е. 5–15%.

Определение объема выборки

Исследуется генеральная совокупность объемом N. Задана доверительная вероятность P и доверительная ошибка (требуемая точность) Δ. Известна генеральная дисперсия σ^2 или некоторая выборочная дисперсия

s^2. Требуется определить объем выборки n, который обеспечит требуемую точность с заданной вероятностью.

Для повторной выборки требуемая точность равна

$$\Delta = t \times s_{\bar{x}} = \frac{t\sigma}{\sqrt{n}}, \quad отсюда \quad n_0 = \frac{t^2\sigma^2}{\Delta^2},$$

где n_0 – искомый объем выборки при повторном отборе.

Для бесповторной выборки требуемая точность равна

$$\Delta = t \times s_{\bar{x}} = \frac{t\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}}, \quad отсюда \quad n = \frac{n_0}{1 + \frac{n_0}{N}},$$

где n – искомый объем выборки при бесповторном отборе.

Пример. Магазин получил 200 ящиков картофеля и проверяет их вес. Из предыдущего опыта известна дисперсия веса ящиков – 3,5. Определим, сколько надо проверять ящиков, чтобы с вероятностью 95% отклонение веса от номинальной (оплаченной) величины не превышало 0,8 кг. Определим по таблице доверительное отклонение, оно равно 1,96. Имеем:

$$n_0 = \frac{1{,}96^2 \times 3{,}5}{0{,}8^2} = 21, \quad n = \frac{21}{1 + \frac{21}{200}} = 19.$$

Если ящик взвешивают один раз, требуется проверять 19 ящиков. Если после взвешивания его возвращают к остальным и могут взвесить повторно, то требуется проверять 21 ящик.

Для *дихотомического признака* (равен 1 или 0) генеральную дисперсию часто приравнивают ее максимальному значению 0,5×(1-0,5)=0,25. Отсюда

$$n_0 = \frac{t^2}{4\Delta^2}.$$

Пример. Получено 2000 электроламп. Допустимый процент брака – 0,1. Определим, сколько надо проверять ламп, чтобы с вероятностью 95,4% отклонение доли брака от нормы было не более 0,05. Имеем:

$$n_0 = \frac{2^2}{4 \times 0{,}05^2} = 400, \quad n = \frac{400}{1 + \frac{400}{2000}} = 333.$$

Если лампу проверяют один раз, требуется проверять 333 лампы, а при повторном отборе – 400. Тогда с заданной вероятностью брак составит 5-15%. Но есть более рациональное решение: поскольку приближенное значение генеральной средней равно 0,1, рассчитаем генеральную дисперсию – 0,1×(1-0,1)=0,09. Тогда

$$n_0 = \frac{2^2 \times 0{,}09}{0{,}05^2} = 36.$$

Итак, данные о параметрах генеральной совокупности позволяют сократить число проверяемых ламп более чем в 10 раз.

Малая выборка

Таблицы интеграла вероятностей рассчитаны для выборок большого объема из бесконечно большой генеральной совокупности, поэтому при малых выборках погрешность становится значительной.

Малая выборка – выборка объемом менее 30 единиц, полученная из нормально распределенной генеральной совокупности.

Распределение Стьюдента – это распределение значений признака в малой выборке. Оно симметрично, задается функцией плотности распределении $f_n(t)$ и при увеличении объема выборки стремится к нормальному распределению, т.е. вероятность попадания переменной в интервал $(-t,t)$ равна

$$F_n(t) = \int_{-t}^{t} f_n(t)dt \to F(t) \quad при \quad n \to \infty.$$

Степень свободы (*d.f.*) – это число индивидуальных значений признака, которые нужно знать для вычисления искомой характеристики. При расчете дисперсии она равна $n-1$.

Уровень значимости (p) – вероятность, составляющая единицу в сумме с доверительной вероятностью: $p = 1 - P$, где P – доверительная вероятность.

Таблицы распределения Стьюдента устанавливают соответствие между тремя показателями: числом степеней свободы, уровнем значимости и отклонением t, которое называют значением t – критерия Стьюдента.

Табл. 2. Значение t – критерия Стьюдента

Число степеней свободы, *d.f.*	Уровень значимости, p		
	0,10	0,05	0,01
1	6,3138	12,706	63,657
2	2,9200	4,3027	9,9248
3	2,3534	3,1825	5,8409
…	…	…	…
28	1,7011	2,0484	2,7633
29	1,6991	2,0452	2,7564
30	1,6973	2,0423	2,7500

Доверительный интервал для малой выборки рассчитывают по описанному выше алгоритму с использованием табл. 2 и скорректированной формулы средней ошибки выборочной средней:

$$s_{\bar{x}} = \frac{s}{\sqrt{n-1}}.$$

С ростом объема выборки разница между прежним и скорректированным значениями сокращается, при объеме 10 она равна 5%.

Пример. Обследовали 20 рабочих фирмы, из них 4 опоздали на работу. Тогда доля опоздавших – 0,2, дисперсия дихотомической переменной – $0,2 \times 0,8 = 0,16$, средняя ошибка выборки:

$$s_{\bar{x}} = \sqrt{\frac{0{,}16}{20-1}} = 0{,}092.$$

Зададим доверительную вероятность 90%. В таблице распределения Стьюдента на пересечении строки *d.f.* = 19 и столбца p = 0,1 определим t = 1,73 (в табл.2 отсутствует). Тогда доверительная ошибка равна Δ = 1,73×0,092 = 0,16, а доверительный интервал $0{,}2 \pm 0{,}16$, т.е. с заданной вероятностью доля опаздывающих равна 4-36%.

Замечание. При использовании таблиц интеграла вероятностей (табл.1) t = 1,64 и Δ = 0,14, т.е. доверительный интервал меньше – 6-34%.

Статистическая проверка гипотез

Задан интервальный ряд, f_i – частота распределения, т.е. число объектов со значением признака, принадлежащим i-му интервалу. Сумма f_i равна числу единиц ряда n. Построим нормальное распределение с *теоретическими* частотами $\hat{f}_i$, используя параметры заданного *эмпирического* распределения:

1. Найдем нормированное отклонение для концов каждого интервала:

$$t_i = \frac{x_i^{\min} - \bar{x}}{s}; \qquad t_{i+1} = \frac{x_i^{\max} - \bar{x}}{s},$$

где t_i и t_{i+1} – нормированные отклонения для левого и правого конца i-го интервала, $x_i^{\min}$ и $x_i^{\max}$ – минимальное и максимальное значения признака в нем, $\bar{x}$ – среднее значение ряда, s – среднее квадратическое отклонение.

2. Найдем для нормального распределения вероятность попадания единицы наблюдения в каждый интервал:

$$P_i = (F(t_i) - F(t_{i+1}))/2, \text{ если } t_i, t_{i+1} < 0;$$
$$P_i = (F(t_{i+1}) - F(t_i))/2, \text{ если } t_i, t_{i+1} > 0;$$
$$P_i = (F(t_i) + F(t_{i+1}))/2, \text{ если } t_i < 0, t_{i+1} < 0,$$

где $F(t_i)$ – значение интеграла вероятностей, его ищем по таблице (табл.1).

3. Найдем теоретическую частоту для, округлив до целого число:

$$\hat{f}_i = n \times P_i.$$

Пример. Задан ряд, x_i' – середина i–го интервала, $\Delta x_i = x_i' - \bar{x}$.

i	x_i	x_i'	f_i	$(\Delta x_i)^2 f_i$	t_i	t_{i+1}	$F(t_i)$	$F(t_{i+1})$	P_i	$\hat{f}_i$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1	0-2	1	6	74,76	-1,84	-1,03	0,934	0,697	0,118	3
2	2-4	3	6	14,04	-1,03	-0,21	0,697	0,166	0,265	8
3	4-6	5	10	2,2	-0,21	0,60	0,166	0,451	0,309	9
4	6-8	7	5	30,5	0,60	1,41	0,451	0,841	0,195	6
5	8-10	9	3	59,94	1,41	2,22	0,841	0,974	0,066	2
∑	-	-	30	181,4	-	-	-	-	0,953	29

1. Найдем среднее значение, используя столбцы 3 и 4:

(1×6+3×6+5×10+7×5+9×3)/30 = 4,53.

2. Найдем дисперсию, разделив на 30 сумму элементов 5-го столбца: 181,4/30=6,05, отсюда s = 2,46.

3. Найдем нормированное отклонение для левого конца 1-го интервала:

(0-4,53)/2,46= - 1,84 и т.д. (столбцы 7, 8). Поскольку отклонение на правом конце предыдущего интервала равно его значению на левом конце последующего интервала, объем вычислений можно сократить.

4. Найдем значения интеграла вероятностей по таблице (столбцы 8, 9).

5. Найдем вероятности: в интервалах 1-2 используем первую формулу, в интервалах 4-5 – вторую, в интервале 3 – третью (столбец 10).

6. Найдем теоретические частоты, умножив на 30 элементы 10-го столбца и округлив до целых значений (столбец 11).

Проверим гипотезу о том, что распределение является нормальным, для этого рассчитаем *значение критерия* χ^2 (критерия Пирсона) по формуле:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i},$$

где f_i и $\hat{f}_i$ – частота фактического и теоретического распределений для i-го интервала, k – число интервалов. Чем больше χ^2, тем больше вероятность расхождения заданного распределения и нормального распределения.

Пример. Используя элементы столбцов 4 и 11 таблицы 2, получим:

$$\chi^2 = \frac{(6-3)^2}{3} + \frac{(6-8)^2}{8} + \frac{(10-9)^2}{9} + \frac{(5-6)^2}{6} + \frac{(3-2)^2}{2} = 4{,}28.$$

Имеется специальная таблица, которая по числу степеней свободы $d.f. = k - 3$ и значению χ^2 позволяет оценить вероятность расхождения заданного распределения с нормальным. Из нее следует, что при значениях χ^2 меньше 4,61 (к ним принадлежит 4,28) эта вероятность меньше 10%.

Задачи

1. Из генеральной совокупности 4, 8, 3, 5 составлены выборки (4,8,3), (4,3,5), (8,3,5) и (4,8,5). Найти (с точностью до 0,01):

а) генеральную среднюю и генеральную дисперсию;

б) выборочные средние и их среднее значение;

в) выборочные дисперсии и их среднее значение;

г) вероятность того, что произвольная выборочная средняя отличается от генеральной средней не более чем на: 0,1; 0,5; 0,8.

2. Из единиц генеральной совокупности объемом 200 составлена выборка объемом 60. Доверительная вероятность равна 96,43%. В предположении, что выборка составлена методом повторного отбора, рассчитан доверительный интервал – (12,54; 14,86). Найти (с точностью до 0,01):

а) доверительную ошибку и среднюю ошибку выборочной средней;

б) выборочную среднюю и выборочную дисперсию;

в) доверительную ошибку и доверительный интервал в предположении, что выборка составлена методом бесповторного отбора.

3. Покупателям супермаркета предлагают оценить качество обслуживания, для чего создано 4 пункта опроса в разных его отделах. 400 человек оказались довольны обслуживанием, а 100 – нет. За время опроса обслужено 2000 человек. Найти:

а) интервал, в который попадает средняя доля неудовлетворенных покупателей с вероятностью 92%;

б) лишние данные задачи.

4. Завод получил 400 медных листов и проверяет долю примесей. Из предыдущего опыта известно среднее квадратическое отклонение доли примесей – 3%. Требуется, чтобы с вероятностью 98% отклонение доли примесей от нормы не превышало 1,4%. Найти необходимый объем выборки:

а) при повторном отборе;

б) при бесповторном отборе.

5. Число посетителей ночного клуба – 250 чел. Среди них периодически выявляют и задерживают наркоманов, их доля обычно равна 9%. Требуется, чтобы с вероятностью 90% отклонение этой доли не превышало 4%. Сколько посетителей клуба требуется проверять?

6. На выходе из театра организован опрос: 10 зрителей поставили оценку «5», 15 зрителей – «4». Всего зрителей – 500. Найти:

а) среднюю ошибку выборочной средней по формулам большой и малой выборки (с точностью до 0,0001);

б) доверительное отклонение – для большой и малой выборки;

в) интервал, в который попадает средняя оценка с вероятностью 95% (с точностью до 0,1).

7. Ихтиологи ловят рыбу в озере, а затем сразу отпускают. 10 рыб имели икру, а 5 – нет. Найти интервал, в который попадает средняя доля рыб с икрой с вероятностью 90% (с точностью до 0,1%).

8. Из стопки 40 курсовых работ выбрали 10: три из них оценены на «5», остальные – на «4». Найти:

а) среднюю ошибку выборочной средней (с точностью до 0,001);

б) интервал, в который попадает средняя оценка хранящихся работ с вероятностью 99% (с точностью до 0,01).

9. В группе 8 студентов проведен выборочный устный опрос 6 человек со следующими оценками: 4, 5, 5, 4, 2, 4. Доверительная вероятность – 99%.

а) найти доверительный интервал;

б) принадлежит ли ему средняя оценка, если два других получат «5»?

в) принадлежит ли ему средняя оценка, если два других получат «1»?

г) не противоречат ли выводы «б» и «в» выводу «а»?

10. Частоты: 8, 12, 7, 6, 4. Первый интервал – (2,4). Найти:

а) среднее значение и дисперсию;

б) нормированные отклонения для концов первого интервала;

в) значения интеграла вероятностей для концов первого интервала;

г) теоретические частоты нормального распределения;

д) значение критерия χ^2,

е) наиболее близкое распределение к нормальному распределению из заданного распределения и распределения с частотами 6, 6, 10, 5, 3 и первым интервалом (0,2).

4. КОРРЕЛЯЦИЯ

Коэффициент корреляции

Исследуется n объектов, (x_i, y_i) – значения признаков i-го объекта (например, рост и вес студентов группы). Объекты изображаются точками плоскости.

Коэффициент корреляции – мера тесноты статистической связи признаков, рассчитывается по формуле:

$$r = s_{xy}/s_x s_y,$$

где в числителе – *ковариация* признаков

$$s_{xy} = (1/n)\sum(x_i - \bar{x})(y_i - \bar{y}),$$

а в знаменателе – произведение их средних квадратических отклонений:

$$s_x^2 = (1/n)\sum(x_i - \bar{x})^2.$$

Если коэффициент корреляции положителен, то между признаками имеется *прямая* статистическая связь, если он отрицателен – *обратная* связь, если он равен нулю, то корреляция отсутствует. Чем больше модуль коэффициента корреляции, тем сильнее статистическая связь признаков.

Коэффициент корреляции

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Из формулы коэффициента корреляции следуют его свойства.

Свойство 1. Коэффициент корреляции не определен, если хотя бы один из признаков принимает одинаковые значения (дисперсия равна нулю).

Свойство 2. Знак коэффициента корреляции совпадает со знаком ковариации. Поэтому коэффициент корреляции рассматривают как нормированную ковариацию.

Свойство 3. Если все точки лежат на восходящей прямой, то коэффициент корреляции равен 1, а если на нисходящей прямой – то он равен -1.

Доказательство. Для точек прямой отношение приростов переменных неизменно: $\Delta y_i = k\Delta x_i$, где число k положительно для восходящих прямых и отрицательно для нисходящих. Отсюда $s_{xy} = k \times s_x^2$, $s_y = |k| \times s_x$, поэтому $r = k/|k|$. Данное отношение равно 1 при положительном значении k, и равно -1 при отрицательном значении.

Пример. Дано: $x = 12,13,19,16$; $y = 4,9,1,2$. Обозначим: $\Delta x_i = x_i - \bar{x}$,

$\Delta y_i = y_i - \bar{y}$. Произведем расчеты в таблице:

i	x_i	y_i	Δx_i	Δy_i	$(\Delta x_i)^2$	$(\Delta y_i)^2$	$\Delta x_i \Delta y_i$
1	12	4	-3	0	9	0	0
2	13	9	-2	5	4	25	-10
3	19	1	4	-3	16	9	-12
4	16	2	1	-2	1	4	-2
Сумма	60	16	0	0	30	38	-24
Среднее	15	4	-	-	7,5	9,5	-6
Показатель	$\bar{x}$	$\bar{y}$	-	-	s_x^2	s_y^2	s_{xy}

Коэффициент корреляции равен: $r = (-6)/(7{,}5 \times 9{,}5)^{0,5} = 0{,}71$, его также можно рассчитать с использованием элементов строки таблицы «Сумма»:

$$r = (-24)/(30 \times 38)^{0,5} = 0{,}71.$$

Коэффициент Фехнера

Коэффициент Фехнера – грубый измеритель силы статистической связи признаков, он рассчитывается по формуле:

$$r_f = (\text{C} - \text{H})/(\text{C} + \text{H}),$$

где C – число совпадений знаков Δx_i и Δy_i, а H – число несовпадений ($\Delta x_i = x_i - \bar{x}$, $\Delta y_i = y_i - \bar{y}$). В знаменателе дроби – число объектов n.

Замечание. Если для некоторого объекта одно из отклонений равно нулю, то для него полагаем: 0,5C и 0,5H.

Свойство 1. Если коэффициент корреляции равен 1, то коэффициент Фехнера также равен 1, но не наоборот.

Свойство 2. Если коэффициент корреляции равен -1, то коэффициент Фехнера также равен -1, но не наоборот.

Свойство 3. Коэффициент Фехнера положителен, если число совпадений знаков больше, чем число несовпадений, и отрицателен в противном случае. Он равен нулю, если число совпадений равно числу несовпадений.

Пример. Дано: x = 0,9,11,20; y = 9,0,20,11. Найдем коэффициент Фехнера и коэффициент корреляции.

i	x_i	y_i	Δx_i	Δy_i	Знаки Δx_i и Δy_i	$\Delta x_i \Delta y_i$	$(\Delta x_i)^2$	$(\Delta y_i)^2$
1	0	9	-10	-1	Совпадают	10	100	1
2	9	0	-1	-10	Совпадают	10	1	100
3	11	20	1	10	Совпадают	10	1	100
4	20	11	10	1	Совпадают	10	100	1
Среднее	10	10	0	0	-	10	50,5	50,5

Коэффициент Фехнера равен (4-0)/(4+0) = 1, а коэффициент корреляции равен 10/50,5 = 0,2. Таким образом, данные показатели могут существенно различаться.

Коэффициент корреляции рангов

Исследуется n объектов, (x_i, y_i) – значения признаков i-го объекта. Расположим все значения x_i по убыванию, тогда порядковый номер в этом ряду i-го объекта называется *рангом* x_i и обозначается p_i^x. Аналогично определяется показатель p_i^y. Теперь вместо точек (x_i, y_i) будем исследовать точки (p_i^x, p_i^y). Обозначим разность рангов i–го объекта через $d_i = p_i^x - p_i^y$.

Коэффициент корреляции рангов – грубый измеритель силы статистической связи признаков, он рассчитывается по *формуле Спирмена*:

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n}.$$

Свойства коэффициента ранговой корреляции.

Свойство 1. Если точки (x_i, y_i) расположены на восходящей кривой, то коэффициент корреляции рангов равен 1. В частности, он равен 1, если коэффициент корреляции равен 1 (точки расположены на восходящей прямой).

Свойство 2. Если точки (x_i, y_i) расположены на нисходящей кривой, то коэффициент корреляции рангов равен -1. В частности, он равен -1, если коэффициент корреляции равен -1 (точки расположены на нисходящей прямой).

Свойство 3. Если у каждого объекта оба ранга равны между собой, то коэффициент корреляции рангов равен нулю.

Свойство 4. Если у всех объектов сумма рангов одинакова (равна n+1), то коэффициент корреляции рангов равен -1.

Пример. Дано: x =40,50,60; y = 7,8,6.

i	x_i	y_i	p_i^x	p_i^y	d_i	d_i^2
1	40	7	3	2	1	1
2	50	8	2	1	1	1
3	60	6	1	3	-2	4
∑	-	-	-	-	0	6

Коэффициент корреляции рангов равен $1 - 6\times6/(3^3-3) = -0{,}5$. Как видно из таблицы, использование формулы Спирмена не требует расчета средних значений признаков и отклонений от средних значений.

Замечание. При расчете коэффициента корреляции рангов полезно сделать простую проверку, используя тождество: $\sum d_i = 0$ (см. предпоследний столбец таблицы). Докажем это свойство:

$$\sum d_i = \sum (p_i^x - p_i^y) = \sum p_i^x - \sum p_i^y = (1+\ldots+n) - (1+\ldots+n) = 0.$$

Корреляция булевых признаков

Переменную называют *булевой (дихотомической, альтернативной)*, если она принимает значения 0 и 1. Обычно булеву переменную используют, когда объект может обладать неким качеством (1) или не обладать им (0).

Тогда среднее значение признака равно удельному весу объектов, обладающих соответствующим качеством.

Коэффициент корреляции булевых признаков равен

$$r_b = \frac{C - H}{C + H},$$

где С – количество совпадений значений признаков (оба равны 1 или 0), Н – количество несовпадений признаков (один равен 1, другой – 0).

Замечание. Формула расчета коэффициента совпадает с формулой Фехнера, однако они имеют разный смысл: в нашем случае сравниваются *значения* признаков, а в формуле Фехнера – *знаки отклонений* значений от средних значений. В отличие от коэффициента Фехнера, для расчета коэффициента корреляции булевых признаков не требуется рассчитывать средние значения признаков.

Свойство 1. В отличие от «классического» коэффициента корреляции, коэффициент корреляции булевых признаков определен в случае, когда один или оба признака принимают одинаковые значения.

Свойство 2. Коэффициент корреляции булевых признаков равен коэффициенту Фехнера.

Пример. Антон имеет квартиру и машину, Виктор имеет машину, но не имеет квартиры, Семен не имеет ни квартиры, ни машины. Обозначим наличие квартиры признаком x, а наличие машины – признаком y.

i	Индивид	x_i	y_i	Сравнение x_i и y_i
1	Антон	1	1	Совпадение
2	Виктор	0	1	Несовпадение
3	Семен	0	0	Совпадение

Число совпадений признаков – 2, несовпадений – 1, тогда коэффициент корреляции булевых признаков равен (2-1)/(2+1) = 0,33, т.е. статистическая связь признаков «есть квартира» и «есть машина» характеризуется показателем 33%.

Метод корреляционной решетки

Пусть число исследуемых объектов n велико, а число возможных значений признаков x и y мало. Тогда все объекты можно разделить на небольшое число групп с одинаковыми парами значений признаков и рассчитать коэффициент корреляции по специальным формулам, которые требуют значительно меньше вычислений, чем стандартные формулы. Описанный ниже метод называют *методом корреляционной решетки*.

Обозначим число значений признака x через m, тогда $x_1, \ldots, x_m$ – возможные значения этого признака. Обозначим число значений признака y через k, тогда $y_1, \ldots, y_k$ – его возможные значения. Число групп равно $m \times k$.

Обозначим через f_{ii} число объектов с одинаковыми парами (x_i, y_j). Тогда число объектов со значением x_i равно сумме элементов i-й строки матрицы $\{f_{ij}\}$:

$$f_i^{\,x} = \sum_{j=1}^{k} f_{ij}.$$

Число объектов со значением y_i равно сумме элементов j-го столбца матрицы $\{f_{ij}\}$:

$$f_j^{\,y} = \sum_{i=1}^{m} f_{ij}.$$

Сумма всех значений f_i^x равна сумме всех значений f_j^y равна сумме всех элементов матрицы f_{ij} и равна количеству объектов n.

Среднее значение каждого признака рассчитывают с учетом удельных весов объектов с тем или иным возможным значением признака:

$$\bar{x} = \sum_{i=1}^{m} \frac{f_i^{\,x}}{n} x_i, \qquad \bar{y} = \sum_{j=1}^{k} \frac{f_i^{\,y}}{n} y_j.$$

Коэффициент корреляции рассчитывается по старой формуле:

$$r = s_{xy}/s_x s_y,$$

где ковариация признаков рассчитывается с учетом численности групп:

$$s_{xy} = (1/n)\sum(x_i - \bar{x})(y_i - \bar{y}) f_{ij},$$

а средние квадратические отклонения – с учетом численности объектов с тем или иным значением признака:

$$s_x^{\,2} = (1/n)\sum(x_i - \bar{x})^2 f_i^x.$$

Коэффициент корреляции
(метод корреляционной решетки)

$$r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y}) f_{ij}}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 f_i^{\,x} \sum_{j=1}^{k}(y_j - \bar{y})^2 f_j^{\,y}}}.$$

Пример. Имеется 10 студентов, причем каждый характеризуется десятибалльной оценкой по предмету (8 или 9) и возрастом (20 или 25 лет). Имеется 4 студента со значениями признаков (8,20), 1 студент – (8,25), 2 студента – (9,20) и 3 студента – (9,25), т.е. студенты разбиты на 4 группы.

x_i	$y_1 = 20$	$y_2 = 25$	$f_i^{\,x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i^{\,x}$
8	4	1	5	-0,5	1,25
9	2	3	9	0,5	1,25
$f_j^{\,y}$	6	4	10	–	2,5
$y_j - \bar{y}$	-2	3	–		
$(y_j - \bar{y})^2 f_j$	24	36	60		

1. Рассчитаем средние значения признаков:

$$\bar{x} = (5/10)\times 8 + (5/10)\times 9 = 8{,}5,$$
$$\bar{y} = (6/10)\times 20 + (4/10)\times 25 = 22.$$

2. Рассчитаем сумму квадратов отклонений от среднего значения признака с учетом численности групп. Для первого признака этот показатель равен 2,5, для второго – 60 (см. таблицу).

3. Умножим численность каждой группы на соответствующие отклонения от средних значений признаков и сложим эти произведения:

$$4\times(-0{,}5)\times(-2) + 1\times(-0{,}5)\times 3 + 2\times 0{,}5\times(-2) + 3\times 0{,}5\times 3 = 5.$$

4. Рассчитаем коэффициент корреляции:

$$r = \frac{5}{\sqrt{2{,}5\times 60}} = 0{,}41.$$

Итак, чем больше возраст студента, тем больше оценка по предмету.

Замечание. По виду матрицы $\{f_{ij}\}$ можно было догадаться, что между возрастом и оценкой имеется положительная корреляция, поскольку ее диагональные элементы больше не диагональных. Однако в случае диагональной матрицы коэффициент корреляции не обязательно равен единице. Для этого необходимо, чтобы все точки (x_i, y_j) располагались на прямой линии.

Применение: теория портфеля

Инвестор вкладывает средства в акции m видов. Объектом наблюдения служат ежедневные значения доходности акций. Имеется n наблюдений.

Обозначим через R_i^k доходность акций k-го вида в i-й день, тогда *средняя доходность акций* этого вида за n дней равна

$$R_k = \frac{\sum_{i=1}^{n} R_i^k}{n}.$$

Риск акций k–го вида оценивают дисперсией доходности:

$$s_k^2 = \frac{\sum_{i=1}^{n} (R_i^k - R_k)^2}{n}.$$

Портфелем называют вектор $(x_1, \ldots, x_m)$, где x_k – удельный вес средств, вложенных в акции k-го вида, сумма его элементов равна единице ($0 \le x_i \le 1$).

Доходность портфеля (R) есть средневзвешенная доходность акций, рассчитанная с учетом их долей в портфеле:

$$R = \sum_{k=1}^{m} x_k R_k.$$

Риск портфеля (s^2) рассчитывают по формуле:

$$s^2 = \sum_{k=1}^{m} x_k^2 s_k^2 + 2\sum_{k,j=1 k\neq j}^{m}\sum x_k x_j s_{kj},$$

где s_{kj} – ковариация доходностей акций k-го и j-го вида.

Рассмотрим три случая задачи *минимизации риска* портфеля.

Случай 1. Имеется два вида акций, доходность портфеля не учитывается. Поскольку сумма x_1 и x_2 равна единице, риск портфеля есть квадратичная функция удельного веса акций 1-го вида:

$$s^2 = x_1^2 s_1^2 + (1 - x_1)^2 s_2^2 + 2x_1(1 - x_1)s_{12}.$$

Определим точку минимума параболы и получим формулу расчета оптимальной доли акций первого вида:

$$x_1^* = \frac{s_2^2 - s_{12}}{s_1^2 + s_2^2 - 2s_{12}} \ .$$

Замечания:

1. Данная формула неприменима, если ковариация превосходит риск доходности акций второго вида. В этом случае корреляция доходности акций настолько велика, что их можно рассматривать как акции одного вида. Вообще, многие алгоритмы статистического анализа не допускают рассмотрение высоко коррелированных переменных. Если данная формула дает число большее единицы, то в качестве оптимального принимаем портфель, состоящий только из акций первого вида. Если же она дает отрицательное число, то оптимальный портфель состоит из акций второго вида.

2. Если коэффициент риски некоторых акций те же, но ковариация доходностей равна нулю, тогда доля первых акций равна $s_2^2/(s_1^2 + s_2^2)$, доля вторых – $s_1^2/(s_1^2 + s_2^2)$, а риск наименее рисованного портфеля равен $s_1^2 s_2^2/(s_1^2 + s_2^2)$, что меньше риска акций каждого вида.

Пример 1. Заданы доходности акций двух видов за три дня (в %):

i	R_i^1	R_i^2	$(\Delta R_i^1)^2$	$(\Delta R_i^2)^2$	$\Delta R_i^1 \Delta R_i^2$
1	4	7	1	1	-1
2	5	8	0	4	0
3	6	3	1	9	-3
Сумма	15	18	2	14	-4
Среднее	5	6	2/3	14/3	-4/3
Показатель	R_1	R_2	s_1^2	s_1^2	s_{12}

1. Оптимальные доли акций равны:

$x_1^* = (2/3 – (-4/3))/(2/3 +14/3 – 2(-4/3)) = 0{,}75$, отсюда $x_2^* = 0{,}25$.

2. Доходность портфеля равна: $0{,}75\times5 + 0{,}25\times6 = 5{,}25$ (%).

3. Риск наименее рискованного портфеля равен

$$s_*^2 = 0{,}75^2\times2/3 + 0{,}25^2\times14/3 + 2\times0{,}75\times0{,}25\times(-4/3) = 0{,}17.$$

Как мы видим, риск оптимального портфеля (0,75; 0,25) значительно меньше, чем риск наименее рискованных акций ($0{,}17 \le 0{,}67$). Кроме того, он также существенно меньше риска наименее рискованного портфеля (0,875; 0,125), рассчитанного для случая акций с теми же рисками, но нулевой ковариацией ($0{,}17 \le 0{,}58$). Таким образом, учет фактора корреляции доходностей позволил снизить риск портфеля более чем в три раза.

Вывод. Для минимизации риска портфеля необходимо включать в него пары активов с отрицательным коэффициентом корреляции доходностей. Так, объемы выручки кафе от продажи чая и от продажи минеральной воды зависят от погоды и при этом характеризуются отрицательной корреляцией, поскольку в холодные дни покупают преимущественно чай, а в теплые дни – воду. Поэтому в меню должны быть чай и вода, и тогда выручка будет слабо зависеть от погоды, т.е. ее дисперсия будет меньше, чем дисперсия выручки от чая и дисперсия выручки от воды.

Эффект «слон-муравьей». Величина риска акций зависит от их средней доходности: при прочих равных условиях акции с большей средней доходностью имеют большую дисперсию. Приведем аналогию: дисперсия роста слонов больше, чем дисперсия роста муравьев, хотя среднее квадратическое отклонение, деленное на средний рост (вариация), может оказаться меньше у слонов. В случае портфеля этот эффект негативно влияет на доходность оптимального портфеля, поскольку из-за завышенной дисперсии (риска) доля более доходных акций («слонов») оказывается необоснованно заниженной, а оптимальный портфель – деформированным в пользу менее доходных акций («муравьев»). Для преодоления данной сложности следует перед выполнением алгоритма определения оптимального портфеля скорректировать исходные значения доходности таким образом, чтобы все акции имели одинаковую среднюю доходность. В случае двух акций этой цели можно достичь, умножив значения доходности более доходных акций на отношение средней доходности менее доходных акций и средней доходности более доходных акций.

Используем данные Примера 1 и скорректируем заданные значения доходности второй акции, которая имеет наибольшую доходность (6>5). Умножим их на 5/6, получим следующие значения: 5,83%, 6,67%, 2,50%. Составим таблицу, повторим расчеты и получим: риск акций второго вида – 3,24 (вместо 4,67) а ковариация доходностей – -1,11 (вместо -1,33), оптимальная доля акций первого вида – 0,71 (вместо 0,75), риск оптимального портфеля – 0,15 (вместо 0,17). Как и ожидалось, устранение эффекта «слона-муравья» привело к росту удельного веса более доходных акций в оптимальном портфеле на 4 п.п.

Случай 2. Число акций произвольно, доходность портфеля не учитывается. Минимизируем квадратичную функцию риска портфеля s^2 при линейных ограничениях на переменные:

$$0 \le x_i \le 1, \quad \sum_{k=1}^{m} x_k = 1, \quad \sum_{k=1}^{m} x_k R_k \ge R_0,$$

где R_0 – требуемая минимальная доходность портфеля. Эта задача квадратичного программирования может иметь два вида решений: внутренние и угловые. Для углового решения хотя бы одно неравенство превращается в равенство, а для внутреннего решения все неравенства превращаются в строгие неравенства.

Внутреннее решение представляет собой условный экстремум задачи минимизации целевой функции риска s^2 при условии, что сумма независимых переменных равна единице. Функция Лагранжа имеет вид:

$$L = \sum_{k=1}^{m} x_k^{\,2} s_k^{\,2} + 2\sum_{k,j=1}^{m}\sum_{k\neq j}^{m} x_k x_j s_{kj} - \lambda(\sum_{k=1}^{m} x_k - 1),$$

где λ – множитель Лагранжа. Приравняв нулю ее частные производные, получим систему линейных уравнений с m+1 неизвестными для определения внутреннего оптимального решения (показан случай трех видов акций):

$$A \times X = B, \text{ где}$$

$$A = \begin{pmatrix} s_1^{\,2} & s_{12} & s_{13} & -1 \\ s_{12} & s_2^{\,2} & s_{23} & -1 \\ s_{13} & s_{23} & s_3^{\,2} & -1 \\ 1 & 1 & 1 & 0 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Матрица системы A является квадратной и имеет блочный вид, ее первый блок размерности $m \times m$ является симметричной матрицей, у которой на главной диагонали расположены риски акций, а недиагональные элементы равны значениям ковариации доходности для соответствующих пар акций.

Определим экономический смысл переменной x_4. Для этого первое уравнение системы умножим на x_1, второе – на x_2, третье – на x_3, затем полученные равенства сложим. Убедимся, что решение системы x_4 равно минимальному риску портфеля, т.е.

$$(x_4)_* = s^2_{min}.$$

Обозначим: $C = A^{-1}$, тогда решение системы: $X = C \times B$. Учитывая вид вектора B, заключаем, что координаты оптимального портфеля и его риск находятся в последнем столбце матрицы С:

$$x_i = c_{i,m+1}\ (i \le m), \quad s^2_{min} = c_{m+1,m+1}.$$

Частный случай. Из матрицы системы следует, что если все коэффициенты ковариации равны 0 (первый блок – диагональная матрица), то удельный вес каждой акций в оптимальном портфеле обратно пропорционален ее риску. Пусть риски акций – 0,5, 1,5 и 2,5, а все коэффициенты ковариации равны 0. Тогда доли акций в оптимальном портфеле пропорциональны числам 1/0,5, 1/1,5 и 1/2,5. Их сумма равна 3,067, поэтому доля акций 1-го вида равна 2/3,067=0,652, 2-го вида – 0,667/3,067=0,217, 3-го вида – 0,4/3,067=0,131. Риск оптимального портфеля равен 0,326, что меньше риска наименее рискованных акций (0,5).

Пример 2. Заданы доходности акций трех видов за четыре дня (в %), требуемая доходность портфеля не задана:

i	R_i^1	R_i^2	R_i^3
1	4	5	2
2	3	5	1
3	4	4	5
4	5	2	4

Расширим заданную таблицу, произведем необходимые вычисления (см. пример выше) и получим матрицу системы:

$$A = \begin{pmatrix} +0{,}50 & -0{,}75 & +0{,}75 & -1{,}00 \\ -0{,}75 & +1{,}50 & -1{,}25 & -1{,}00 \\ +0{,}75 & -1{,}25 & +2{,}50 & -1{,}00 \\ +1{,}00 & +1{,}00 & +1{,}00 & 0{,}00 \end{pmatrix}.$$

Решим систему уравнений с четырьмя неизвестными, обратив матрицу А с помощью программы Microsoft Excel согласно алгоритму:

1. Введите обращаемую матрицу.

2. Подготовьте место для обратной матрицы, используя левую кнопку мыши. Клетки пустой матрицы – голубые, за исключением верхней левой.

3. Нажмите кнопку «f_x», вызовите функцию «МОБР», при этом появится окно «Аргумент функции» с полем «Массив».

4. Выделите обращаемую матрицу с помощью левой кнопки мыши, при этом в поле «Массив» появятся координаты матрицы (например, «R[-4]…»). Под ними можно увидеть первый элемент обратной матрицы.

5. Нажмите на кнопку «ОК» в окне «Аргумент функции», при этом в верхней левой клетке подготовленной пустой матрицы появятся соответствующий элемент обратной матрицы.

6. Для того чтобы получить все элементы обратной матрицы нажмите клавишу «F2», а затем одновременно три клавиши: «Ctrl», «Shift» и «Enter».

Обратная матрица A^{-1} представлена в табл.1.

Табл.1. Расчет оптимального портфеля: доходность не задана

№	обратная матрица системы				портфель
1	1,386667	-0,48	-0,90667	0,626667	0,627
2	-0,48	0,32	0,16	0,36	0,360
3	-0,90667	0,16	0,746667	0,013333	0,013
4	-0,62667	-0,36	-0,01333	0,053333	s^2 = 0,053

Отсюда следует, что наименее рискованный портфель:

$$X_* = (0{,}627;\ 0{,}360,\ 0{,}013).$$

Его риск равен x_4 = 0,053, что в десять раз меньше риска наименее рискованных акций (первого вида): 0,053≤0,5. Сравним полученное решение с оптимальным портфелем, полученным в предыдущем примере для случая акций с теми же значениями риска, но нулевыми значениями корреляции доходностей: риск такого портфеля X_0 = (0,652; 0,217; 0,131) равен 0,326, что в шесть раз больше, чем в случае с ненулевыми коэффициентами ковариации. Как мы видим, учет фактора корреляции доходностей привел к снижению доли первых акций в оптимальном портфеле на 2,5 процентных пункта, увеличению доли вторых акций на 14,3 п.п. и сокращению доли

третьих акций на 11,8 п.п., или в десять раз. Доходность оптимального портфеля равна 3,987.

Случай 3. Число акций произвольно, доходность портфеля задана. Предположим, что требуемая доходность портфеля в точности равна R_0, тогда ограничения функции риска портфеля имеют вид линейных уравнений:

$$\sum_{k=1}^{m} x_k = 1, \quad \sum_{k=1}^{m} x_k R_k = R_0.$$

Функция Лагранжа принимает более сложный вид:

$$L = \sum_{k=1}^{m} x_k{}^2 s_k{}^2 + 2\sum_{k,j=1k\neq j}^{m}\sum^{m} x_k x_j s_{kj} - \lambda(\sum_{k=1}^{m} x_k - 1) - \mu(\sum_{k=1}^{m} x_k R_k - R_0),$$

где μ – второй множитель Лагранжа. Приравняв нулю ее частные производные, получим систему линейных уравнений с m+2 неизвестными для определения внутреннего оптимального решения (показан случай трех видов акций):

$$A \times X = B, \text{ где}$$

$$A = \begin{pmatrix} s_1^2 & s_{12} & s_{13} & -1 & -0{,}5R_1 \\ s_{12} & s_2^2 & s_{23} & -1 & -0{,}5R_2 \\ s_{13} & s_{23} & s_3^2 & -1 & -0{,}5R_3 \\ 1 & 1 & 1 & 0 & 0 \\ R_1 & R_2 & R_3 & 0 & 0 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ R_0 \end{pmatrix}.$$

Умножим первое уравнение системы умножим на x_1, второе – на x_2, третье – на x_3, затем полученные равенства сложим. Убедимся, что минимальный риск портфеля является линейной комбинацией компонентов оптимального портфеля x_{m+1} и x_{m+2} и рассчитывается по формуле

$$s^2{}_{min} = x_{m+1} + 0{,}5R_0 x_{m+2}.$$

Обозначим: $C = A^{-1}$, тогда $X = C \times B$. Учитывая вид вектора B, заключаем, что компоненты оптимального портфеля равны:

$$x_i = c_{i,m+1} + R_0 c_{i,\mathrm{m+2}} \ (i \le m).$$

Пример 3. Воспользуемся данными Примера 2 и предположим, что задана требуемая доходность 3,9%. Дополним матрицу A пятой строкой (4; 4; 3; 0; 0) и пятым столбцом (-2; -2; -1,5; 0; 0). Вектор правых частей системы равен B = (0; 0; 0; 1; 3,9). Обратим матрицу A (см. табл.2).

Табл.2. Расчет оптимального портфеля: доходность задана

№	обратная матрица системы					портфель
1	0,285714	-0,28571	8,97E-16	-4,21429	1,214286	0,521
2	-0,28571	0,285714	-3,9E-16	1,214286	-0,21429	0,379
3	1,34E-15	0	-6,7E-16	4	-1	0,100
4	4,214286	-1,21429	-4	21,33929	-5,33929	-
5	-2,42857	0,428571	2	-10,6786	2,678571	s^2 = 0,064

Рассчитаем удельный вес акций первого вида в наименее рискованном портфеле: x_1 = -4,21439 + 3,9×1,214286 = 0,521. Аналогично: x_2 = 0,379, x_3 = 0,1, x_4 = 0,516, x_5 = - 0,232. Риск оптимального портфеля: s^2_{min} = 0,516 + 0,5×3,9×(- 0,232) = 0,064. Как мы видим, требование обеспечить доходность портфеля на уровне 3,9% привело к увеличению его риска (0,064>0,053).

Замечания:

1. Перед решением описанной системы уравнений рекомендуется исключить из рассмотрения акции с доходностью ниже требуемого значения, а также исключить акции с высоким (близким к единице) значением парной корреляции: для таких акций некий коэффициент ковариации больше или равен дисперсии. Для оставшихся акций необходимо скорректировать значения доходности, т.е. устранить эффект «слона-муравья».

2. Если требуется учесть в решении все заданные виды акций, то метод может оказаться неприменимым (удельные веса акций больше 1 или меньше 0). В этом случае можно использовать готовые программы (пакеты) для решения задач квадратичного программирования либо применить более грубый *метод перебора*: разбить единичный промежуток изменения каждой переменной x_i на равные промежутки длиной h и с помощью несложной программы рассчитать на компьютере риски и доходности каждого портфеля, выбрав среди них оптимальный. Количество всех возможных портфелей равно h^{-m}. Так, при выбранной длине шага 0,1 и числе видов акций 6 общее количество портфелей равно 10^6, что позволяет достаточно быстро получить приближенную оценку параметров оптимального портфеля.

Метод перебора позволяет решить задачу, которая не имеет простых математических или программных алгоритмов, – задачу о нахождении оптимального портфеля по критерию доходность-риск, т.е. максимизации функции R/s при прежних ограничениях.

Задачи

1. Найдите коэффициент корреляции:
а) x =12,10,14,9; y = 6,3,-2,5;
б) x =9,13,7,5; y = 18,19,15,20;
в) x =20,13,17,14; y = 6,2,5,-1.
2. Найдите коэффициент Фехнера:
а) x =12,10,14,9; y = 6,3,-2,5;
б) x =9,13,7,5; y = 18,19,15,20;
в) x =20,13,17,14; y = 6,2,5,-1.
3. Найдите коэффициент корреляции рангов:
а) x =12,10,14,9; y = 6,3,-2,5;
б) x =9,13,7,5; y = 18,19,15,20;
в) x =20,13,17,14; y = 6,2,5,-1.
4. Найдите коэффициент корреляции булевых признаков:
а) x =1,1,1; y = 0,0,1;
б) x =1,1,0,0; y = 0,0,1,1;

в) x = 1,0,1,0,0; y = 0,0,1,1,0.

5. Найдите коэффициент корреляции, если имеется 50 студентов с оценками по двум предметам 2 и 3 балла, 20 студентов с оценками 3 и 5 баллов и 30 студентов с оценками 5 и 9 баллов.

6. Найдите удельный вес акций первого вида в наименее рискованном портфеле и его риск, если доходность акций за 4 дня составила:

а) акции 1-го вида: 3,9,7,5%; акции 2-го вида: 10,8,16,14%;

б) акции 1-го вида: 20,12,15,17%; акции 2-го вида: 3,8,7,4%;

в) акции 1-го вида: 11,18,17,14%; акции 2-го вида: 6,10,8,4%.

7. Заданы доходности акций трех видов за четыре дня. Используя программу Microsoft Excel, найдите (с точн. до 0,001):

а) риски акций;

б) коэффициенты ковариации доходности между акциями 1-го вида и акциями 2-го и 3-го видов;

в) наименее рискованный портфель и его риск.

i	R_i^1	R_i^2	R_i^3
1	3	6	5
2	6	4	7
3	8	4	5
4	3	6	3

8. Дайте ответ (верно/неверно)

1) Коэффициент корреляции измеряет силу воздействия одного признака на другой

2) Ковариация определена для любых значений признаков

3) Если дисперсии признаков равны 1, то коэффициент корреляции равен ковариации

4) Ковариация не может равняться 20

5) Если коэффициент корреляции равен – 1, все точки лежат на прямой

6) Коэффициент корреляции определен для любых значений признаков

7) Ковариация равна среднему произведению значений x и y минус произведение средних x и y

8) Дисперсия равна произведению средней минус средний квадрат x

9) X = 3,1,5, Y = 4,2,7. Тогда коэффициент корреляции положителен

10) X = 0,2,2,0,0, Y = 0,2,0,2,2. Тогда коэффициент корреляции равен 0

11) Коэффициент Фехнера измеряет силу статистической связи показателей

12) Отклонения от средних: 2,-4,2 (x), -6,3,3 (y). Тогда коэффициент Фехнера меньше 0,1

13) Коэффициент Фехнера: сравнивают значения признаков

14) X = 4,3,5, Y = 5,4,3. Тогда коэффициент Фехнера больше 0,1

15) Ранговая корреляция: рассчитывают средние значения признаков

16) Для всех объектов ранги равны, тогда корреляция рангов равна 1

17) Точки лежат на кривой $1/x$, тогда корреляция рангов равна -1

18) Корреляция рангов равна 1, тогда коэффициент корреляция равен 1

19) Знаменатель в формуле Спирмена зависит от числа признаков

20) При четырех объектах знаменатель в формуле Спирмена равен 12

21) При 10 объектах знаменатель в формуле Спирмена равен 990

22) Булева корреляция больше 0,58, если число совпадений в четыре раза больше числа несовпадений

23) Метод корреляционной решетки (МКР): объектов - 16, групп - 4. Тогда ковариация равна сумме 16 слагаемых

24) МКР: дисперсия x зависит от перемещения объектов между группами с одинаковым значением x

25) МКР: ковариация зависит от перемещения объектов между группами с одинаковым значением x

26) Доходность портфеля – это средневзвешенная значений доходности акций

27) Риск портфеля – это средневзвешенная величина риска акций

28) Теория портфеля: число значений признаков равно числу видов акций

29) Теория портфеля: при 5 видах акций рассчитывают 10 показателей ковариации

30) Для минимизации риска портфеля выбирают пары акций с отрицательной ковариацией

31) Риск портфеля из двух акций – гиперболическая функция от доли акций первого вида

32) Ковариация равна 0, риски акций равны 1, тогда риск портфеля (0,4, 0,6) меньше 0,45

5. ПАРНАЯ РЕГРЕССИЯ

Метод наименьших квадратов (МНК)

Исследуется n объектов, (x_i, y_i) – значения признаков i-го объекта (например, рост и вес студентов группы). Объекты изображаются точками плоскости. Имеется несколько классов функций: прямые, параболы и т.д.

Для некоторой функции f сумма квадратов отклонений $f(x_i)$ от y_i равна

$$d = (y_1 - f(x_1))^2 + \ldots + (y_n - f(x_n))^2.$$

Число d показывает, насколько близко точки (объекты) расположены к графику функции, т.е. насколько точно график описывает множество (x_i, y_i).

Регрессия в классе функций ($\tilde{y}$) – функция с наименьшим значением d. Также регрессией называют график этой функции.

Наилучший вид регрессии для данных объектов – тот, которому отвечает минимум из минимальных значений d.

Алгоритм МНК. Пусть функции некоторого класса задаются параметрами a и b (например, прямые: $a + bx$), тогда $d(a,b)$ - функция двух переменных. Для расчета параметров регрессии определим значения a и b,

обеспечивающие минимум функции *d*. Для этого приравняем нулю ее частные производные и решим систему уравнений относительно *a* и *b:*

$$\partial d/\partial a = 0, \quad \partial d/\partial b = 0.$$

Параметры регрессии – решение данной системы. Если функции задаются тремя параметрами, то решаем систему с тремя неизвестными и т.д. Основные виды регрессии – линейная и параболическая.

Линейная регрессия

Линейная регрессия – это прямая линия:

$$\tilde{y} = a + bx.$$

Используется, когда заданные точки лежат вблизи прямой.

Применим алгоритм МНК, для этого запишем функцию:

$$d = (y_1 - a - bx_1)^2 + \ldots + (y_n - a - bx_n)^2.$$

Уравнение 1. Продифференцируем *d* по *a*, получим:

$$(-2)(y_1 - a - bx_1) + \ldots + (-2)(y_n - a - bx_n) = 0, \text{ или}$$

$$\sum y_i - na - b\sum x_i = 0.$$

Разделим последнее равенство на *n*, получим:

$$\bar{y} = a + b\bar{x}.$$

Свойство 1 линейной регрессии следует из последнего равенства: линейная регрессия проходит через точку $(\bar{x}, \bar{y})$ с координатами, равными средним значениям признаков.

Уравнение 2. Продифференцируем *d* по *b*, получим:

$$(-2x_1)(y_1 - a - bx_1) + \ldots + (-2x_n)(y_n - a - bx_n) = 0, \text{ или}$$

$$\sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0.$$

Разделим последнее равенство на *n*, получим:

$$\overline{xy} - a\bar{x} - b\overline{x^2} = 0.$$

Решение системы. Подставим *a* из первого уравнения во второе:

$$\overline{xy} - (\bar{y} - b\bar{x})\bar{x} - b\overline{x^2} = 0, \text{ отсюда } b = s_{xy}/s_x^2, \text{ где}$$

- в числителе – ковариация признаков:

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}, \text{ или } s_{xy} = (1/n)\sum (x_i - \bar{x})(y_i - \bar{y}).$$

- в знаменателе – дисперсия *x*:

$$s_x^2 = \overline{x^2} - \bar{x}^2, \text{ или } s_x^2 = (1/n)\sum (x_i - \bar{x})^2.$$

Из формулы для *b* следуют еще три свойства линейной регрессии.

Свойство 2. Линейная регрессия не существует, если все значения *x* одинаковы (дисперсия *x* равна нулю).

Свойство 3. Линейная регрессия совпадает с осью абсцисс ($\tilde{y} = 0$), если все значения признака *y* равны нулю.

Свойство 4. Линейная регрессия совпадает с осью абсцисс ($\tilde{y} = 0$), если ковариация (корреляция) признаков равна нулю.

Докажем основное свойство линейной регрессии.

Свойство 5. Сумма отклонений значений линейной регрессии от соответствующих значений *y* равна нулю:

$$\sum(\tilde{y}_i - y_i) = 0.$$

Доказательство. Подставим a из Уравнения 1 в формулу регрессии:

$$\sum(\tilde{y}_i - y_i) = \sum(a + bx_i - y_i) = \sum(\bar{y} - b\bar{x} + bx_i - y_i) = n\bar{y} - nb\bar{x} + bn\bar{x} - n\bar{y} = 0.$$

Свойство 6. Угловой коэффициент линейной регрессии равен

$$b = r\frac{s_y}{s_x},$$

где r – коэффициент корреляции. Отсюда следует, что знаки углового коэффициента регрессии и коэффициента корреляции совпадают. Данные показатели равны в случае, когда дисперсии признаков равны.

Линейная регрессия $\tilde{y} = a + bx$

$$a = \bar{y} - b\bar{x}, \quad b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Пример. Дано: x = 1,2,3; y = 3,5,4. Обозначим: $\Delta x_i = x_i - \bar{x}$, $\Delta y_i = y_i - \bar{y}$. Произведем расчеты в таблице:

i	x_i	y_i	Δx_i	Δy_i	$\Delta x_i \Delta y_i$	$(\Delta x_i)^2$	$\tilde{y}_i$
1	1	3	-1	-1	1	1	3,5
2	2	5	0	1	0	0	4
3	3	4	1	0	0	1	4,5
Сумма	6	12	0	0	1	2	-
Среднее	2	4	-	-	1/3	2/3	-
Показатель	$\bar{x}$	$\bar{y}$	-	-	s_{xy}	s_x^2	-

Рассчитаем параметры линейной регрессии:

$$b = s_{xy} : s_x^2 = 1/3 : 2/3 = 0{,}5,$$

$$a = 4 - 0{,}5 \times 2 = 3.$$

Линейная регрессия: $\tilde{y} = 3 + 0{,}5x$. Рассчитаем ее 1-е значение: $\tilde{y}_1 = 3 + 0{,}5 \times 1 = 3{,}5$. Аналогично, ее 2-е и 3-е значения равны 4 и 4,5.

Параболическая регрессия

Параболическая регрессия – это парабола:

$$\tilde{y} = a + bx + cx^2.$$

Используется, когда заданные точки лежат вблизи кривой, имеющей выраженный локальный экстремум (максимум или минимум), либо когда данная кривая показывает замедленный рост или ускоренное падение.

Применим алгоритм МНК, для этого запишем функцию:

$$d = (y_1 - a - bx_1 - cx_1^2)^2 + \ldots + (y_n - a - bx_n - cx_n^2)^2.$$

Уравнение 1. Продифференцируем d по a, получим:

$$(-2)(y_1 - a - bx_1 - cx_1^2) + \ldots + (-2)(y_n - a - bx_n - cx_n^2) = 0, \text{ или}$$

$$\sum y_i - na - b\sum x_i - c\sum x_i^2 = 0.$$

Разделим последнее равенство на n, получим:

$$\bar{y} = a + b\bar{x} + c\overline{x^2}$$

Свойство 1 параболической регрессии следует из последнего равенства: параболическая регрессия *не проходит* через точку $(\bar{x}, \bar{y})$ с координатами, равными средним значениям признаков, т.к. для этого требуется, чтобы средний квадрат x был равен квадрату среднего значения x, а это возможно лишь в случае, когда дисперсия x равна нулю, т.е. все значения данного признака равны между собой.

Уравнения 2 и 3. Продифференцируем d последовательно по b и c, получим два новых уравнения. В итоге система уравнений для определения параметров параболической регрессии примет вид:

$$an + b\sum x_i + c\sum x_i^2 = \sum y_i$$
$$a\sum x_i + b\sum x_i^2 + c\sum x_i^3 = \sum y_i x_i$$
$$a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 = \sum y_i x_i^2$$

Матрица системы уравнений симметрична и составлена из сумм значений признака x, возведенных соответственно в степень 0,1,2,3,4. Таким образом, матрица системы не зависит от значений признака y.

Свойство 2 параболической регрессии: сумма отклонений значений параболической регрессии от соответствующих значений y равна нулю:

$$\sum(\tilde{y}_i - y_i) = 0.$$

Доказательство. Подставим a из Уравнения 1 в формулу регрессии:

$$\sum(\tilde{y}_i - y_i) = \sum(a + bx_i + cx_i^2 - y_i) = \sum((\bar{y} - b\bar{x} - c\overline{x^2}) + bx_i + cx_i^2 - y_i) =$$
$$n\bar{y} - nb\bar{x} - nc\overline{x^2} + nb\bar{x} + nc\overline{x^2} - n\bar{y} = 0.$$

Пример. Дано: x = 1,2,3; y = 3,5,4. Произведем расчеты в таблице:

i	x_i	y_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$
1	1	3	1	1	1	3	3
2	2	5	4	8	16	10	20
3	3	4	9	27	81	12	36
∑	6	12	14	36	98	25	59

Коэффициенты системы для определения параметров параболической регрессии расположены в последней строке таблицы. Система имеет вид:

$$3a + 6b + 14c = 12,$$
$$6a + 14b + 36c = 25,$$
$$14a + 36b + 98c = 59.$$

Решение системы: a = -2; b = 6,5; c = -1,5, т.е. искомая параболическая регрессия задается формулой $\tilde{y}$ =-2+6,5x-1,5x^2. Рассчитаем первое значение регрессии: $\tilde{y}_1$ = -2+6,5×1-1,5×1^2 = 3. Её 2-е и 3-е значения равны 5 и 4.

Замечание. Здесь значения параболической регрессии совпали с заданными значениями признака y, поскольку имеется всего лишь три объекта, а через три произвольные точки всегда проходит некая парабола.

Гиперболическая регрессия

Гиперболическая регрессия – это гипербола:

$$\tilde{y} = a + b/x.$$

Используется, когда заданные точки лежат вблизи гиперболы, т.е. значения y убывают и стремятся к постоянной величине при увеличении значений x. При этом признак x не должен принимать нулевых значений.

Введем новую переменную $t = 1/x$, а затем по точкам (t_i, y_i) построим линейную регрессию, используя полученные ранее формулы:

$$\tilde{y} = a + bt.$$

Согласно свойству линейной регрессии, ее график проходит через точку со средними координатами признаков t и y. Среднее значение t равно:

$\bar{t} = (1/n)\sum t_i = (1/n)\sum 1/x_i = (1/n)(n/\ddot{x}) = 1/\ddot{x}$, где

$$\ddot{x} = \frac{n}{\sum 1/x_i} \quad -$$

среднее гармоническое значение x. Итак, $\bar{y} = a + b/\ddot{x}$, т.е. гиперболическая регрессия проходит через точку $(\ddot{x}, \bar{y})$, координатами которой являются среднее гармоническое первого признака и среднее арифметическое второго признака.

Гиперболическая регрессия $\tilde{y} = a + b/x$
$t = 1/x, \quad a = \bar{y} - b\bar{t}, \quad b = \dfrac{\sum (t_i - \bar{t})(y_i - \bar{y})}{\sum (t_i - \bar{t})^2}$

Логарифмическая регрессия

Логарифмическая регрессия – это логарифмическая кривая:

$$\tilde{y} = a + b \times lnx$$

Используется, когда значения y слабо возрастают, но в то же время не стремятся к постоянной величине при увеличении значений x (например, спортивные достижения в прыжках или метании диска). При этом признак x не должен принимать отрицательных значений.

При построении регрессии можно использовать логарифм с любым основанием в зависимости от заданных значений признаков.

Пример. Заданы признаки объектов: (10,2), (101,4), (1003, 6). Заметим, что десятичный логарифм значения x составляет приблизительно половину соответствующего значения y, т.е. между ними существует линейная зависимость. Поэтому в данном случае выбираем десятичный логарифм, а формула логарифмической регрессии имеет вид $\tilde{y} = a + b \times lqx$.

Введем новую переменную $t = lnx$, а затем по точкам (t_i, y_i) построим линейную регрессию, используя полученные ранее формулы:

$$\tilde{y} = a + bt.$$

Согласно свойству линейной регрессии, ее график проходит через точку со средними координатами признаков t и y. Среднее значение t равно

$$\bar{t} = (1/n)\sum t_i = (1/n)\sum \ln x_i = (1/n)\ln \prod x_i = \ln \hat{x}, \text{ где}$$

$$\hat{x} = \sqrt[n]{\prod x_i} \quad -$$

среднее геометрическое значение *x*. Итак, $\bar{y} = a + b \times \ln \hat{x}$, т.е. логарифмическая регрессия проходит через точку $(\hat{x}, \bar{y})$, координатами которой являются среднее геометрическое первого признака и среднее арифметическое второго признака.

Упростим расчетные формулы, используя свойства логарифма:

$$t_i - \bar{t} = \ln x_i - \ln \hat{x} = \ln x_i / \hat{x}, \text{ тогда}$$

$$b = \sum \ln x_i / \hat{x} (y_i - \bar{y}) / \sum \ln^2 x_i / \hat{x}.$$

Пример. Дано: x = 15,30,70; y = 7,8,9. Построим логарифмическую регрессию. Найдем среднее геометрическое *x*: $(15 \times 30 \times 70)^{1/3}$ = 31,58. Произведем расчеты в таблице:

i	x_i	y_i	$\ln x_i / \hat{x}$	Δy_i	$\ln x_i / \hat{x} \, \Delta y_i$	$\ln^2 x_i / \hat{x}$
1	15	7	-0,794	-1	0,744	0,553
2	30	8	-0,051	0	0	0,003
3	70	9	2,216	1	2,216	4,911
∑	-	24	-	0	2,96	5,467

Рассчитаем *b*: разделим сумму элементов предпоследнего столбца на сумму элементов последнего столбца: *b* = 2,96 : 5,457 = 0,542.

Рассчитаем $a = \bar{y} - b \ln \hat{x} = 8 - 0{,}542 \times 3{,}452 = 6{,}129$.

Рассчитаем значения регрессии: $\tilde{y}(15) = 6{,}129 + 0{,}542 \ln 15 = 7{,}597$. Второе и третье значения регрессии равны 7,972 и 8,432 соответственно.

Логарифмическая регрессия $\tilde{y} = a + b \times \ln x$

$$a = \bar{y} - b \times \ln \hat{x}, \quad b = \frac{\sum \ln(x_i / \hat{x})(y_i - \bar{y})}{\sum \ln^2 (x_i / \hat{x})}$$

Задачи

1. Отклонения значений признака *y* от соответствующих значений регрессии составили: в классе функций «А»: 5, -2, - 2; в классе функций «В»: 2, -4, 3. Определите, какой вид регрессии лучше.

2. Отклонения значений признака *y* от соответствующих значений параболической регрессии составили: 3, -1, *t*. Найти *t*.

3. Дано: *x* = 2,2,5; *y* = 4,6,8. Не вычисляя параметры линейной регрессии, определите точку, через которую проходит ее график.

4. Определите без вычислений наилучший вид регрессии (линейная, параболическая, гиперболическая), если заданы признаки трех объектов:

а) (1,1), (2,2), (3,1);

б) (4,6), (7,9), (9,12);

в) (4,16), (16,9), (64,6).

5. Определите формулу линейной регрессии:

а) x = 3,9,7,5; y = 10,8,16,14;

б) x = 3,8,11,2; y = 6,17,8,5;

в) x = 11,15,16,10; y = 1,3,7,5;

6. Исследуются значения двух признаков для пяти объектов, причем x=0,1,2,3,4. Определите элементы последней строки матрицы системы для расчета параметров параболической регрессии.

7. Найти формулу параболической регрессии, если:

а) x = 1,2,3; y = 2,1,2;

б) x = 0,1,2; y = 1,0,2;

8. Задано распределение численности занятых по возрастным группам в России в 1995 г. Найти формулу параболической регрессии, приняв в качестве значений факторного признака средний возраст в группе:

Группа	Возраст, лет	Удельный вес, %
1	16-19	3
2	20-24	10,3
3	25-29	11,3
4	30-34	15,1
5	35-39	16,5
6	40-44	15,3
7	45-49	11,4
8	50-54	7,1
9	55-59	6,9

Источник: РСЕ. 2009. С. 139.

9. Кривые Энгеля строят как график зависимости удельного веса расходов на товар (y) от номера группы домохозяйств (с ростом номера доходы возрастают). Номера групп: x = 0,1,2,3,4. Определите параметр при x^2 параболической регрессии для предметов роскоши (данные Росстата):

а) покупка транспортных средств (2010 г.): y = 0; 0,1; 0,3: 1,6; 16,3%;

б) гостиницы, кафе, рестораны (2010 г.): y = 1; 1,4; 1,9: 3,3; 4,9%;

в) предметы домашнего обихода (2006 г.): y = 3; 3,9; 5,2: 8,4; 9%;

10. Гиперболическая регрессия построена по точкам (2,23), (12,4), (24,2). Не вычисляя параметры регрессии, определите точку, через которую проходит ее график (с точностью до 0,01).

11. Логарифмическая регрессия построена по точкам (3,1), (9,2), (81,5). Не вычисляя параметры регрессии, определите точку, через которую проходит ее график (с точностью до 0,01).

12. Дано: (10,4), (20,5), (80,6). Построить логарифмическую регрессию.

13. Дайте ответ (верно/неверно):

1) МНК – метод наименьших коэффициентов

2) Для расчета параметров регрессии минимизируемую функцию дисконтируют

3) Угол наклона линейной регрессии равен 45^o, если ковариация равна дисперсии x

4) МНК: минимизируется сумма разностей значений признаков

5) Квадрат средней не меньше среднего квадрата признака

6) Сумма модулей отклонений y от среднего значения y равна 0

7) A = 1,2,3, B = 0,1,0. Тогда параболическая регрессия лучше линейной

8) A = 7,2,1,4 B = 3,8,9,6. Тогда параболическая регрессия лучше линейной

9) Свободный член линейной регрессии равен среднему значению *y*

10) Ковариация равна среднему произведению *x* и *y* минус произведение средних *x* и *y*

11) Дисперсия *x* равна квадрату среднего *x* минус средний квадрат *x*

12) Среднее произведение не меньше произведения средних значений

13) Гиперболическую регрессию не используют при наличии экстремумов

14) Параболическая регрессия строится с помощью формул линейной регрессии

15) Параболическая регрессия: матрица системы симметрична

16) Построение гиперболической регрессии: «x» заменяют на противоположную величину

17) Параболическая регрессия задается двумя параметрами

18) Параболическая регрессия: для расчета матрицы системы не требуются значения y

19) Параболическая регрессия проходит через точку (x среднее, y среднее)

20) Сумма значений параболической регрессии равна сумме исходных значений *y*

6. ВЗАИМОСВЯЗЬ НЕСКОЛЬКИХ ПРИЗНАКОВ

Множественное уравнение регрессии

Имеется n объектов, каждый из которых характеризуется значением результативного признака y и значениями факторных признаков x_1, x_2, …, x_m. Предполагается, что факторные признаки влияют на значение результативного признака. Поскольку всего имеется m+1 признаков, объекты можно рассматривать как точки (m+1)-мерного пространства.

Множественная регрессия – это линейная функция m переменных:

$$\tilde{y} = a + \sum_{i=1}^{m} b_i x_i,$$

где a, b_1, …, b_m – параметры регрессии, определяемые методом наименьших квадратов с помощью функции d, которая имеет m+1 переменных:

$$d(a, b_1, \ldots b_n) = (a + \sum b_i x_i^{\,1} - y_1)^2 + \ldots + (a + \sum b_i x_i^{\,m} - y_m)^2,$$

где x_i^k – значение k-го факторного признака, y_i – значение результативного признака для i-го объекта.

Определим значения параметров, обеспечивающие минимум функции d, для этого приравняем нулю ее частные производные и решим систему уравнений. Первое уравнение:

$$\partial d/\partial a = 0, \quad \text{отсюда } a = \bar{y} - \sum b_i \bar{x}_i,$$

где $\bar{x}_i$ – среднее значение i–го факторного признака, $\bar{y}$ – среднее значение результативного признака, рассчитанное по всем объектам.

Множественная регрессия $\tilde{y} = a + b_1 x + b_2 z$
$a = \bar{y} - b_1\bar{x} - b_2\bar{z},$
$b_1 \sum (\Delta x_i)^2 + b_2 \sum \Delta x_i \Delta z_i = \sum \Delta y_i \Delta x_i$
$b_1 \sum \Delta x_i \Delta z_i + b_2 \sum (\Delta z_i)^2 = \sum \Delta y_i \Delta z_i$

Матрица системы m линейных уравнений для определения параметров b_i симметрична, на ее главной диагонали расположены дисперсии факторных признаков s_i^2, а недиагональные элементы равны значениям ковариации факторных признаков s_{ij}. Вектор правых частей уравнений составлен из значений ковариации результативного признака с факторными признаками:

$$s_{iy} = \sum_{k=1}^{n} (x_i^k - \bar{x}_i)(y_k - \bar{y}),$$

где $\bar{x}_i$ – среднее значение i-го факторного признака.

Для случая двух факторных признаков система имеет вид:

$$\begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} s_{1y} \\ s_{2y} \end{pmatrix},$$

где $s_{12} = s_{21}$ – ковариация факторных признаков. Приравняем нулю определитель матрицы и получим условие, когда система не имеет решений:

$$\left(\frac{s_{12}}{s_1 s_2} \right)^2 = 1.$$

Отсюда множественная линейная регрессия *не существует*, если модуль коэффициента корреляции между факторными признаками равен 1. Если же коэффициенты корреляции между всеми парами факторных признаков равны 0, тогда матрица диагональная и параметры определяют по формулам, аналогичным формуле коэффициента парной регрессии:

$$b_i = s_{iy}/s_i^2.$$

Пример. Имеется три объекта. Значения результативного признака y и факторных признаков x и z заданы в таблице.

i	y_i	x_i	z_i	Δy_i	Δx_i	Δz_i	$(\Delta x_i)^2$	$(\Delta z_i)^2$	$\Delta x_i \Delta z_i$	$\Delta y_i \Delta x_i$	$\Delta y_i \Delta z_i$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1	3	1	4	-3	-1	-4	1	16	4	3	12

2	5	2	7	-1	0	-1	0	1	0	0	1
3	10	3	13	4	1	5	1	25	5	4	20
∑	18	6	24	0	0	0	2	42	9	7	33

Определим параметры линейной регрессии: $\tilde{y} = a + b_1 x + b_2 z$.

1. Рассчитаем средние значения признаков: результативного: 18:3=6, факторных: 2 и 8 (столбцы 2-4)

2. Рассчитаем отклонения значений каждого признака от среднего значения: $\Delta y_i = y_i - \bar{y}$ и т.д. (столбцы 5-7).

3. Рассчитаем квадраты отклонений для факторных признаков, найдем их сумму (столбцы 8-9).

4. Рассчитаем произведения отклонений для каждой пары признаков, найдем их сумму (столбцы 10-12).

5. Составим систему уравнений:

$$2b_1 + 9b_2 = 7,$$
$$9b_1 + 42b_2 = 33.$$

6. Решим систему: (-1, 1).

7. Рассчитаем свободный член регрессии: $a = 6 - (-1)\times 6 - 1\times 8 = 0$.

8. Составим уравнение регрессии: $\tilde{y} = -x + z$. Оно задает плоскость в трехмерном пространстве, проходящую через начало координат.

9. Рассчитаем значения регрессии: для 1-го объекта: -1+4=3, для 2-го: -2+7=5, для 3-го: -3+13=10. Как мы видим, значения регрессии совпали с заданными значениями результативного признака. Этого следовало ожидать, т.к. через три произвольные точки всегда можно провести плоскость, и за редким исключением она единственна.

Коэффициент множественной детерминации

Коэффициент множественной детерминации (R^2) – общий показатель тесноты связи факторных признаков с результативным признаком, он равен отношению определителей двух матриц:

$$R^2 = -\Delta^*/\Delta.$$

Матрица с определителем Δ – это *матрица парных коэффициентов корреляции*. Она симметрична: ее диагональ составлена из единиц, а недиагональные элементы равны соответствующим значениям парной корреляции между факторными признаками, ее размерность $m \times m$, где m – число факторных признаков.

Матрица с определителем Δ^* получается расширением предыдущей матрицы дополнительными строкой и столбцом, которые составлены из коэффициентов парной корреляции между результативным признаком и факторными признаками. Ее размерность $(m+1)\times(m+1)$.

Свойство 1. Коэффициент множественной детерминации лежит в пределах от нуля до единицы.

Свойство 2. Коэффициент множественной детерминации равен нулю, если все коэффициенты корреляции между факторными признаками и результативным признаком равны нулю.

В случае двух факторных признаков матрицы имеют вид:

$$\Delta = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}, \quad \Delta^* = \begin{vmatrix} r_{1y} & r_{2y} & 0 \\ 1 & r_{12} & r_{1y} \\ r_{21} & 1 & r_{2y} \end{vmatrix},$$

где r_{12} – коэффициент корреляции между факторными признаками, r_{1y} – коэффициент корреляции между первым факторным признаком и результативным признаком. В данном случае коэффициент множественной детерминации может быть рассчитан по формуле

$$R^2 = \frac{r_{1y}^{\ 2} + r_{2y}^{\ 2} - 2r_{1y}r_{2y}r_{12}}{1 - r_{12}^{\ 2}}.$$

Из формулы следует, что коэффициент множественной детерминации *не существует*, если коэффициент корреляции между факторными признаками равен единице.

Свойство 3. Коэффициент множественной детерминации не определен, если факторные признаки функционально связаны друг с другом.

Если коэффициент корреляции между факторными признаками равен нулю, то формула упрощается:

$$R^2 = r_{1y}^{\ 2} + r_{2y}^{\ 2}.$$

Из нее следует, что чем больше корреляция факторного и результативного признаков, тем больший вклад вносит данный факторный признак в общий показатель тесноты связи факторных признаков и результативного признака.

Пример. Исследуется связь между уровнем образования человека и его родителей. Имеется три булевых признака: результативный – «наличие высшего образования у человека», факторные – «наличие высшего образования у отца» и «наличие высшего образования у матери». Результаты обследования пяти человек приведены в таблице.

Признак	A	B	C	D	E
Высшее образование человека (y)	0	1	0	0	1
Высшее образование отца (x_1)	0	1	1	0	1
Высшее образование матери (x_2)	0	1	0	1	0

1. Коэффициенты корреляции равны:
- между уровнем образования отца и матери: (2-3)/(2+3) = -0,2,
- между уровнем образования человека и его отца: (4-1)/(4+1) = 0,6,
- между уровнем образования человека и его матери: (3-2)/(3+2) = 0,2.

2. Коэффициент множественной детерминации равен:

$$R^2 = (0{,}6^2 + 0{,}2^2 - 2\times 0{,}6\times 0{,}2\times(-0{,}2))/(1 - (-0{,}2)^2) = 0{,}47.$$

Применение: функция Кобба-Дугласа

Производственная функция Кобба-Дугласа имеет вид $Q = AL^{\alpha}K^{\beta}$, где Q – выпуск при затратах труда L и капитала K, а числа A, α и β положительны.

Пусть получено n наблюдений: i-е наблюдение характеризуется значением результативного признака Q_i и значениями факторных признаков L_i и K_i. На основании этих данных определим параметры функции Кобба-Дугласа, для этого прологарифмируем ее:

$$\ln Q = \ln A + \alpha \ln L + \beta \ln K.$$

Мы видим, что логарифм результативного признака линейно зависит от логарифмов факторных признаков, что позволяет использовать множественную регрессию. Обозначим: $y=\ln Q$, $x=\ln L$, $z=\ln K$. Прологарифмируем заданные значения признаков и по этим данным построим множественную регрессию:

$$\tilde{y} = a + b_1 x + b_2 x.$$

Определим параметры искомой функции: $\alpha = b_1$, $\beta = b_2$, $A = e^a$.

Получаем *эмпирическую* функцию Кобба-Дугласа $Q = e^a L^{b1} K^{b2}$, которая позволяет получить достоверный прогноз выпуска при любых ожидаемых значениях затрат труда и капитала.

Применение: функция Минцера

Функция заработной платы Минцера имеет вид:

$$w = w_0 p^E q^S,$$

где w – зарплата квалифицированного работника с продолжительностью образования E и продолжительностью трудового стажа S, w_0 – зарплата неквалифицированного работника без образования и стажа. Параметр p показывает, во сколько раз увеличится зарплата при увеличении образования на 1 год, а параметр q – при увеличении трудового стажа на 1 год.

Пусть i-е наблюдение характеризуется значением результативного признака w_i и значениями факторных признаков E_i и S_i. По этим данным определим параметры функции Минцера, для этого прологарифмируем ее:

$$\ln w = \ln w_0 + E \ln p + S \ln q.$$

Мы видим, что логарифм результативного признака линейно зависит от значений факторных признаков, что позволяет нам использовать множественную регрессию. Обозначим: $y=\ln w$, $x=E$, $z=S$. Прологарифмируем заданные значения результативного признака и по этим данным и значениям факторных признаков построим множественную регрессию:

$$\tilde{y} = a + b_1 x + b_2 x.$$

Определим параметры искомой функции:

$$a = \ln w_0, \text{ отсюда } w_0 = e^a,$$
$$b_1 = \ln p, \text{ отсюда } p = e^{b1},$$
$$b_2 = \ln q, \text{ отсюда } q = e^{b2}.$$

Получаем *эмпирическую* функцию Минцера $w = w_0 e^{a+b_1 E+b_2 S}$, которая позволяет получить достоверную оценку заработной платы при любых значениях продолжительности образования и трудового стажа.

Замечание. Результаты статистических исследований показывают, что параметр p равен приблизительно 1,08, т.е. каждый дополнительный год обучения увеличивает заработную плату в среднем на 8%.

Задачи

1. Найти коэффициент множественной детерминации:

Признак	A	B	C	D	E	F
Высшее образование человека (y)	1	1	1	0	1	1
Высшее образование отца (x_1)	1	1	1	1	1	1
Высшее образование матери (x_2)	0	1	1	1	0	1

2. Антон обучался 8 лет и получает 13906 руб., Борис обучался 10 лет и получает 15749 руб., Иван планирует учиться 14 лет. Найти:

а) зарплату неквалифицированного работника;

б) на сколько процентов увеличится зарплата после двух лет обучения;

в) ожидаемую зарплату Ивана.

3. Дайте ответ (верно/неверно):

1) Множественная регрессия: рассматривается один результативный признак

2) Множественная линейная регрессия: число параметров равно числу признаков

3) Факторных признаков – 2, тогда параметров линейной регрессии – 2

4) Параметры множественной линейной регрессии определяют с помощью МНК

5) Построение множественной линейной регрессии: диагональ матрицы составлена из дисперсий

6) Построение множественной линейной регрессии: матрица системы асимметрична

7) Факторных признаков 2, корреляция между ними равна -1, тогда множественная линейная регрессия существует

8) Построение множественной линейной регрессии: матрица системы зависит от y.

9) Построение множественной линейной регрессии: правые части системы зависят от y

10) Построение функция Кобба-Дугласа: всего имеется 2 признака

11) Функция Кобба-Дугласа: логарифм выпуска – линейная функция затрат труда и капитала

12) Построение функции Кобба-Дугласа: регрессия задается тремя параметрами

13) Для определения параметров функции Кобба-Дугласа ее дифференцируют

14) Построение функции Кобба-Дугласа: регрессия – плоскость в трехмерном пространстве

15) Для расчета коэффициента множественной детерминации используют две матрицы равной размерности

16) Коэффициент множественной детерминации показывает силу связи между факторными признаками

17) Корреляция между всеми парами факторных признаков равна нулю, тогда коэффициент множественной детерминации равен нулю

18) При расчете коэффициента множественной детерминации с двумя факторными признаками нужно знать три коэффициента корреляции

7. ДИНАМИКА: ОСНОВНЫЕ ПОНЯТИЯ

Уровень – это признак, зависящий от времени, совокупность его значений называют *рядом динамики (динамическим рядом).* Уровень является результативным признаком, а время – факторным.

Если время задано в виде промежутков, ряд называют *интервальным*, а если в виде моментов времени – *моментным*.

Динамика характеризуется *тенденцией (трендом)* и *колеблемостью* – отклонением от тенденции.

Основные показатели динамики

1. *Абсолютный прирост* – разность между сравниваемым уровнем и уровнем более раннего периода. Может быть цепным или базисным.

Цепной абсолютный прирост – разность сравниваемого уровня и предыдущего уровня:

$$a_i = y_i - y_{i-1}.$$

Базисный абсолютный прирост – разность сравниваемого уровня и некого фиксированного, базисного уровня:

$$a_i' = y_i - y_0.$$

Замечание. Абсолютные приросты можно рассчитывать, начиная со 2-го уровня, поэтому их общее число равно n-1, где n – число уровней.

Свойство. Для каждого момента времени базисный абсолютный прирост равен сумме предыдущих цепных абсолютных приростов для моментов времени, следующих за базисным годом:

$$a_i' = \sum_{j=i_0+1}^{i} a_j.$$

Число слагаемых в правой части равно числу лет, прошедших между базисным и исследуемым годом.

2. *Темп роста* – отношение сравниваемого уровня и уровня более раннего периода. Может быть цепным или базисным.

Цепной темп роста – отношение сравниваемого уровня и предыдущего уровня:

$$k_i = \frac{y_i}{y_{i-1}}.$$

Базисный темп роста – отношение сравниваемого уровня и базисного уровня:

$$k_i' = \frac{y_i}{y_0}.$$

Свойство. Для каждого момента времени базисный темп роста равен произведению предыдущих цепных темпов роста для моментов времени, следующих за базисным годом:

$$k_i' = \prod_{j=i_0+1}^{i} k_j.$$

Число множителей в правой части равно числу лет, прошедших между базисным и исследуемым годом.

3. *Темп прироста* – отношение абсолютного прироста и уровня более раннего периода. Может быть цепным или базисным.

Цепной темп прироста – отношение цепного абсолютного прироста и предыдущего уровня, равное цепному темпу роста минус 1:

$$l_i = \frac{y_i - y_{i-1}}{y_{i-1}} = \frac{y_i}{y_{i-1}} - 1 = k_i - 1.$$

Базисный темп прироста – отношение базисного абсолютного прироста и базисного уровня, равное базисному темпу роста минус 1:

$$l_i' = \frac{y_i - y_0}{y_0} = \frac{y_i}{y_0} - 1 = k_i' - 1.$$

Темпы прироста обычно выражают в процентах.

Пример. Объемы выручки заданы в таблице, базисный год – 2002 г.

Показатель		Ед.измерения	2001	2002	2003	2004	2005
Время	i	Год	1	2	3	4	5
Выручка	y_i	Млн.руб.	10	20	25	21	28
Абсол. прирост (ц)	a_i	Млн.руб.	-	10	5	-4	7
Абсол. прирост (б)	a_i'	Млн.руб.	-	-	5	1	8
Темп роста (ц)	k_i	Разы	-	2	1,25	0,84	1,33
Темп роста (б)	k_i'	Разы	-	-	1,25	1,05	1,4
Темп прироста (ц)	l_i	%	-	100	25	-16	33
Темп прироста (б)	l_i'	%	-	-	25	5	40

Замечание. Если базисный год совпадает с предыдущим, то цепные показатели равны базисным.

Динамика относительных уровней

Рассмотрим три динамических ряда: $A_i = B_i + C_i$.

Альтернативными долями называют показатели:

$$x_i = \frac{B_i}{A_i}, \quad y_i = \frac{C_i}{A_i}.$$

Альтернативные доли – это удельные веса слагаемых, их сумма равна единице. Исходные уровни являются объемными показателями, а альтернативные доли – относительными показателями.

Свойство 1. Абсолютные приросты альтернативных долей противоположны:

$$a_{xi} = - a_{yi},$$

где a_{xi} – цепной абсолютный прирост удельного веса первого слагаемого (B), a_{yi} – цепной абсолютный прирост удельного веса второго слагаемого (C) в i-й момент времени. Аналогичное равенство верно для базисных приростов.

Пример. Удельный вес мужчин вырос на 5%, тогда удельный вес женщин снизился на 5%.

Замечание. При увеличении доли мужчин с 20% до 25% она возросла на четверть своей начальной величины, т.е. на 25%. Но с другой стороны, ее увеличение составило 5%. Чтобы избежать путаницы, абсолютные приросты относительных уровней выражают в *процентных пунктах* (п.п.). Итак, доля мужчин выросла на 5 процентных пунктов.

Свойство 2. Темпы роста альтернативных долей связаны равенством:

$$k_{yi} = \frac{1 - x_{i-1} k_{xi}}{1 - x_{i-1}},$$

где k_{xi} – темп роста первой доли, k_{yi} – темп роста второй доли в i-й момент времени, x_{i-1} – величина первой доли в предыдущий момент времени. Аналогичное равенство верно для базисных приростов.

Пример. Доля мужчин в начале года была 40%, а к концу года она выросла в 1,3 раза. Тогда темп роста доли женщин составит:

$$k_y = \frac{1 - 0{,}4 \times 1{,}3}{1 - 0{,}4} = 0{,}8.$$

Данный результат можно получить иначе. Поскольку доля мужчин возросла до 40×1,3=52%, доля женщин снизилась с 60% до 48%, т.е. изменилась в 48/60=0,8 раз.

Свойство 3. Темп роста альтернативной доли и темпы роста объемных показателей связаны равенством:

$$k_x = \frac{k_B}{k_A},$$

где k_B и k_A – темпы роста объемных показателей A и B. Аналогичное равенство справедливо для базисных темпов роста первой доли и для цепных и базисных темпов роста второй доли.

Пример. Число мужчин выросло на 32%, а численность населения – на 10%. Тогда доля мужчин выросла в 1,32/1,1 = 1,2 раза. Если она составляла 50%, то увеличится до 50×1,2=60%.

Свойство 4. Темп прироста альтернативной доли и темпы прироста объемных показателей связаны приближенным равенством:

$$l_x \approx l_B - l_A,$$

где l_B и l_A – темпы прироста объемных показателей A и B. Аналогичное равенство справедливо для базисных темпов прироста первой доли и для цепных и базисных темпов прироста второй доли.

Пример. Число мужчин выросло на 32%, а численность населения – на 10%. Тогда доля мужчин выросла приблизительно на 32-10=22%. Если она составляла 50%, то увеличится приблизительно до 50×1,22=61%.

Замечание. Свойства 3 и 4 можно использовать для любого показателя, который являются отношением двух других показателей.

Пример. Номинальный ВВП вырос на 4%, а уровень инфляции составил 9%. Определим процентное изменение реального ВВП. Поскольку уровень инфляции есть темп прироста дефлятора ВВП, а реальный ВВП есть отношение номинального ВВП и дефлятора ВВП, темп роста реального ВВП равен 1,04/1,09 = 0,954 (свойство 3), т.е. он сократился на 1-0,954=0,046-ю часть, или 4,6%. Приближенный темп прироста (свойство 4): 4-9=-5%.

Средние показатели тенденции динамики

Средний уровень моментного ряда с равными промежутками времени:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

где n – количество промежутков времени.

Средний уровень интервального ряда с различными промежутками времени:

$$\bar{y} = \frac{\sum y_i T_i}{\sum T_i},$$

где T_i – продолжительность промежутка времени, в течение которого уровень неизменно составлял y_i. Если все промежутки одинаковы, то мы получим предыдущую формулу.

Хронологическая средняя – среднее значение уровней, при расчете которого внутренние уровни учитываются с коэффициентом 1, а два крайних уровня – с коэффициентами 0,5:

$$\bar{y}_x = \frac{0{,}5y_1 + \sum_{i=2}^{n-1} y_i + 0{,}5y_n}{n-1}.$$

Средний цепной абсолютный прирост равен:

$$\bar{a} = \frac{\sum_{i=2}^{n} a_i}{n-1} = \frac{y_n - y_1}{n-1}.$$

Средний цепной абсолютный прирост равен результирующему приросту уровня (разности последнего и первого уровня), деленному на количество исследуемых промежутков времени. Он показывает среднее абсолютное изменение уровня за один промежуток времени.

Средний цепной темп роста равен:

$$\bar{k} = \sqrt[n-1]{\prod_{i=2}^{n} k_i} = \sqrt[n-1]{\frac{y_n}{y_1}}.$$

Средний цепной темп роста равен корню (n-1)-й степени из результирующего темпа роста (отношения последнего и первого уровней). Он показывает средний темп роста уровня за один промежуток времени.

Средний цепной темп прироста равен: $\bar{l} = \bar{k} - 1$. Он показывает средний темп прироста уровня за один промежуток времени.

Пример. Уровни: 2,4,5,10. Тогда средний уровень равен 5,25, хронологическая средняя – (1+4+5+5)/3=5, средний цепной абсолютный прирост – (10-2)/3 = 2,67, средний цепной темп роста – $\sqrt[3]{10/2} = 1{,}71$, средний цепной темп прироста – 1,71-1=0,71, или 71%.

Задачи

1. Дайте ответ (верно/неверно):

1) Уровни и абсолютные приросты измеряются в одинаковых единицах

2) Базисный абсолютный прирост больше цепного прироста

3) Сумма базисных приростов равна цепному приросту

4) Цепной прирост не определен, если предыдущий уровень равен 0

5) Уровень вырос с 10 до 12, тогда темп роста равен 20%

6) Уровень снизился с 8 до 6, тогда темп роста равен 0,75

7) Уровень изменился с 9 до 8, тогда темп прироста отрицателен

8) Базисный темп роста в 1-м и 2-м году равен 1,2, тогда уровень за два года вырос на 44%

9) Цепные темпы роста в 1-м и 2-м году равны 1,4 и 1,5, тогда уровень за два года вырос в 2,1 раза

10) Инфляция в 1-м и 2-м году составила 15%, тогда за два года она превысит 31%

11) Приросты альтернативных долей – обратные величины

12) Сумма темпов роста альтернативных долей равна единице

13) Сумма приростов альтернативных долей равна нулю.

14) Темп роста альтернативной доли зависит от ее значения в предыдущий момент

15) Темпы роста альтернативных долей выражают в процентных пунктах

16) Доля мужчин выросла с 50% до 60%, тогда она выросла на 10 п.п.

17) Доля мужчин выросла вдвое, тогда она выросла на 50 п.п.

18) Доля мужчин выросла с вдвое, тогда доля женщин сократилась вдвое

19) Темп роста доли мужчин равен отношению темпов роста числа мужчин и населения

20) Темп прироста доли мужчин равен разности темпов прироста числа мужчин и населения

21) Темп прироста мужчин равен 80%, населения – 50%, тогда доля мужчин выросла на 20 п.п.

22) Средний прирост динамического ряда 4,*x*,5 не зависит от *x*

23) Средний прирост динамического ряда 4,5,*x* не зависит от *x*

24) Средний темп прироста динамический ряда 2,5,2,16 больше 90%

8. ДИНАМИКА: ТРЕНД

Общие положения

Тренд – уравнение регрессии для динамического ряда.

Рассмотрим динамический ряд с равными промежутками между моментами времени, причем их число *n* нечетно. Среднему по порядку моменту (году) присвоим номер 0, последующему – 1, предыдущему – -1 и т.д. Тогда факторный признак «время» принимает следующие значения:

$$t = -\frac{n-1}{2}, \dots, -2, -1, 0, 1, 2, \dots, \frac{n-1}{2}.$$

В силу симметрии значений сумма нечетных степеней переменной *t* равна нулю, отсюда: $\bar{t} = 0, \quad \overline{t^3} = 0.$

Выведем формулы линейного, параболического, экспоненциального, логистического и степенного трендов, используя формулы регрессии.

Линейный тренд

Линейный тренд – это прямая линия:

$$\tilde{y} = a + bt.$$

Определим свободный член линейного тренда. Для этого запишем формулу свободного члена линейной регрессии и учтем, что среднее значение *t* равно нулю:

$$a = \bar{y} - b\bar{t} = \bar{y}.$$

Итак, свободный член линейного тренда равен среднему уровню.

$\underline{\textit{Линейный тренд}} \quad \underline{\tilde{y} = a + bt}$
$a = \bar{y}, \quad b = \dfrac{\sum y_i t_i}{\sum t_i^2}$

Определим угловой коэффициент линейного тренда. Запишем формулу для линейной регрессии и учтем, что среднее значение *t* равно нулю:

$$b = \frac{\sum (t_i - \bar{t})(y_i - \bar{y})}{\sum (t_i - \bar{t})^2} = \frac{\sum t_i (y_i - \bar{y})}{\sum t_i^2} = \frac{\sum t_i y_i - \bar{y}\sum t_i}{\sum t_i^2} = \frac{\sum y_i t_i}{\sum t_i^2}.$$

Знаменатель дроби равен:

- при 3-х моментах времени: $(-1)^2 + 0^2 + 1^2 = 2$;
- при 5-ти моментах времени: $(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10$ и т.д.

Свойство 1. Линейный тренд проходит через точку $(0, \bar{y})$.

Свойство 2. Сумма отклонений значений линейного тренда от соответствующих уровней равна нулю:

$$\sum(\tilde{y}_i - y_i) = 0.$$

Отсюда следует, что сумма значений тренда равна сумме уровней.

Пример. Уровни заданы во второй строке таблицы:

Время	t	-2	-1	0	1	2	$\sum$
Уровень	y_i	8	7	9	10	9	43
Тренд	$\tilde{y}_i$	7,6	8,1	8,6	9,1	9,6	43

Рассчитаем параметры линейного тренда:

$$a = (8 + 7 + 9 + 10 + 9)/5 = 8{,}6,$$

$$b = (8\times(-2) + 7\times(-1) + 9\times0 + 10\times1 + 9\times2)/10 = 0{,}5.$$

Линейный тренд: $\tilde{y} = 8{,}6 + 0{,}5t$. Рассчитаем его значения: $\tilde{y}(-2) = 8{,}6 + 0{,}5\times(-2) = 7{,}6$. Значения тренда показаны в 3-й строке таблицы. Проверка: сумма уровней равна сумме значений тренда (последний столбец).

Параболический тренд

Параболический тренд – это парабола:

$$\tilde{y} = a + bt + ct^2.$$

Используется, когда точки лежат вблизи кривой, имеющей локальный экстремум, или показывающей замедленный рост или ускоренное падение.

В системе уравнений для определения параметров параболической регрессии заменим x на t:

$$an + b\sum t_i + c\sum t_i^2 = \sum y_i$$
$$a\sum t_i + b\sum t_i^2 + c\sum t_i^3 = \sum y_i t_i$$
$$a\sum t_i^2 + b\sum t_i^3 + c\sum t_i^4 = \sum y_i t_i^2$$

Учитывая, что $\sum t_i = \sum t_i^3 = 0$, упростим систему:

$$an + 0 + c\sum t_i^2 = \sum y_i$$
$$0 + b\sum t_i^2 + 0 = \sum y_i t_i$$
$$a\sum t_i^2 + 0 + c\sum t_i^4 = \sum y_i t_i^2$$

Решим систему. Из второго уравнения определим параметр b:

$$b = \frac{\sum y_i t_i}{\sum t_i^2}.$$

Из первого и третьего уравнений системы определим параметр c:

$$c = \frac{\overline{yt^2} - \bar{y}\times\overline{t^2}}{\overline{t^4} - (\overline{t^2})^2}, \quad \text{где } \overline{yt^2} = \frac{\sum y_i t_i^2}{n} \text{ и т.д.}$$

Из первого уравнения системы получаем: $a = \bar{y} - c\overline{t^2}$.

Свойство. Сумма отклонений значений параболического тренда от соответствующих уровней равна нулю.

Частный случай - пять уровней. Среднее значение t^2 равно 10/5=2, а среднее значение t^4 равно (16+1+0+1+16)/5=6,8. Тогда $b = 0{,}1\sum y_i t_i$. В

формуле расчета c знаменатель равен $6{,}8-2^2=2{,}8$, отсюда: $c = 0{,}357(\overline{yt^2} - 2\bar{y})$. Свободный член тренда равен $a = \bar{y} - 2c$.

Пример. Уровни: 9,16,21,24,25.

i	y_i	t_i	t_i^2	$y_i t_i$	$y_i t_i^2$
1	9	-2	4	-18	36
2	16	-1	1	-16	16
3	21	0	0	0	0
4	24	1	1	24	24
5	25	2	4	50	100
∑	95	0	10	40	176

Используем формулы для частного случая и данные таблицы: b = 40/10=4, c = 0,357(176/5-2×19) = -1, a = 95/5- 2×(-1) = 21. Тогда формула параболического тренда: $\tilde{y} = 21 + 4t - t^2$. Его значение для начального момента времени равно $21+4(-2)-(-2)^2$ = 9, т.е. совпадает с заданным значением уровня.

Экспоненциальный тренд

Экспоненциальный тренд – это график функции $\tilde{y} = mk^t$, где числа m и k положительны. Используется в случае ускоренного неограниченного возрастания или ускоренного ограниченного снизу убывания уровней.

1. Прологарифмируем формулу тренда: $\ln \tilde{y} = \ln m + t \ln k$.

2. Введем переменную $z = \ln \tilde{y}$, которая линейно зависит от времени.

3. Построим линейный тренд для динамического ряда z_i: $\tilde{z} = a + bt$.

4. Из сравнений формул $\ln \tilde{y}$ и $\tilde{z}$ следует, что a = ln$m$, а из свойства линейного тренда: $a = \bar{z}$. Отсюда

$$m = \exp(\bar{z}) = \exp(\ln \hat{y}),$$

где $\hat{y}$ - среднее геометрическое значение заданных уровней y_i.

5. Из сравнений формул $\ln \tilde{y}$ и $\tilde{z}$ также следует, что b = ln$k$. Используя формулу углового коэффициента линейной регрессии, получим:

$$k = \exp b = \exp \frac{\sum z_i t_i}{\sum t_i^2} = \exp \frac{\sum t_i \ln y_i}{\sum t_i^2}.$$

Итак, формула экспоненциального тренда:

$$\tilde{y} = \hat{y} \exp(t \frac{\sum t_i \ln y_i}{\sum t_i^2}).$$

Пример. Динамический ряд: 1,2,4.

i	1	2	3	∑
t_i	-1	0	1	-

y_i	1	2	4	-
$\ln y_i$	0	ln2	2ln2	-
$t_i \ln y_i$	0	0	2ln2	2ln2

1. Среднее геометрическое значение уровней: $(1\times2\times4)^{1/3} = 2$.
2. Параметр b равен $2\ln2/((-1)^2+0^2+1^2) = \ln2$.
3. Экспоненциальный тренд: $\tilde{y} = 2\exp(t\ln 2)$, или $\tilde{y} = 2^{t+1}$.

Логистический тренд

Логистический тренд – это график функции

$$\tilde{y} = \frac{1}{e^{a+bt}+1}.$$

При положительном значении параметра b она убывает, а при отрицательном возрастает. Тренд используется для описания процессов внедрения и устаревания товаров, насыщения рынков. Так, данная функция описывает изменение удельного веса жителей, имеющих сотовый телефон: сначала она близка к нулю и медленно растет, затем рост ускоряется, а после перегиба темп роста замедляется и затем она стремится к единице.

1. Обозначим: $y = \frac{1}{x+1}$, отсюда $x = \frac{1}{y} - 1$.
2. Новая переменная описывается экспоненциальной функцией:

$$x = e^a e^{bt}.$$

3. Используем полученную выше формулу экспоненциального тренда, заменив в ней y на x, получим:

$$\tilde{x} = \hat{x}\exp(t\frac{\sum t_i \ln x_i}{\sum t_i^{\,2}}).$$

4. Тогда искомый тренд задается формулой

$$\tilde{y} = \frac{1}{\tilde{x}+1} = \frac{1}{\hat{x}\exp(t\frac{\sum t_i \ln x_i}{\sum t_i^{\,2}})+1}.$$

Пример. Динамический ряд y: 0,11; 0,5; 0,8. Найдем значения x и их среднее геометрическое: $\sqrt[3]{8{,}09\times1\times0{,}25} = 1{,}265$. Найдем логарифмы x и запишем их в таблицу.

t_i	-1	0	1
y_i	0,11	0,5	0,8
x_i	8,09	1	0,25
$\ln x_i$	2,09	0	-1,39
$t_i \ln x_i$	-2,09	0	-1,39
$\tilde{y}_i$	0,12	0,44	0,82

Параметр b равен: $(-2{,}09+0-1{,}39)/((-1)^2+0^2+1^2) = -1{,}74$.

Формула логистического тренда: $\tilde{y} = \dfrac{1}{1{,}265e^{-1{,}74t} + 1}$.

Рассчитаем значения тренда: первое – 0,12, второе – 0,44, третье – 0,82.

Степенной тренд

Степенной тренд – это график функции $\tilde{y} = mT^k$, где числа m и k положительны, а время T принимает положительные значения, причем число уровней может быть четным. Используется в случае ускоренного возрастания (если скорость возрастания больше, чем у логарифмической и меньше, чем у экспоненциальной функции) или замедленного убывания.

1. Прологарифмируем формулу тренда: $\ln \tilde{y} = \ln m + k \ln T$.

2. Введем переменные $z = \ln \tilde{y}$ и $x = \ln T$, они связаны линейной зависимостью.

3. Построим линейную регрессию для признака z_i: $\tilde{z} = a + bx$.

4. Из сравнений формул $\ln \tilde{y}$ и $\tilde{z}$ следует, что $a = \ln m$, а из свойства линейной регрессии: $a = \bar{z} - b\bar{x}$. Учтем, что средний логарифм значений равен логарифму среднего геометрического значений, тогда

$$m = \exp(a) = \exp(\bar{z} - b\bar{x}) = \exp(\ln \hat{y} - b \ln \hat{T}) = \exp \ln(\hat{y} / \hat{T}^b) = \hat{y} / \hat{T}^b.$$

где $\hat{y}$ и $\hat{T}$ - средние геометрические значение заданных значений y_i и T_i.

5. Из сравнений формул $\ln \tilde{y}$ и $\tilde{z}$ также следует, что $b = k$. Используем формулу углового коэффициента линейной регрессии, получим:

$$k = b = \frac{\sum (z_i - \bar{z})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (\ln y_i - \ln \hat{y})(\ln T_i - \ln \hat{T})}{\sum (\ln T_i - \ln \hat{T})^2} = \frac{\sum \ln(y_i / \hat{y}) \ln(T_i / \hat{T})}{\sum \ln^2 (T_i / \hat{T})}.$$

Пример. В моменты времени 4,9,16 уровни равны 6,9,12.

i	T_i	y_i	$\ln(T_i / \hat{T})$	$\ln(y_i / \hat{y})$	$\ln(T_i / \hat{T}) \ln(y_i / \hat{y})$	$\ln^2(T_i / \hat{T})$
1	4	6	-0,732	-0,366	0,268	0,536
2	9	9	0,079	0,040	0,003	0,006
3	16	12	0,654	0,327	0,214	0,428
∑	-	-	-	-	0,485	0,970

1. Средние геометрические значения признаков:

$$\hat{T} = \sqrt[3]{4 \times 9 \times 16} = 8{,}32,$$

$$\hat{y} = \sqrt[3]{6 \times 9 \times 12} = 8{,}65.$$

2. Произведем расчеты в таблице. Параметр k степенного тренда равен отношению суммы предпоследнего и последнего столбцов таблицы:

$$k = b = 0{,}485/8{,}65 = 0{,}5.$$

3. Параметр m степенного тренда равен: $m = 8{,}65/8{,}32^{0{,}5} = 3$.

4. Итак, формула степенного тренда: $\tilde{y} = 3\sqrt{T}$.

Задачи

1. В январе-мае объем ежемесячной прибыли фирмы составил 19, 21, 19, 13 и 3 млн. руб. Найти: а) формулу параболического тренда; б) ожидаемую прибыль в июне.

2. В период 2006-2012 гг. ежегодная выручка фирмы составляла 24, 33, 40, 45, 48, 49 и 48 млн. руб. Найти:

а) расчетные формулы параметров параболического тренда для произвольного динамического ряда из семи уровней;

б) формулу параболического тренда выручки данной фирмы;

в) ожидаемый объем выручки в 2017 г.

3. Приведите пример динамического ряда, для которого средний цепной абсолютный прирост: а) равен угловому коэффициенту линейного тренда; б) больше его; в) меньше его.

4. Динамический ряд: 33, 15, 7, 4, 2. Найти формулу экспоненциального тренда (параметры рассчитать с точностью до 0,01).

5. Динамический ряд: 0,1; 0,2; 0,5; 0,7; 0,8. Найти формулу логистического тренда (параметры рассчитать с точностью до 0,01).

6. В моменты времени 4, 8, 15 уровни равны 10, 14, 21. Найти формулу степенного тренда (параметры рассчитать с точностью до 0,01).

7. Дайте ответ (верно/неверно):

1) Тренд: факторный признак называют уровнем

2) Основные виды трендов: число наблюдений нечетно

3) Тренд – это тенденция изменения уровня во времени

4) Регрессия – это частный случай тренда

5) Тренд: число факторных признаков не больше 1

6) Колеблемость – это отклонение уровня от среднего уровня

7) Линейный тренд: свободный член равен среднему уровню

8) Линейный тренд: свободный член равен хронологической средней

9) Число лет равно 5, тогда наклон линейного тренда не зависит от уровня в 3-м году

10) Параболический тренд: свободный член равен среднему уровню

11) Линейный тренд: 6+2t, тогда средний уровень равен 2

12) Число уровней линейного тренда равно 5, тогда средний квадрат аргумента равен 2

13) Построение параболического тренда: в матрице системы уравнений имеется 4 нулевых коэффициента

14) Формулы параметров экспоненциального тренда выводятся из формул параметров линейного тренда

15) Построение экспоненциального тренда: логарифм уровня есть линейная функция логарифма аргумента

16) Динамический ряд: 4,5,6,3,5, тогда линейный тренд возрастает

17) Динамический ряд: 6,8,5, тогда параболический тренд лучше экспоненциального

18) Линейный тренд возрастает, если последний уровень больше первого уровня.

9. ДИНАМИКА: КОЛЕБЛЕМОСТЬ

Основные понятия

Колеблемость – отклонение уровней динамического ряда от значений линейного тренда: $u_i = y_i - \tilde{y}_i$, где y_i – значение уровня, $\tilde{y}_i$ – значение линейного тренда в момент времени i.

Силу колебаний измеряет *среднее линейное отклонение* от тренда:

$$a = \frac{\sum_{i=2}^{n-1} |u_i|}{n-2}.$$

Если уровни динамического ряда образуют арифметическую прогрессию, то данный показатель равен нулю. Также используют аналогичный показатель – среднее квадратическое отклонение от тренда.

Поворотная точка – локальный экстремум последовательности u_i: максимум характеризуют как *пик*, минимум – как *дно*. Поворотной точкой может быть только внутренняя точка ряда, поэтому максимально возможное число поворотных точек равно n-2, где n – число уровней.

Колебание называют *циклическим*, если промежутки времени между всеми соседними точками «пик» и всеми соседними точками «дно» одинаковы и неизменны во времени. Это число называют *длиной цикла*.

Основные типы колеблемости.

Маятниковая колеблемость – каждая внутренняя точка является поворотной, при этом «пик» и «дно» сменяют друг друга в строгой последовательности, число поворотных точек равно n-2. Это краткосрочная циклическая колеблемость с длиной цикла 2.

Случайная колеблемость – «дно» и «пик» чередуются хаотично, число поворотных точек равно (2/3)(n-2).

Долгопериодическая циклическая колеблемость – это циклическая колеблемость с длиной цикла больше 2, число поворотных точек равно $2n/l$, где l – длина цикла.

Пример. В период 2001-2005 гг. объемы ВВП составили 34, 38, 45. 49, 54, а уровни инфляции – 6, 8, 11, 8, 5%. Найдем дефляторы ВВП, объемы реального ВВП, линейный тренд реального ВВП.

Показатель	2001	2002	2003	2004	2005
ВВП номинальный	34	38	45	49	54
Уровень инфляции, %	6	8	11	8	5
Дефлятор ВВП	1,06	1,145	1,271	1,372	1,441

ВВП реальный (y_i)	32,1	33,2	35,4	35,7	37,5
Тренд реального ВВП ($\tilde{y}_i$)	32,2	33,5	34,8	36,1	37,4
Отклонение от тренда (u_i)	-0,1	-0,2	0,6	-0,4	0,1

Линейный тренд для реального ВВП: $\tilde{y} = 1{,}3 + 34{,}8t$. Рассчитаем отклонения от тренда. Из таблицы следует, что в 2002 г. и 2004 г. имело место «дно», в 2003 г. – «пик», т.е. выявлена маятниковая колеблемость. Среднее линейное отклонение от тренда: (0,2+0,6+0,4)/3 = 0,4.

Автокорреляция

1. Рассмотрим динамической ряд со слабо выраженной тенденцией динамики, т.е. с небольшим угловым коэффициентом линейного тренда.

Автокорреляция – это корреляция между значениями одного и того же динамического ряда, но со сдвигом во времени.

Имеется динамический ряд $y_1, y_2, \ldots, y_n$. Число $m<n$ назовем *сдвигом* и построим две последовательности одинаковой длины n-m:

- последовательность исходных значений x: $y_1, \ldots, y_{n-m}$;
- последовательность сдвинутых значений z: $y_{1+m}, \ldots, y_n$.

Коэффициент автокорреляции при сдвиге m – это коэффициент корреляции между признаками x и z:

$$r_m = \frac{s_{xz}}{s_x s_z},$$

где s_{xz} – ковариация признаков, s_x и s_z – их средние квадратические отклонения. Поскольку последовательности x и z получены из одного динамического ряда, их дисперсии близки, и знаменатель можно заменить на дисперсию y. Получим приближенную формулу:

$$r_m \cong \frac{\sum_{i=1}^{n-m}(y_i - \bar{x})(y_{i+m} - \bar{z})}{\sum_{j=1}^{n}(y_j - \bar{y})^2} \times \frac{n}{n-m},$$

где $\bar{x} = \frac{\sum_{i=1}^{n-m} y_i}{n-m}$, $\bar{z} = \frac{\sum_{i=1}^{n-m} y_{i+m}}{n-m}$, $\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$.

Если данную формулу применять при значительном наклоне линейного тренда, то возникает эффект *ложной автокорреляции*, который приводит к неверным выводам.

Пример. Для динамического ряда 1,1; 1,9; 3,1; 3,9; 5,1 значения линейного тренда равны 1, 2, 3, 4, 5, а отклонения от тренда равны 0,1; -0,1, 0,1; -0,1, 0,1. Налицо маятниковая колеблемость: при единичном сдвиге точки «пик» совпадают с точками «дно», т.е. по смыслу следует ожидать, что коэффициент автокорреляции равен -1. Однако при таком сдвиге приведенная формула дает значение коэффициента автокорреляции (r_1) между исходной (1,1; 1,9; 3,1; 3,9) и сдвинутой последовательностями (1,9;

3,1; 3,9; 5,1) близкое к 1, поскольку точки (1,2), (2,3), (3,4), (4,5) лежат на одной прямой. Мы пришли к противоречию.

2. Рассмотрим динамический ряд с сильно выраженной тенденцией динамики, т.е. со значительным угловым коэффициентом линейного тренда.

Автокорреляция – это корреляция между отклонениями от тренда одного и того же динамического ряда, но со сдвигом во времени.

Коэффициент автокорреляции для такого ряда учитывает тенденцию динамики, а в знаменателе его формулы дисперсия отклонений от тренда заменена на «хронологическую дисперсию», при расчете которой отклонения в крайние моменты времени складываются с весовым коэффициентом 0,5. Коэффициент автокорреляции при сдвиге m равен:

$$r_m' = \frac{\sum_{i=1}^{n-m} u_i u_{i+m}}{\frac{u_1^2}{2} + \sum_{j=2}^{n-1} u_j^2 + \frac{u_n^2}{2}} \times \frac{n-1}{n-m},$$

где $u_i = y_i - \tilde{y}_i$ – отклонение от тренда в момент i.

Коэффициент автокорреляции позволяет выявлять *циклические колебания*: если при сдвиге m его значение равно единице, то колеблемость динамического ряда является циклической с длиной цикла не менее m. При этом длина цикла в точности равна m, если для всех сдвигов, меньших m, коэффициенты автокорреляции не равны единице. Если имеется несколько единичных значений коэффициента автокорреляции для различных сдвигов, то длина цикла равна минимальному сдвигу. Так, в случае маятниковой колеблемости (длина цикла равна двум) все коэффициенты автокорреляции с четными сдвигами равны единице, а с нечетными сдвигами – минус единице.

Пример. Для динамического ряда 3, 7, 5, 6, 4 линейный тренд задается формулой 5+0,1t. Определим коэффициент автокорреляции при сдвиге 2.

i	1	2	3	4	5	$\sum$
t_i	-2	-1	0	1	2	0
y_i	3	7	5	6	4	25
$\tilde{y}_i$	4,8	4,9	5	5,1	5,2	25
u_i	-1,8	2,1	0	0,9	-1,2	0
u_{i+2}	0	0,9	-1,2	-	-	-
$u_i u_{i+2}$	0	1,9	0	-	-	1,9
u_i^2	3,24	4,41	0	0,81	1,44	-

Коэффициент автокорреляции при сдвиге 2 равен:

$$r_2' = \frac{1{,}9}{1{,}62 + 4{,}41 + 0 + 0{,}81 + 0{,}72} \times \frac{5-1}{5-2} = 0{,}33.$$

Таким образом, с вероятностью 33% динамический ряд является циклическим с длиной цикла 2, т.е. он проявляет свойства маятниковой колеблемости. Этот вывод подтверждается тем, что все внутренние точки

являются поворотными: во 2-м и 4-и годах наблюдались пики (отклонения – 2,1 и 0,9), а в 3-м году – дно (отклонение – 0).

Корреляция рядов динамики

Рассмотрим динамические ряды x_i и y_i, построим для каждого линейный тренд и обозначим отклонения от тренда:

$$u_{xi} = x_i - \bar{x}; \quad u_{yi} = y_i - \bar{y}.$$

Коэффициент корреляции динамических рядов – это коэффициент корреляции между их отклонениями от трендов:

$$r_u = \frac{\sum u_{xi} u_{yi}}{\sqrt{\sum u_{xi}^2 \sum u_{yi}^2}}.$$

Уравнение регрессии для отклонений:

$$\tilde{u}_y = b u_x, \quad где \quad b = \frac{\sum u_{xi} u_{yi}}{\sum u_{xi}^2}.$$

Знак коэффициента корреляции динамических рядов цен товаров позволяет выявить товары-заменители (он отрицателен), взаимодополняемые товары (положителен) несопряженные товары (равен нулю).

Пример. Среднегодовая цена товара *X*: 14, 16, 19, 20, 21; товара *Y*: 10, 13, 15, 14, 18. Линейный тренд первой цены: 18+1,8*t*, второй цены 14+1,7*t*.

t_i	x_i	y_i	$\tilde{x}_i$	$\tilde{y}_i$	u_{xi}	u_{yi}	$u_{xi}u_{yi}$	u_{xi}^2	u_{yi}^2
-2	14	10	14,4	10,6	-0,4	-0,6	0,24	0,16	0,36
-1	16	13	16,2	16,2	-0,2	0,7	-0,14	0,04	0,49
0	19	15	18	18	1	1	1	1	1
1	20	14	19,8	19,8	0,2	-1,7	-0,34	0,04	2,89
2	21	18	21,6	21,6	-0,6	0,6	-0,36	0,36	0,36
∑	90	70	90	70	0	0	0,4	1,6	5,1

Коэффициент корреляции динамических рядов цен товаров равен:

$$r_u = \frac{0{,}4}{\sqrt{1{,}6 \times 5{,}1}} = 0{,}14.$$

Поскольку он положителен, товары являются взаимодополняемыми. Однако поскольку корреляция невелика, регрессия между отклонениями цен от тренда здесь лишена содержания: параметр регрессии положителен (0,4/1,6=0,25), в то время как в три года из пяти (2-й. 4-й, 5-й) отклонения цен товаров имели противоположные знаки.

Сезонные колебания

Сезонные колебания – это отклонения уровней динамического ряда от линейного тренда, вызванные сменой времен года. Сезонные колебания – циклические с длиной цикла 12 месяцев. Пример – цена яблок.

Опишем алгоритм построения нелинейного тренда, учитывающего фактор сезонности. Пусть уровни динамического ряда y_i охватывают m циклов (лет), т.е. $12m$ месяцев ($i = 1, \dots, 12m$).

Индекс сезонности для i–го месяца есть отношение уровня и значения линейного тренда:

$$I_i = \frac{y_i}{\tilde{y}_i}.$$

Средний индекс сезонности для k-го месяца года ($k = 1, \dots, 12$) равен среднему индексу сезонности для этого месяца, рассчитанному по всем годам:

$$\bar{I}_k = \frac{1}{m}(I_k + I_{k+12} + I_{k+24} + \dots) = \frac{1}{m}\sum_{i=0}^{m-1} I_{k+12i}.$$

Например, для трехлетнего периода средний индекс сезонности для января ($k = 1$) равен $\bar{I}_1 = (I_1 + I_{13} + I_{25})/3$, где I_1, I_{13} и I_{25} – индексы сезонности в январе 1-го, 2-го и 3-го года соответственно.

Тренд с учетом сезонности ($\tilde{y}_i^{\,1}$) получают умножением значений линейного тренда и среднего индекса сезонности для каждого месяца (i) рассматриваемого периода:

$$\tilde{y}_i^{\,1} = \bar{I}_i \times \tilde{y}_i.$$

Общая колеблемость (сумма квадратов отклонений значений уровня от среднего уровня) приблизительно равна сумме трех слагаемых, характеризующих фактор тренда, фактор сезонности и прочие факторы:

$$\sum_{i=1}^{12m}(y_i - \bar{y})^2 = \sum_{i=1}^{12m}(\tilde{y}_i - \bar{y})^2 + \sum_{i=1}^{12m}(\tilde{y}_i^{\,1} - \tilde{y}_i)^2 + \sum_{i=1}^{12m}(y_i - \tilde{y}_i^{\,1})^2, \text{ или}$$

$$\textit{Колеблемость} = \textit{Фактор тренда} + \textit{Фактор сезонности} + \textit{Прочие факторы}.$$

Пример. Исследуем динамику цены яблок, когда цикл состоит из зимы и лета, т.е. имеется маятниковая колеблемость. Данные о пяти периодах охватывают 2,5 цикла: 8, 2, 7, 2, 6, т.е. зимой цена была 8, 7, 6, летом – 2, 2.

1. Тренд цены яблок: $5\text{-}0{,}4t$, его значения: 5,8; 5,4; 5; 4,6; 4,2.
2. Индексы сезонности: $I_1 = 8/5{,}8 = 1{,}38$, …, $I_5 = 1{,}43$ (см. табл.).
3. Средний индекс сезонности для зимы: $\bar{I}_1 = (1{,}38 + 1{,}4 + 1{,}43)/3 = 1{,}4$.
4. Средний индекс сезонности для лета: $\bar{I}_2 = (0{,}37 + 0{,}43)/2 = 0{,}4$.
5. Тренд с учетом сезонности для зимних периодов:
$$\tilde{y}_1^{\,1} = 1{,}4 \times 5{,}8 = 8{,}1, \quad \tilde{y}_3^{\,1} = 1{,}4 \times 5 = 7, \quad \tilde{y}_5^{\,1} = 1{,}4 \times 4{,}2 = 5{,}9.$$
6. Тренд с учетом сезонности для летних периодов:
$$\tilde{y}_2^{\,1} = 0{,}4 \times 5{,}4 = 2{,}2, \quad \tilde{y}_4^{\,1} = 0{,}4 \times 4{,}6 = 1{,}8.$$
7. Средняя цена - 5, общая колеблемость равна $(8\text{-}5)^2 + \dots + (6\text{-}5)^2 = 32$.
8. Фактор тренда равен $(5{,}8\text{-}5)^2 + \dots + (4{,}2\text{-}5)^2 = 1{,}6$.
9. Фактор сезонности равен $(8{,}1\text{-}5{,}8)^2 + \dots + (5{,}9\text{-}4{,}2)^2 = 30{,}3$.
10. Прочие факторы: $(8\text{-}8{,}1)^2 + \dots + (6\text{-}5{,}9)^2 = 0{,}1$.

11. Итак, общая колеблемость (32) равна сумме фактора тренда (1,6), фактора сезонности (30,3) и прочих факторов (0,1). Фактор сезонности определяет 30,3/32 = 0,95 (95%) общей колеблемости.

i	1	2	3	4	5	$\sum$
t_i	-2	-1	0	1	2	0
y_i	8	2	7	2	6	25
$\tilde{y}_i$	5,8	5,4	5	4,6	4,2	25
I_i	1,38	0,37	1,4	0,43	1,43	-
$\tilde{y}_i^1$	8,1	2,2	7	1,8	5,9	25
$(y_i - \bar{y})^2$	9	9	4	9	1	32
$(\tilde{y}_i - \bar{y})^2$	0,64	0,16	0	0,16	0,64	1,6
$(\tilde{y}_i^1 - \tilde{y}_i)^2$	5,4	10,5	4	7,6	2,8	30,3
$(y_i - \tilde{y}_i^1)^2$	0,01	0,04	0	0,04	0,01	0,1

Задачи

1. Объемы ВВП в 2005-2009 гг.: 40, 46, 49, 56, 63, уровни инфляции: 8, 11, 5, 6, 8%. Найти:

а) линейный тренд для реального ВВП;

б) поворотные точки;

в) среднее линейное отклонение от тренда.

2. Для динамического ряда 9, 7, 5, 6, 5 найдите:

а) коэффициент автокорреляции при сдвиге 1;

б) поворотные точки.

3. Средняя цена за неделю: товара *X* – 14, 12, 12, 10, 9; товара *Y* – 28, 26, 25, 23, 22. Определите:

а) линейный тренд для цены каждого товара;

б) коэффициент корреляции динамических рядов цен товаров;

в) являются ли товары заменителями или взаимодополняемыми;

г) среднее отношение отклонений от трендов цен товаров *Y* и *X*;

д) для каких конкретных товаров наблюдается сходная динамика цен.

4. Цикл состоит из зимы и лета, т.е. имеется маятниковая колеблемость, уровни динамического ряда: 4, 8, 5, 9, 6. Найти:

а) значения линейного тренда в первую зиму и первое лето;

б) индексы сезонности для первой зимы и первого лета;

в) средние индексы сезонности для зимы и лета;

г) откорректированные значения тренда с учетом сезонности для первой зимы и первого лета;

д) общую колеблемость;

е) удельный вес фактора сезонности в общей колеблемости.

10. ИНДЕКСЫ

Индексы цен

Простой индекс цены – это отношение цены товара в текущем году к цене в базисном году. Сложный индекс показывает, во сколько раз изменилась стоимость набора товаров при неизменных объемах продаж.

Индекс Ласпейреса (I_L) показывает, во сколько раз увеличилась стоимость набора товаров, если их объемы равны базисным значениям:

$$I_L = \frac{\sum p_i^1 \times Q_i^0}{\sum p_i^0 \times Q_i^0},$$

где p_i^0 и p_i^1 – базисная и текущая цена i-го товара, Q_i^0 – базисный объем продаж i-го товара (суммирование производится по всему набору товаров).

Индекс Пааше (I_P) показывает, во сколько раз увеличилась стоимость набора товаров, если их объемы равны текущим значениям:

$$I_P = \frac{\sum p_i^1 \times Q_i^1}{\sum p_i^0 \times Q_i^1},$$

где Q_i^1 – текущий объем продаж i-го товара.

Индекс Фишера (I_F) равен среднему геометрическому значению индексов Ласпейреса и Пааше:

$$I_F = \sqrt{I_L \times I_P}.$$

Он устраняет недостаток этих двух индексов, который состоит в том, что они не учитывают изменения объемов продаж товаров. *Свойства* трех индексов:

1. Если все цены увеличились, то все индексы больше 1.
2. Если все цены уменьшились, то все индексы меньше 1.
3. Если все цены изменились в a раз, то все индексы равны a. Так, при неизменных ценах все индексы равны 1.
4. Если все объемы продаж изменились в a раз, то все три индекса одинаковы. Так, при неизменных объемах продаж все индексы равны отношению нового и старого значений суммарной выручки.

Пример. Цены и объемы продаж хлеба и молока заданы в таблице.

i	Продукт	p_i^0	p_i^1	Q_i^0	Q_i^1	$p_i^0 \times Q_i^0$	$p_i^0 \times Q_i^1$	$p_i^1 \times Q_i^0$	$p_i^1 \times Q_i^1$
1	Хлеб	10	12	5	6	50	60	60	72
2	Яблоки	20	25	3	4	60	80	75	100
Сумма		-	-	-	-	110	140	135	172

Индекс Ласпейреса – 135/110=1,227, Пааше – 172/140=1,229, Фишера – 1,228.

Индекс потребительских цен (ИПЦ) – отношение стоимости потребительской корзины (известного набора товаров и услуг) в текущих ценах к ее стоимости в базисных ценах. Относится к типу Ласпейреса.

Потребительская корзина разрабатывается на основе Международной классификации индивидуального потребления по целям (Classification Of Individual Consumption By Purpose), содержащей 11 товарных групп. В России отсутствуют группы: 1) алкогольные напитки, табачные изделия, 2) здравоохранение, 3) образование, 4) гостиницы, кафе, рестораны.

Отраслевой индекс цен показывает, во сколько раз выросли цены продукции отрасли (индекс цен промышленных товаров, сельскохозяйственной продукции и др.).

Дефлятор ВВП – отношение стоимости рыночной корзины в ценах текущего года к ее стоимости в ценах базисного года. Рыночная корзина содержит основные группы товаров, представленные с учетом их доли в ВВП текущего года. Дефлятор показывает, во сколько раз увеличился уровень цен за период после базисного года. В базисном году он равен единице, при инфляции он больше ее, при дефляции – меньше. Относится к типу Пааше.

Индексы неравенства доходов

Домохозяйства располагают по возрастанию дохода и разбивают на равные группы: от «беднейшей» до «богатейшей».

Коэффициент фондов (коэффициент дифференциации доходов) – отношение среднего (суммарного) дохода 10% богатейших домохозяйств к среднему (суммарному) доходу 10% беднейших. При абсолютном равенстве он равен 1. В России в 2010 г. он составил 16,5. Недостаток показателя – слабая чувствительность к изменениям доходов домохозяйств. Так, он не изменится, если доход домохозяйства из богатейшей группы увеличится или доход домохозяйства из беднейшей группы уменьшится.

Квинтильный коэффициент – отношение среднего (суммарного) дохода 20% богатейших домохозяйств к среднему (суммарному) доходу 20% беднейших домохозяйств. В России в 2010 г. он составил 9,2. Аналогично определяют *квартильный коэффициент* (квинта – 1/5, кварта – 1/4).

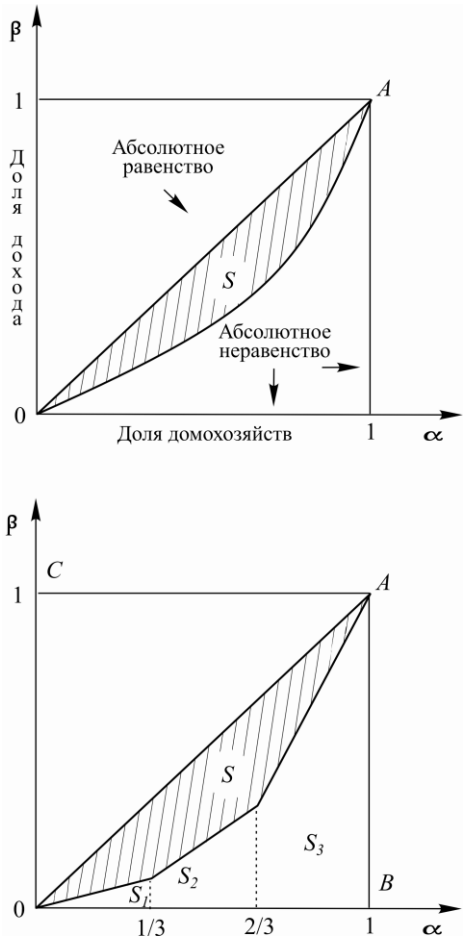

Рис.1. Кривая Лоренца и коэффициент Джини

Коэффициент Джини – чувствительный показатель неравенства, он учитывает большое число групп, которое здесь совпадает с числом домохозяйств. Пусть их число равно n, тогда удельный вес i беднейших домохозяйств равен $\alpha_i = i/n$. Для беднейшего домохозяйства он равен $1/n$, а для всех домохозяйств – 1. Пусть β_i – удельный вес дохода i беднейших домохозяйств в суммарном доходе всех домохозяйств, тогда очевидно $\beta_i \le \alpha_i$, причем равенство достигается при равенстве всех доходов.

Кривая Лоренца – ломаная, соединяющая соседние точки $(\alpha_i;\beta_i)$. При большом числе домохозяйств она имеет форму дуги (рис.1а). Ее свойства:

- ее концами служат точки O(0;0) и A(1;1);
- она возрастает и расположена под биссектрисой OA;
- при равном распределении доходов она лежит на биссектрисе OA;
- при абсолютном неравенстве (одно домохозяйство получает весь доход) она состоит из двух отрезков, один из которых лежит на оси абсцисс. Чем больше число домохозяйств, тем ближе угол между отрезками к 90^0;

- чем выше неравенство, тем больше площадь между ней и ОА.

Коэффициент Джини (индекс концентрации доходов) (*G*) равен удвоенной площади (*S*) между линией абсолютного равенства (ОА) и кривой Лоренца (см. рис.1). Он равен нулю при равенстве доходов и близок к единице при абсолютном неравенстве:

$$G = 2 \times S = 1 - \frac{2}{n}(\sum_{i=1}^{n-1} \beta_i + 0{,}5),$$

где n – число домохозяйств, β_i – доля дохода i беднейших домохозяйств.

Доказательство (n = 3). Искомая площадь S равна площади ΔAOB за вычетом площади треугольника S_1 и трапеций S_2 и S_3 (рис.1), отсюда

$$G = 2 \times S = 2 \times (\frac{1}{2} - S_1 - S_2 - S_3) =$$

$$= 2 \times (\frac{1}{2} - \frac{1}{2} \times \frac{1}{3} \times \beta_1 - \frac{1}{3} \times \frac{\beta_1 + \beta_2}{2} - \frac{1}{3} \times \frac{\beta_2 + 1}{2}) = 1 - \frac{2}{3}(\beta_1 + \beta_2 + 0{,}5).$$

Максимальное значение коэффициента Джини равно 1 – 1/n.

Пример. Доход студентов: Ивана – 6, Федора – 3, Глеба – 1. Тогда суммарный доход равен 10. Глеб – беднейший: β_1 = 1/10, Глеб и Федор – два беднейших: β_2 = 4/10. Тогда G = 1 – (2/3)×(1/10 + 4/10 + 0,5) = 0,33.

В России в 2010 г. коэффициент Джини составил 0,42.

Малоимущее домохозяйство (бедное) – имеет доход ниже прожиточного минимума. *Дефицит дохода* – сумма, необходимая для доведения дохода всех малоимущих хозяйство до прожиточного минимума.

В России в 2010 г. численность малоимущего населения составила 17,9 млн. человек, или 12,6% от общей численности населения. Дефицит дохода составил 380,2 млрд. руб., или 2,1% от расходов государственного бюджета.

Индекс Герфиндаля

Индекс Герфиндаля (Херфиндаля-Хиршнера) измеряет уровень концентрации производства (уровень конкуренции), он равен

$$H = \sum_{i=1}^{n} \alpha_i^{\,2},$$

где α_i – удельный вес (доля) объема продаж i-ой фирмы в общем объеме продаж всех фирм (в %), n – число фирм на рынке.

Свойства индекса:

- достигает максимума 10000 в случае монополии;
- достигает минимума 10000/N, если все фирмы одинаковы (N – число фирм на рынке);
- увеличивается после слияния двух фирм: $(d_1 + d_2)^2 > d_1^2 + d_2^2$. Слева – квадрат доли новой фирмы, справа – сумма квадратов долей старых фирм;
- уменьшается после разделения фирмы на несколько фирм.

Сложность сравнения индекса Герфиндаля для рынков разных стран состоит в том, что статистические данные содержат данные о долях разного числа крупнейших фирм: в России – 3, 4, 6 и 8, в США – 4, 8, 20 и 50.

Если фирмы разбиты на группы фирм с равным объемом продаж, то:

$$H = \sum_{j=1}^{m} n_j \beta_j^{\,2},$$

где β_j – удельный вес одной фирмы j-ой группы на рынке, n_j – число фирм в j-ой группе (сумма n_j равна n), m – число групп.

На практике концентрацию производства часто измеряют удельными весами объемов продаж нескольких крупнейших фирм. Пусть B_j – доля k_j крупнейших фирм ($B_0 = 0$, $B_m = 100$, $k_0 = 0$, $k_m = n$), тогда рынок разбивается на m групп, причем доля j-й группы равна

$$A_j = B_j - B_{j-1},$$

а число фирм в j–ой группе равно

$$n_j = k_j - k_{j-1}.$$

Последняя группа содержит самые мелкие фирмы, их число обычно составляет более 90% числа всех фирм.

Пример. Три крупнейшие фирмы занимают 30% рынка, а пять крупнейших – 40%. Всего на рынке 120 фирм. Тогда имеются три группы фирм с удельными весами 30%, 10% и 60%, причем $m = 3$, $k_1 = 3$, $k_2 = 5$, $k_3 = 120$, $B_1 = 30$, $B_2 = 40$, $B_3 = 100$, $A_1 = 30$, $A_2 = 10$, $A_3 = 60$, $n_1 = 3$, $n_2 = 2$, $n_3 = 115$. Если в каждой группе доли фирм одинаковы, то в первой группе эта доля равна $\beta_1 = 30/3=10\%$, второй – $10/2=5\%$, третьей – $60/115 = 0{,}52\%$.

Опишем алгоритмы нижней и верхней границ индекса Герфиндаля для случая, когда концентрация измеряется показателями B_j.

Нижняя граница. Предполагается, что все фирмы одной группы имеют равные объемы продаж, тогда доля фирмы j–ой группы равна

$$\beta_j = A_j/n_j.$$

Используем вторую формулу H для оценки нижней границы индекса Герфиндаля для электроэнергетики России в 2001 г. Доля крупнейших 3-х фирм – 15,2%, 4-х – 18,5%, 6-ти – 23,9%, 8-ми – 28,6%. Всего на рынке 1464 фирм. Всего 5 групп, для 2-й группы имеем:

$A_2 = B_2 - B_1 = 18{,}5 - 15{,}2 = 3{,}3\%$, $n_2 = k_2 - k_1 = 4 - 3 = 1$, $\beta_2 = A_2/n_2 = 1\%$.

Табл.1. Расчет нижней границы индекса Герфиндаля

j	k_j	B_j	A_j	n_j	β_j	$n_j\beta_j^2$
1	3	15,2	15,2	3	5,07	77,02
2	4	18,5	3,3	1	3,3	10,89
3	6	23,9	5,4	2	2,7	15,58
4	8	28,6	4,7	2	2,35	11,04
5	1464	100	71,4	1456	0,05	3,49
∑	-	-	100	1464	-	117

Источник: Промышленность России. 2002. Стат.сб./Госкомстат России. М., 2002.С.20,48.

Нижняя граница индекса Герфиндаля равна 117 (табл.1).

Верхняя граница. Предполагается, что в первой и последней группах дифференциация объемов продаж максимальная, в частности:

- в первой группе доля каждой фирмы кроме одной равна доле фирмы во второй группе, т.е. первая (крупнейшая) фирма образует отдельную группу, а оставшиеся фирмы первой и второй групп (старых) образуют новую вторую группу фирм с равными доля объема продаж;

- в последней группе часть фирм имеет нулевой объем продаж, а остальные имеет максимально возможную долю продаж, равную доле в предпоследней группе.

Алгоритм расчета верхней границы индекса состоит в определении новых значений числа фирм и долей фирм для трех групп:

- во 2-й группе: $n_2' = n_1 + n_2 - 1$; $\beta_2' = \beta_2$;
- в 1-й группе: $n_1' = 1$; $\beta_1' = B_2 - n_2'\beta_2$;
- в m–ой группе: $n_m' = A_m : \beta_{m-1}$; $\beta_m' = \beta_{m-1}$.

Табл.2. Расчет верхней границы индекса Герфиндаля

j	A_j'	n_j'	β_j'	$n_j'(\beta_j')^2$
1	8,6	1	8,6	73,96
2	9,9	3	3,3	32,67
3	5,4	2	2,7	14,58
4	4,7	2	2,35	11,04
5	71,4	30	2,35	165,67
∑	100	-	-	297,9

Источник: Там же.

Верхняя граница индекса Герфиндаля равна 297,9 (табл.2).

Индекс Герфиндаля приближенно равен средней величине нижней и верхней границ: Н = (117 + 297,9)/2 = 207,4.

Индекс развития человеческого потенциала

Индекс развития человеческого потенциала (ИРЧП), или *индекс человеческого развития* (ИЧР) – отражает главные факторы развития человека: долголетие, образование и доход. Он равен простой средней трех компонентных индексов:

$$\text{ИРЧП} = (\text{И1} + \text{И2} + \text{И3})/3,$$

где И1 – индекс ожидаемой продолжительности жизни, И2 – индекс уровня образования, И3 – индекс ВВП на душу населения. Индексы И1 и И3 рассчитывают по формуле

$$И = \frac{fact - \min}{\max - \min},$$

где *fact* – фактическое значение показателя, max и min – его максимальное и минимальное значения по странам мира.

Индекс уровня образования равен И2 = 2/3×У1 + 1/3×У2, где У1 – уровень грамотности (удельный вес грамотных в населении в возрасте 15 лет

и старше), У2 – удельный вес обучающихся в соответствующих возрастных группах (обычно 6-22 года).

ИРЧП и компонентные индексы лежат в пределах от 0 до 1.

Выделяют *три группы* стран: в первую входят страны с высоким уровнем ИЧРП, занимающие места с 1 по 70, во вторую (со средним уровнем) – с 71 по 155, в третью (с низким уровнем) – с 156 по 177.

Табл.1. Индекс развития человеческого потенциала в 2005 г.

Рейтинг	Страна	Индекс			
		ИРЧП	ожидаемой продолжитель-ности жизни	уровня образова-ния	ВВП на душу населения
1	Исландия	0,968	0,941	0,978	0,985
2	Норвегия	0,968	0,913	0,991	1,000
37	Польша	0,870	0,836	0,951	0,823
67	Россия	0,802	0,667	0,956	0,782
68	Албания	0,801	0,853	0,887	0,663
70	Бразилия	0,800	0,779	0,883	0,740
128	Индия	0,619	0,633	0,669	0,503

Индекс экономического настроения

В рамках Единой Европейской гармонизированной программы исследования настроений предпринимателей и потребителей в России с 1998 г. рассчитывают ежеквартально *индекс экономического настроения* (ИЭН):

$$ИЭН = (1/5)\times(ИПУ_п + ИПУ_с + ИПУ_т + ИПУ_у + ИПУ),$$ где

$ИПУ_п$ – *индекс предпринимательской уверенности в промышленности*:

$$ИПУ_п = (1/3)\times(ИС + ИЗ + ИОВ),$$

где ИС – индекс уровня спроса, ИЗ – индекс запасов готовой продукции, ИОВ – индекс ожидаемого выпуска продукции;

$ИПУ_с$ – индекс предпринимательской уверенности в строительстве;

$ИПУ_т$ – индекс предпринимательской уверенности в розничной торговле;

$ИПУ_у$ – индекс предпринимательской уверенности в сфере услуг;

ИПУ – *индекс потребительской уверенности*:

$$ИПУ = (1/5)\times(И_1 + И_2 + И_3 + И_4 + И_5),$$

где $И_1$ – индекс произошедших изменений личного материального положения, $И_2$ – индекс ожидаемых изменений личного материального положения, $И_3$ – индекс произошедших изменений в экономической ситуации в России, $И_4$ – индекс ожидаемых изменений в экономической ситуации в России, $И_5$ – благоприятность условий для крупных покупок.

Пример. Пусть 55% респондентов отметили увеличение уровня спроса, 35% – его уменьшение, 10% – нейтральные оценки (не учитываются), тогда ИС = 55 – 35 = 20%.

Пример. На вопрос об улучшении личного материального положения ответили: определенно положительно – 34%, скорее положительно – 28%,

нейтрально – 8% (не учитываются), скорее отрицательно – 18%, определенно отрицательно – 12%. Тогда $И_1$ = (34+28)–(18+12) = 30%.

Индекс экономического настроения в августе 2012 г. составил: Греция – - 65%, Италия – - 38%, Россия – - 6%, Швеция – + 10%.

Максимумы и минимумы композитных индексов предшествуют разворотам экономики (спаду или подъему в экономическом цикле), т.е. они служат «барометрами» экономики в краткосрочном периоде.

Задачи

1. Доходы домохозяйств: 36, 32, 30, 28, 42, 26, 6. Прожиточный минимум равен 27. Для покрытия дефицита дохода введен налог на доходы, равные 1,5 прожиточного минимума и выше. Найти:

а) коэффициент Джини до введения налога (до 0,01);

б) дефицит дохода;

в) ставку налога (до 0,1%);

г) во сколько раз сократится отношение чистого максимального дохода к минимальному после введения налога и устранения бедности.

2. Топливная промышленность России в 2001 г.: доля крупнейших фирм на рынке: 3 фирмы – 26,5%, 4 – 29,4%, 6 – 34,5%, 8 – 38,7%. Всего фирм 4804. Найти нижнюю и верхнюю границы индекса Герфиндаля.

3. Рынок подержанных автомобилей США в 2007 г.: доля крупнейших фирм на рынке: 4 фирмы – 13,3%, 8 – 14,3%, 20 – 16,2%, 50 – 18,8%. Всего фирм 26293. Найти нижнюю и верхнюю границы индекса Герфиндаля.

4. Найти индекс уровня образования:

а) Панама 2005 г.: уровень грамотности – 91,9%, удельный вес обучающихся – 79,5%;

б) Вьетнам 2005 г.: уровень грамотности – 90,3%, удельный вес обучающихся – 63,9%.

5. Болгария 2005 г.: уровень грамотности – 98,2%, удельный вес обучающихся – 81,5%, индекс ожидаемой продолжительности жизни – 0,795, индекс ВВП на душу населения – 0,752. Найти индекс развития человеческого потенциала.

11. СТАТИСТИЧЕСКИЕ МОДЕЛИ

Модель миграции населения

Рассматривается миграция населения между n регионами страны, естественное движение (рождаемость, смертность) не учитывается.

Пусть a_{ij} – вероятность переезда жителя i-го региона в j-й регион в течение года, тогда матрицу $A = \{a_{ij}\}$ назовем *переходной,* ее свойства:

а) число на главной диагонали равно вероятности остаться в регионе в течение года, он является максимальным элементом в строке и столбце,

б) матрица является квадратной;

в) сумма элементов каждой строки равна 1, поскольку человек может либо остаться проживать в своем регионе, либо переехать в другой регион;

Графическая интерпретация переходной матрицы дана на рис.1.

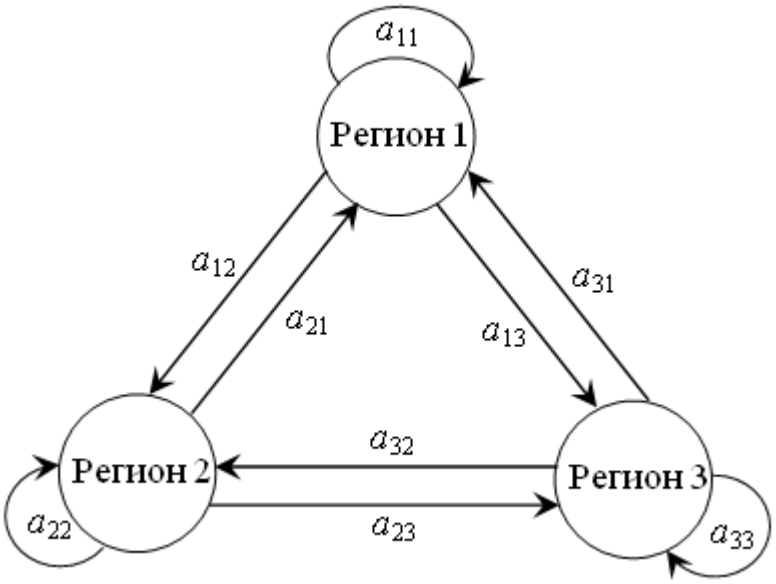

Рис.1. Переходная матрица миграции

Переходная матрица миграции *k*-го порядка A^k обладает свойствами «б» и «в» переходной матрицы миграции. Ее элемент, расположенный на пересечении *i*-ой строки и *j*-го столбца, равен вероятности переезда жителя *i*-го региона в *j*-й регион в течение *k* лет. В табл.1 представлены переходные матрицы миграции 1-го и 2-го порядка, из нее следуют два вывода:

- диагональные элементы второй матрицы меньше, чем первой и т.д., т.е. вероятность остаться в своем регионе уменьшается со временем;

- если вероятность переезда в течение года из одного региона в другой равна нулю ($a_{12} = a_{31} = 0$), то соответствующий переезд возможен за более длительный период времени через другие регионы.

Табл.1. Переходная матрица миграции

Регион	Переходная матрица миграции	
	1-го порядка (A)	2-го порядка (A^2)
1	0,9 0 0,1	0,81 0,03 0,16
2	0,1 0,7 0,2	0,16 0,55 0,29
3	0 0,3 0,7	0,03 0,42 0,55

Рассмотрим *замкнутую систему* – без внешней миграции. Считаем переходную матрицу миграции неизменной во времени. *Вектор распределения* населения по регионам в начальном году:

$$X_0 = (x_1^0; x_2^0; ...; x_n^0),$$

где x_i^0 - численность населения *i*-го региона в начальном году. Численность населения страны равна сумме координат данного вектора.

Пусть число регионов равно трем, тогда численность населения первого региона в конце первого года равна сумме трех слагаемых:

$x_1^0 a_{11}$ – число жителей 1-го региона, которые не изменили место проживания в течение первого года;

$x_2^0 a_{21}$ – число жителей 2-го региона, которые переехали в 1-й регион в течение первого года;

$x_3^0 a_{31}$ – число жителей 3-го региона, которые переехали в 1-й регион в течение первого года. Тогда число жителей 1-го региона через год равно

$$x_1^1 = x_1^0 a_{11} + x_2^0 a_{21} + x_3^0 a_{31}.$$

Число жителей 2-го и 3-го регионов в конце первого года равно:

$$x_2^1 = x_1^0 a_{12} + x_2^0 a_{22} + x_3^0 a_{32};$$

$$x_3^1 = x_1^0 a_{13} + x_2^0 a_{23} + x_3^0 a_{33}.$$

Отсюда вектор распределения в конце 1-го года X_1 равен произведению вектора распределения в начале 1-го года и переходной матрицы миграции:

$$X_1 = X_0 \times A.$$

Сумма координат X_1 равна неизменной численности населения страны.

Теперь примем начало 2-го года за начальный момент. Тогда вектор распределения в конце 2-го года X_2 равен:

$$X_2 = X_1 \times A = X_0 \times A \times A = X_0 \times A^2,$$

Обобщив данную формулу, получим:

$$X_k = X_0 \times A^k,$$

где X_k – вектор распределения в конце *k*-го года.

Пример. Начальное распределение – (10;30;20). Переходные матрицы даны в табл.1. Рассчитаем численность населения регионов в конце 1-го года:

$$x_1^1 = 10 \times 0{,}9 + 30 \times 0{,}1 + 20 \times 0 = 12;$$

$$x_2^1 = 10 \times 0 + 30 \times 0{,}7 + 20 \times 0{,}3 = 27;$$

$$x_3^1 = 10 \times 0{,}1 + 30 \times 0{,}2 + 20 \times 0{,}7 = 21.$$

Рассчитаем численность населения регионов в конце 2-го года:

$$x_1^2 = 12 \times 0{,}81 + 27 \times 0{,}16 + 21 \times 0{,}03 = 14{,}67;$$

$$x_2^2 = 12 \times 0{,}03 + 27 \times 0{,}55 + 21 \times 0{,}42 = 24{,}03;$$

$$x_3^2 = 12 \times 0{,}16 + 27 \times 0{,}29 + 21 \times 0{,}55 = 21{,}30.$$

Численность населения страны неизменно равна 60.

Переходная матрица миграции бесконечного порядка (A^∞), как можно доказать, имеет одинаковые строки, а в каждом ее столбце все элементы одинаковы:

$$A^{\infty} = \lim A^k = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_1 & a_2 & a_3 \\ a_1 & a_2 & a_3 \end{pmatrix},$$

где a_i – *миграционная привлекательность* i-го региона, или ожидаемый удельный вес населения региона в общей численности населения страны в долгосрочной исторической перспективе. Приближенные оценки миграционной привлекательности можно получить, сложив элементы столбцов переходной матрицы A. Чем больше эта сумма, тем интенсивнее иммиграционные потоки в регион, тем больше его миграционная привлекательность.

Пример. Задана переходная матрица:

$$A = \begin{pmatrix} 0{,}8 & 0{,}2 & 0{,}0 \\ 0{,}0 & 0{,}9 & 0{,}1 \\ 0{,}2 & 0{,}1 & 0{,}7 \end{pmatrix}.$$

Наибольшей миграционной привлекательностью обладает 2-й регион, т.к. сумма элементов 2-го столбца максимальна (1,1), а наименьшей – 3-й регион, т.к. сумма элементов 3-го столбца минимальна (0,8). Расчет точных значений миграционной привлекательности (a_i) связан с техническими сложностями, т.к. при возведении матрицы в большие степени ошибки накапливаются, что приводит к искажению элементов искомой матрицы A^{∞}.

Рассмотрим две модели *открытой системы* – с внешней миграцией.

1. *Модель с вектором внешней миграции*. Пусть неизменным остается вектор внешней миграции $B = (b_1 ; \ldots ; b_n)$, где b_i – сальдо внешней миграции для i-го региона, т.е. разность численности въехавших в него иностранцев и выехавших из него за границу россиян за год. Сумма элементов вектора B равна миграционному приросту населения, если она положительна, то население сокращается, в противном случае оно увеличивается. Вектор распределения населения по регионам в конце первого года:

$$X_1 = X_0 \times A + B .$$

Теперь примем начало 2-го года за начальный момент, тогда:

$$X_2 = X_1 \times A + B = X_0 \times A^2 + B \times A + B .$$

Обобщив данную формулу, получим распределение в конце k-го года:

$$X_k = X_0 \times A^k + B \times A^{k-1} + \ldots + B \times A + B .$$

Пример. Начальное распределение – (10;30;20), внешняя миграция – (3;0;–1). Переходная матрица – в табл.1. Используем полученные данные:

$$x_1^1 = 12 + 3 = 15 ; \qquad x_2^1 = 27 + 0 = 27 ; \qquad x_3^1 = 21 - 1 = 20 .$$

Рассчитаем вспомогательный вектор: $B \times A = (2{,}7;-0{,}3;-0{,}4)$.

Используем полученные ранее данные:

$$x_1^2 = 14{,}67 + 2{,}7 + 3 = 20{,}37 ; x_2^2 = 24{,}03 - 0{,}3 + 0 = 23{,}73 ; x_3^2 = 21{,}30 - 0{,}4 - 1 = 19{,}9 .$$

Итак, население страны по годам составило 60, 62 и 64 (годовой миграционный прирост равен $3 + 0 - 1 = 2$).

2. *Модель с расширенной переходной матрицей*. Обозначим: x_{n+1}^{k} – численность населения внешнего мира в k-ом году, $a_{i,n+1}$ – вероятность эмиграции жителя i–го региона в течение года, $a_{n+1,i}$ – вероятность иммиграции жителя внешнего мира в i-й регион страны в течение года. Тогда матрица A' размерности $(n+1)\times(n+1)$ называется расширенной переходной матрицей миграции, а вектор X_k' размерности $(n+1)$ – расширенным вектором распределения населения в k-ом году. Сумма его первых n координат равна численности населения страны, а сумма всех координат - численности населения всех стран. Поскольку система всех стран мира замкнута, данный подход позволяет применять полученные выше формулы. Так, расширенный вектор распределения населения через k лет равен $X_k' = X_0' \times (A')^k$.

Модель естественного движения населения

Рассматривается рождаемость и смертность, а миграция не учитываются. Имеются четыре возрастные группы: дети, младшая и старшая фертильные группы и пенсионеры. Население фертильного возраста способно создавать потомство, в отличие от детей и пенсионеров. Половой состав возрастных групп не влияет на рождаемость и смертность.

Опишем переходную матрицу естественного движения населения. Ее элемент c_{ij} равен вероятности перехода представителя i-ой возрастной группы в j-ю группу в течение года. В отличие от модели миграции, многие переходы между возрастными группами не возможны, а соответствующие элементы переходной матрицы равны нулю. Так, представитель группы «дети» не может через год стать «пенсионером» ($c_{14} = 0$). Переходная матрица имеет две группы ненулевых элементов: первая описывает смертность, вторая – рождаемость.

Смертность. Обозначим через $c_{i,i+1}$ вероятность остаться в живых через год для представителя i-ой возрастной группы и назовем данный показатель *коэффициентом дожития*. Он определяется для всех групп, за исключением старшей. Сумма коэффициента дожития и *коэффициента смертности* равна единице. Коэффициенты дожития расположены над главной диагональю переходной матрицы естественного движения населения.

Рождаемость. Обозначим через $c_{i,1}$ вероятность рождения ребенка в течение года для представителей i-ой группы и назовем данный показатель *коэффициентом рождаемости*. Его положительное значение для фертильной группы показывает, что ее представитель способен «перейти» в группу «дети», делегировав туда своего «представителя». Коэффициенты рождаемости расположены в первом столбце переходной матрицы. Теоретически данный коэффициент может быть больше 1. Если он равен 1,2, то на 100 представителей этой группы рождается в среднем 120 детей в год.

Свойства переходной матрицы естественного движения населения: а) число ее строк (столбцов) равно числу возрастных групп; б) ее элементы меньше единицы; в) ненулевые элементы расположены над главной диагональю и в первом столбце; г) коэффициент дожития сначала растет (0-5 лет), а затем уменьшается. Переходная матрица естественного движения

населения показана на рис.2. Переход в следующую возрастную группу изображен сплошной стрелкой, а рождение ребенка – пунктирной стрелкой.

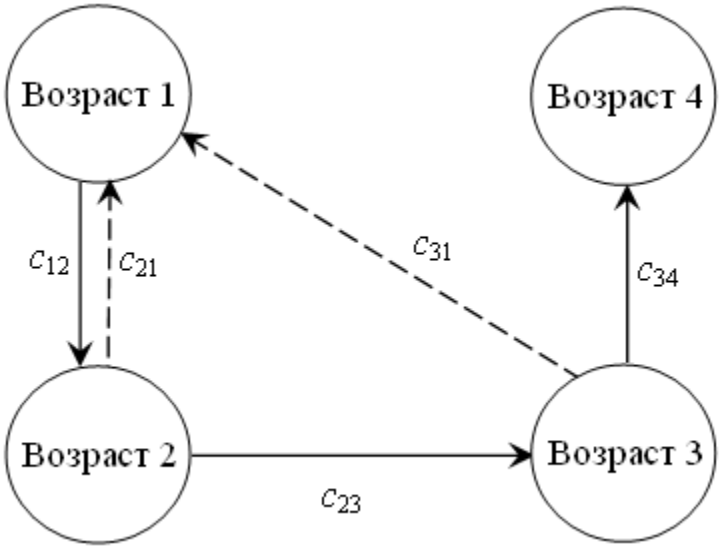

Рис.2. Переходная матрица естественного движения населения

Назовем *k*-ю степень матрицы *C* переходной матрицей *k*-го порядка и обозначим через C^k. Ее элемент, расположенный на пересечении *i*-ой строки и *j*-го столбца, равен вероятности перехода представителя *i*-ой группы в *j*-ю группу через *k* лет. Если он положителен, то возможно наступление одного из двух событий: либо человек в возрасте *i* лет доживает до возраста *j* лет, либо он через *k* лет имеет потомков в возрасте *j* лет.

Табл.2. Переходная матрица естественного движения населения

Возрастная группа	Переходная матрица естественного движения населения							
	первого порядка (*C*)				второго порядка (C^2)			
1	0	0,9	0	0	0,54	0	0,72	0
2	0,6	0	0,8	0	0,16	0,54	0	0,56
3	0,2	0	0	0,7	0	0,18	0	0
4	0	0	0	0	0	0	0	0

В табл.2 показаны переходные матрицы естественного движения населения первого и второго порядка. Из нее следует:

а) для младшей группы вероятность прожить два года равна 72%;

б) для младшей фертильной группы вероятность иметь через два года потомство в той же возрастной группе равна 54%;

в) для старшей фертильной группы вероятность иметь через два года потомство в младшей возрастной группе равно нулю.

Исследуем рождаемость и смертность в *замкнутой экономике*, если переходная матрица естественного движения населения неизменна.

Вектором распределения населения по возрастным группам в начале периода наблюдений называют вектор

$$Y_0 = (y_1^0; \ldots; y_m^0),$$

где y_i^0 - начальная численность *i*-ой возрастной группы, *m* - число групп. Начальная численность населения страны равна сумме координат данного вектора. Численность детей в конце 1-го года равна сумме $y_2^0 c_{21}$ и $y_3^0 c_{31}$ – количества детей, родившихся за год в младшей и старшей фертильной группах соответственно. Численность детей в конце 1-го года:

$$y_1^1 = y_1^0 c_{11} + y_2^0 c_{21} + y_3^0 c_{31} + y_4^0 c_{41}.$$

Численность второй группы в конце 1-го года равна числу детей, доживших до конца первого года, т.е. равна $y_1^0 c_{12}$, или:

$$y_2^1 = y_1^0 c_{12} + y_2^0 c_{22} + y_3^0 c_{32} + y_4^0 c_{42}.$$

Аналогично образом получим соотношения:

$$y_3^1 = y_1^0 c_{13} + y_2^0 c_{23} + y_3^0 c_{33} + y_4^0 c_{43}, \quad y_4^1 = y_1^0 c_{14} + y_2^0 c_{24} + y_3^0 c_{34} + y_4^0 c_{44}.$$

Итак, вектор распределения населения по возрастным группам в конце 1-го года Y_1 равен произведению начального вектора распределения и переходной матрицы естественного движения населения:

$$Y_1 = Y_0 \times C.$$

Сумма координат этого вектора равна численности населения страны в конце 1-го года. Обобщив данную формулу, получим:

$$Y_k = Y_0 \times C^k,$$

где Y_k - вектор распределения населения в конце *k*-го года, C^k - переходная матрица естественного движения населения *k*-го порядка.

Пример. Начальное распределение – (60;40;50;10). Переходные матрицы – в табл.2. Рассчитаем численности групп в конце 1-го года: $y_1^1 = 34$; $y_2^1 = 54$; $y_3^1 = 32$; $y_4^1 = 35$. Рассчитаем численности групп в конце 2-го года: $Y_2 = Y_0 \times C^2 = (38{,}8;\ 30{,}6;\ 43{,}2;\ 22{,}4)$. Численность населения равна: в конце 1-го года – 155, в конце 2-го – 135, т.е. она сокращалась (табл.3).

Табл.3. Естественное движение населения в замкнутой экономике

Возрастная группа	Численность возрастных групп в начале года		
	1-й год	2-й год	3-й год
1	60	34	38,8
2	40	54	30,6
3	50	32	43,2
4	10	35	22,4
Итого	160	155	135

Исследуем совместно естественное движение населения и внешнюю миграцию, полагая неизменными переходную матрицу естественного движения населения и *вектор внешней миграции*

$$D = (d_1;\ d_2;\ \ldots;\ d_m),$$

где d_i - сальдо внешней миграции за год для *i*-ой группы, т.е. разность численности иммигрантов и эмигрантов, принадлежащих к данной группе, *m*

- число возрастных групп. Сумма координат вектора внешней миграции есть сальдо внешней миграции для всей экономики. Здесь, в отличие от модели миграции, данный показатель не равен приросту численности населения страны за год. Вектор распределения в конце 1-го года:

$$Y_1 = Y_0 \times C + D.$$

Примем начало 2-го года за начальный момент, тогда:

$$Y_2 = Y_0 \times C^2 + D \times C + D.$$

Обобщим данную формулу: $Y_k = Y_0 \times C^k + D \times C^{k-1} + ... + D \times C + D$, где $Y_k -$ вектор распределения в конце *k*-го года.

Пример. Начальное распределение – (60;40;50;10), вектор внешней миграции – (2;7;4;-3). Переходная матрица – в табл.2. Тогда сальдо внешней миграции равно $2+7+4-3=10$, а численности групп в конце 1-го года:

$y_1^1 = 34+2 = 36$; $y_2^1 = 54+7 = 61$; $y_3^1 = 32+4 = 36$; $y_4^1 = 35-3 = 32$.

Рассчитаем численность групп в конце 2-го года. Для этого рассчитаем $D \times C = (5;\ 1{,}8;\ 5{,}6;\ 2{,}8)$. Используя полученные ранее формулы и данные:

$y_1^2 = 38{,}8+5+2 = 45{,}8$; $y_2^2 = 30{,}6+1{,}8+7 = 39{,}4$;

$y_3^2 = 43{,}2+5{,}6+4 = 52{,}8$; $y_4^2 = 22{,}4+2{,}8-3 = 22{,}2$.

Как следует из табл.4, динамика численности населения в открытой экономике определялась разнонаправленным действием двух факторов: фактора сокращения населения (смертность превышает рождаемость) и фактора внешней миграции (иммиграция превышает эмиграцию). В итоге численность населения увеличилась в первом году и сократилась во втором году.

Табл.4. Естественное движение населения в открытой экономике

Возрастная группа	Численность возрастных групп в начале года		
	1-й год	2-й год	3-й год
1	60	36	45,8
2	40	61	39,4
3	50	36	52,8
4	10	32	22,2
Итого	160	165	160,2

Задачи

1. Вероятность остаться в регионе в течение года равна: для жителей региона А – 80%, В – 90%, С – 70%. Вероятность переезда равна: из А в В – 20%, из В в С – 10%, из С в А – 20%, из С в В – 10%. Численность населения регионов: А – 20, В – 70, С – 10. Найдите численность регионов через год и укажите регион с наименьшей миграционной привлекательностью.

2. Имеются три возрастные группы, причем средняя группа – фертильная. Коэффициент рождаемости – 0,3, коэффициенты дожития – 0,9, 0,8 и 0. Численность групп – 50, 30 и 20. Найдите:

а) численность возрастных групп через год;

б) изменение численности населения страны за год;

в) вероятность для представителей младшей группы иметь через два года «представителей» (сам человек или его потомство) в каждой группе.

12. АНАЛИЗ СТРУКТУРНЫХ СДВИГОВ

Традиционные статистические методы анализа структурных сдвигов лишены наглядности и слабо адаптированы для графического представления динамических процессов, поскольку они не позволяют достаточно просто интерпретировать характер и направленность данного процесса, происходящего в многомерном пространстве:

- в нем традиционные методы анализа динамических рядов неприменимы, в то время как само содержание динамического процесса требует построения некого аналога линейного тренда и разработки соответствующих статистических методов;
- все точки-структуры пространства лежат в ограниченном подмножестве некоторого гиперпространства, т.к. координаты структуры положительны, а их сумма равна единице. Это налагает ограничения на параметры динамики структуры и, в частности, исключает возможность долгосрочной линейной тенденции.

Опишем метод визуализации структурной динамики и применим его для анализа динамики структуры расходов федерального бюджета США в 2001-2011 гг. Она характеризуется десятью основными статьями: пенсионная система, здравоохранение, образование, оборона, социальное обеспечение, правоохранительная деятельность (полиция, суды, тюрьмы и др.), транспорт, управление, другие расходы, процентные платежи.

Таблица 1. Структура расходов федерального бюджета США

Ста-тья	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Pens	25,1	25,5	24,7	23,6	23,1	22,6	22,1	23,2	22,1	20,8	21,7	21,5
Heal	19,6	20,9	21,2	21,7	22,2	22,2	21,9	23,5	22,5	21,7	23,7	23,8
Edu	3,3	3,4	3,9	4,2	4,2	4,3	4,8	3,7	3,4	2,5	4,0	3,2
Defe	20,0	19,7	21,0	22,3	23,6	24,3	23,4	23,9	24,5	22,6	24,5	24,4
Wel	9,9	10,1	11,4	11,5	10,6	10,2	9,6	9,6	10,8	11,8	14,5	13,1
Prot	1,6	1,6	1,7	1,6	2,0	1,6	1,5	1,6	1,6	1,5	1,6	1,6
Tran	2,6	2,9	3,1	3,1	2,8	2,7	2,6	2,7	2,6	2,3	2,7	2,6
Gov	0,9	0,9	0,9	1,1	1,0	0,8	0,7	0,7	0,7	0,6	0,7	0,8
Othe	4,5	3,9	3,6	3,6	3,3	3,8	4,7	2,6	3,3	10,7	0,8	2,7
Inter	12,4	11,1	8,5	7,1	7,0	7,4	8,5	8,7	8,5	5,3	5,7	6,4

Статьи расходов: Pensions, Health Care, Education, Defense, Welfare, Protection, Transportation, General Government, Other Spending, Interest.
Источник: www.usfederalbudget.us

Метод предполагает расчет трех основных характеристик динамики в многомерном пространстве: сдвиг, отклонение и удаление. Структура бюджетных расходов в *i*–ом году изображается точкой многомерного эвклидова пространства

$$A_i = (a_1^{\,i}, a_2^{\,i}, \ldots, a_n^{\,i}),$$

где $a_k^{\,i}$ – удельный вес расходов на k–ю статью бюджета в i–ом году (в процентах), n - число статей бюджетных расходов. Соединив соседние точки, мы получим ломаную траекторию структуры в многомерном пространстве, описывающую процесс ее динамики.

Сдвиг структуры есть эвклидово расстояние между точками, задающими структуру в текущем и последующем году. Квадрат расстояния между структурами в i–м и последующем году равен

$$d_{i,i+1}^{\;2} = |A_i A_{i+1}|^2 = \sum_{k=1}^{n} (a_k^{\,i+1} - a_k^{\,i})^2.$$

Суммарное значение сдвига равно длине траектории, по которой двигалась точка-структура в многомерном пространстве за весь период времени. Единицей измерения сдвига является процент.

Коэффициент динамического отклонения структуры бюджетных расходов в i–м году есть косинус угла между входящим звеном ломаной $A_{i-1}A_i$ и ее исходящим звеном A_iA_{i+1}, он равен косинусу соответствующего угла в треугольнике $A_{i-1}A_iA_{i+1}$ и рассчитывается по теореме косинусов:

$$c_i = \frac{d_{i,i-1}^{\;2} + d_{i,i+1}^{\;2} - d_{i-1,i+1}^{\;2}}{2 d_{i-1} d_{i+1}}.$$

Угол отклонения α_i равен арккосинусу коэффициента динамического отклонения c_i, он измеряется в градусах, лежит в пределах от 0 до 180^0 и показывает характер изменения структуры бюджетных расходов в каждом году (изменение ее «курса»):

- если он равен минус единице, то входящее звено ломаной и ее исходящее звено лежат на одной прямой, т.е. структура бюджетных расходов изменилась линейным образом, а курс сохранился неизменным. В этом случае говорят о последовательной, преемственной политике бюджетных расходов;

- чем больше отклонение показателя от -1, тем более существенным образом изменяется направление (курс) динамики структуры расходов.

Принята следующая градация значений коэффициента динамического отклонения:

а) от -1 до -1/3 – динамику характеризуем как отклонение (изменение курса, близкое к линейному);

б) от -1/3 до 1/3 – поворот (изменение, близкое к повороту на 90^o);

в) от 1/3 до 1 – разворот (изменение, близкое к развороту на 180^o).

Показатели «сдвиг структуры» и «коэффициент динамического отклонения» являются статистически независимыми.

Удаление – это эвклидово расстояние от текущей структуры до фиксированной базовой структуры:

$$R_i = |A_0 A_i|,$$

где A_0 – базовая структура. В США в 2000 г. она была равна (25,1; 19,6; 3,3; 20,0; 9,9; 1,6; 2,6; 0,9; 4,5; 12,4).

Таблица 2. Параметры динамики структуры бюджетных расходов США

№	Период	Значения параметров		Качество параметров	
		Сдвиг	Курс	Сдвиг	Курс
1	2001/2002	3,36	-0,53	большой	отклонение
2	2002/2003	2,29	-0,89	средний	отклонение
3	2003/2004	1,83	-0,35	небольшой	отклонение
4	2004/2005	1,24	-0,25	небольшой	поворот
5	2005/2006	1,95	-0,36	небольшой	отклонение
6	2006/2007	2,75	0,61	средний	разворот
7	2007/2008	2,10	0,55	средний	разворот
8	2008/2009	8,54	-0,46	большой	отклонение
9	2009/2010	10,78	0,87	большой	разворот
10	2010/2011	2,60	0,84	средний	разворот

В табл.2 приведены данные о динамике расходов федерального бюджета США в 2001-2011 гг., из нее следует, что

- в докризисный период 2001-2006 гг. коэффициент динамического отклонения («курс») был отрицателен и большим по модулю. Политика расходов носила поступательный характер, а траектория динамики была близка к прямой линии (исключение составил 2005 г., когда траектория динамики демонстрировала поворот). При этом значения сдвига структуры также были невелики (за исключением 2002 г.);

- в кризисный период 2006-2011 гг. коэффициент динамического отклонения был положительным и большим по модулю. Политика расходов носила непоследовательный, маятниковый характер и характеризовалась как «разворот» (исключение составил 2009 г., когда траектория динамики демонстрировала отклонение). Наибольшие различия параметров динамики от средних значений были зафиксированы в 2010 г., когда сдвиг структуры и коэффициент отклонения достигли максимальных значений 10,78 и 0,87. В данном году завершились дорогостоящие антикризисные мероприятия, что позволило сократить затраты по статье «Прочие расходы» с 377,1 до 29,7 млрд.долл., т.е. в 12,7 раз, при этом расходы на образование возросли в полтора раза с 89,8 до 139,4 млрд. долл. В предыдущем 2009 г. наблюдалась противоположная тенденция: прочие расходы выросли в 3,8 раза, а затраты на образование сократились на 11%. Столь резкое изменение структуры расходов в 2010 г. отражено максимальными значениями двух основных параметров динамики. В 2009 г. основными тенденциями изменения структуры расходов были увеличение прочих расходов (антикризисные меры), снижение удельного веса процентных выплат, снижение расходов на образование, увеличение расходов на социальное обеспечение. Поскольку указанные тенденции наблюдались в предыдущем 2008 г., политика бюджетных расходов была поступательной (отклонение), что отражено отрицательным коэффициентом структурных сдвигов (-0,46).

Усредненные показатели, характеризующие изменение структуры динамики расходов за исследуемый период и за два периода: докризисный (2001-2006 гг.) и кризисный (2006-2011 гг.):

- среднее значение сдвига $\bar{d}$ характеризует изменчивость структуры и выполняет функцию дисперсии, в докризисный период оно более чем в 2,5 раза меньше, чем в кризисный период (табл.3);

- среднее динамическое отклонение $\bar{c}$ характеризует форму траектории динамики. Если его значение равно -1, то в каждом году изменение носит строго линейный характер, и все точки лежат на одной прямой. Если оно близко к -1, то наблюдается линейная тенденция изменения структуры, и может быть построен многомерный линейный тренд, т.е. точки группируются вокруг некоторой прямой многомерного пространства. Если оно близко к единице, то в каждом году наблюдается возвратная тенденция и динамика структуры является по сути маятниковой. В докризисный период среднее динамическое отклонение составило -0,48, что говорит о поступательном, последовательном характере политики бюджетных расходов. В кризисный период оно равно 0,48, что говорит о непоследовательной, разнонаправленной политике расходов (табл.3);

Таблица 3. Усредненные параметры динамики структуры бюджетных расходов США

№	Показатель	2001-2006	2006-2011	2001-2011
1	Средний сдвиг ($\bar{d}$)	2,13	5,35	3,74
2	Среднее отклонение ($\bar{c}$)	-0,48	0,48	0,00
3	Коэффициент линейности (l)	0,56	0,20	0,20
4	Угловой коэффициент тренда (b)	0,46	0,95	0,54
5	Характер динамики	Отклонение	Разворот	Поворот

- коэффициент линейности динамики структуры l – отношение расстояния между первой и последней структурой и длиной траектории:

$$l = \frac{|A_1A_n|}{|A_1A_2|+...+|A_{n-1}A_n|}.$$

Если данный показатель равен единице, то все точки-структуры лежат на одной прямой, т.е. наблюдается линейная динамика. Чем ближе его значение к единице, тем ближе тенденция динамики к многомерному линейному тренду. Если показатель равен нулю, то начальная и конечная структуры совпадают, т.е. изменение в целом носит возвратный характер. В докризисный период коэффициент линейности был в 2,5 раза больше, чем в кризисный период, что подтверждает выводы, полученные при анализе среднего динамического отклонения;

- угловой коэффициент линейного тренда, построенного для значений показателя удаления b, характеризует среднегодовую скорость изменения расстояния между текущей и базовой структурами. В докризисный период средняя скорость удаления от базовой структуры была в два раза меньше, чем в кризисный период (табл.3).

Алгоритм построения графика динамики структуры расходов:

1. Рассчитать необходимые данные: сдвиг, угол отклонения, расстояние до базовой структуры (табл.4). Установить масштаб построения, в нашем случае один процент равен 1 см.

2. Отметить точками плоскости базовую структуру A_0 и начальную структуру A_1 так, чтобы расстояние между ними равнялось R_1.

3. Отложить от точки A_1 два отрезка A_1A_2 длиной d_{12} = 3,4 см так, чтобы угол между отрезками A_0A_1 и A_1A_2 составлял $\alpha_1 = 122^o$. Из двух отрезков выбрать тот, у которого расстояние от точки A_2 до базовой структуры ближе к R_1 = 4,7 см. Другой отрезок «стереть». Убедиться, что направление изменения структуры невелико (отклонение).

4. Отложить от точки A_2 два отрезка A_2A_3 длиной d_{23} = 2,3 см так, чтобы угол между отрезками A_1A_2 и A_2A_3 составлял $\alpha_2 = 153^o$. Из двух отрезков выбрать тот, у которого расстояние от точки A_3 до базовой структуры ближе к значению R_2 = 6,7 см. Убедиться, что направление изменения структуры меньше, чем на предыдущем этапе. И так далее.

На построенном графике без изменения отображаются две основные характеристики динамики: сдвиг и отклонение. Однако на нем искажается расстояние от текущей структуры до базовой структуры, что не позволяет достоверно оценить степень их различия и определить, удаляется ли она от базовой или приближается к ней с течением времени. Но поскольку предложенный алгоритм требует при построении графика выбирать из двух возможных направлений то, которое в наибольшей степени отвечает критерию удаления от базовой структуры, искажающий эффект ослабляется. Другая особенность метода состоит в том, что пересечение звеньев на графике нельзя трактовать как их пересечение в многомерном пространстве. Так, если на графике произошло совпадение двух структур, относящихся к разным годам, то неверно делать вывод об их равенстве как точек многомерного пространства. Этот недостаток является следствием отображение многомерного пространства в плоскость, которое неизбежно вносит искажающие эффекты.

При построении графической схемы динамики структуры расходов федерального бюджета РФ возникает ряд специфических проблем.

Во-первых, статистические данные о расходах федерального бюджета до 2004 г. несовместимы с данными бюджетов последующих годов. Дело в том, что с 2005 г. изменилась классификация статей расходов, что не позволяет получить общую картину динамики структуры расходов за длительный период времени, например, 2002-2012 гг. Если в бюджете 2004 г. выделены 26 основных статей расходов, то в бюджете 2005 г. их всего 11. Так, в старой классификации имеются статьи расходов, которых нет в новой классификации: «Судебная власть», «Международная деятельность», «Промышленность, энергетика, строительство», «Средства массовой информации», «Обслуживание государственного и муниципального долга», «Исследование и использование космического пространства» и др. Большинство старых статей вошли в агрегированные статьи новой классификации: «Судебная власть» – в «Национальная безопасность и

правоохранительная деятельность», «Международная деятельность» – в «Общегосударственные вопросы», «Промышленность, энергетика, строительство» – в «Национальная экономика» и т.д. Но вместе с тем остаются спорные вопросы, которые требуют специальных исследований и без основательного решения которых невозможно получить достоверные данные об анализе динамики структуры бюджетных расходов РФ. Поэтому в настоящей работе нет анализа динамики структуры расходов бюджета РФ за период 2001-2012 гг., рассмотренный при анализе бюджетных расходов в США.

Таблица 4. Построение графика динамики структуры бюджетных расходов США

Показатель	Периоды									
	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9	9/10	10/11
Сдвиг (*d*), см	3,4	2,3	1,8	1,2	1,9	2,7	2,1	8,5	10,8	2,6
Угол отклонения (α), град.	122	153	110	105	111	52	56	118	29	32
Расстояние до базы (*R*), см	4,7	6,7	7,5	7,6	6,6	7,2	7,4	11,1	11,3	10,0
Характер изменения	О	О	О	П	О	Р	Р	О	Р	Р

Вторая проблема связана с неполнотой статистических отчетов Росстата, которые содержат лишь агрегированные данные о фактических расходах федерального бюджета по статье «Социально-культурные мероприятия». Это не позволяет рассчитать удельные веса фактических расходов раздельно по статьям «Образование», «Культура, кинематография, средства массовой информации», «Здравоохранение, физическая культура и спорт», «Социальная политика» (подробные данные приведены лишь для консолидированного бюджета). Также Росстат не публикует данные о фактических расходах федерального бюджета по статье «Жилищно-коммунальное хозяйство» [5].

Структура расходов федерального бюджета РФ с 2005 г. характеризуется одиннадцатью основными статьями: общегосударственные вопросы, национальная оборона, национальная безопасность и правоохранительная деятельность, национальная экономика, жилищно-коммунальное хозяйство, образование; культура, кинематография, средства массовой информации; здравоохранение, физическая культура и спорт, социальная политика, межбюджетные трансферты. Основные параметры динамики структуры бюджетных расходов РФ составили в период 2006-2007 гг.: сдвиг структуры – 9,0, коэффициент динамического отклонения – 0,9 (разворот), в период 2007-2008 гг.: 2,6 и -0,53 (отклонение), в период 2008-2009 гг.: 7,9 и 0,28 (поворот), в период 2009-2010 гг.: 4,8 и 0,22 (поворот). Усредненные значения за весь период 2006-2010 гг. составили: сдвиг структуры – 6,1; коэффициент динамического отклонения – 0,22 (поворот). Таким образом, в кризисный период среднее значение коэффициента динамического отклонения в России было выше, чем в США (0,22>0), причем в обеих странах динамика структуры расходов характеризовалась как «поворот».

Сравнение динамики структуры расходов бюджета между разными странами затруднено по двум причинам. Во-первых, в разных странах принята различная классификация статей бюджетных расходов. Сравнение статей бюджетных расходов в России и США выявляет ряд таких несоответствий. Так, статьи расходов российского бюджета «Национальная экономика», «Жилищно-коммунальное хозяйство», «Охрана окружающей среды», «Культура, кинематография, средства массовой информации», «Межбюджетные трансферты» не имеют аналогов в классификации расходов американского бюджета. Наоборот, статьи расходов американского бюджета «Пенсии», «Социальное обеспечение» и «Процентные платежи» не имеют аналогов в современной российской классификации. Так, в российском бюджете выплаты по государственному долгу отнесены к «Общегосударственным вопросам», а «Транспорт» является подстатьей статьи «Национальная экономика».

Вторая проблема следует из первой и заключается в том, что структуры бюджетных расходов разных странах имеют различную размерность, что порождает технические сложности при сравнении их динамики. Если угловые показатели инвариантны относительно размерности пространства, то для сравнения длин (линейных показателей) необходимо применять процедуру корректировки их значений. При разработке алгоритма корректировки следует учитывать, что структуры максимальной длины (состоящие из одной единицы и нулей) в различных пространствах имеют равную длину, равную единице. Структура минимальной длины характеризуется абсолютно равномерным распределением расходов, ее длина равна $1/\sqrt{n}$, где n – размерность пространства, или число статей бюджетных расходов. Поскольку на практике структура бюджетных расходов ближе к равномерной, чем к абсолютно неравномерной, корректировочный коэффициент для перевода линейных параметров динамики определяют как отношение минимальных длин структур, он равен $\beta = \sqrt{m/n}$, где m – размерность второго пространства. Тогда расстояние d во втором пространстве отвечает расстоянию βd в первом пространстве.

Для сравнения динамики структур бюджетных расходов в РФ и США рассчитаем поправочный коэффициент. Он равен $\sqrt{11/10} = 1{,}05$, поскольку размерность пространства структуры расходов в РФ равна 11, в США – 10. Тогда после корректировки среднее значение сдвига структуры в РФ за 2006-2010 гг. составит $6{,}1 \times 1{,}05 = 6{,}4$, что больше, чем для кризисного периода в США (5,35). Как видно, при близких значениях размерности поправочный коэффициент близок к единице и не играет существенной роли при межстрановых сравнениях. Но при больших различиях в размерности пространства необходима корректировка линейных параметров динамики.

ОТВЕТЫ К ЗАДАЧАМ ГЛАВЫ 1

1. Средние величины и вариация

1. а) 41,42 лет; б) 45,48 лет; в) 35,52 лет; г) 48,68 лет.
2. а) 7; б) 2; в) 6,5; г) 2,55; д) 0,51.
3. а) 417; б) 20,4 лет; в) 0,84; г) 0,04.
4. а) -0,82; б) -0,82.
5. а) 180; б) 130 000; в) 1,01.

2. Группировка

1. а) 0,54; б) 0,5; в) 0,53.
2. а) 16,67; б) 2,93; в) 13,64; г) 91%.
3. Антон – 1,12, Семен – 0,894.
4. а) 0,074 и 0,031; б) 0.
5. а) 0,0121 и 46504; б) Литва – 1,298; Польша – 1,293; Швеция – 1,461.

3. Выборка

1. а) 5; 3,5; б) 5; 4; 5,33; 5,67; 5; в) 4,67; 0,67; 4,22; 2,89; 3,11; г) 25, 50, 75%.
2. а) 1,16 и 0,55; б) 13,7 и 18,15; в) 0,46 и (13,24; 14,16).
3. а) 10-30%; б) 4 и 2000.
4. а) 25; б) 24.
5. 65.
6. а) 0,0955 и 0,0975; б) 1,96 и 2,064; в) (4,2; 4,6).
7. 44,4–88,8%.
8. а) 0,132; б) (3,87; 4,73).
9. а) (3,1; 4,9); б) да; в) да.
10. а) 6,24; 6,56; б) -1,656; - 0,875; в) 0,902; 0,62; г) 5, 10, 11, 6, 2; д) 5,65; е) 2.

4. Корреляция

1. а) -0,676; б) 0,09; в) 0,83.
2. а) -0,25; б) 0,25; в) 1.
3. а) -0,4; б) -0,2; в) 0,8.
5. 1.
6. а) 65% и 2,88; б) 41% и 0,13; в) 14% и 4,93.
7. а) 4,5; 1 и 2; б) -2 и 1,5; в) (0,232; 0,623; 0,145) и 0,014.
8. 3) Да; 6) Нет; 9) Да; 12) Да; 15) Нет. 18) Нет; 21) Да. 24) Нет. 26) Да.

5. Парная регрессия

1. «В».
2. -2.
3. (3,6).
4. а) параболическая, б) линейная; в) гиперболическая.

5. а) 21,6-1,6x; б) 5+0,67x; в) -1+0,38x.
6. (30, 100, 354).
7. а) $4{,}86 - 3{,}84x + 0{,}96x^2$; б) $1 - 2{,}5x + 1{,}5x^2$.
8. $-24{,}738 + 2{,}132x - 0{,}028x^2$.
9. а) 2,16; б) 0,24; в) 0,09.
10. (4,8; 9,67).
11. (12,98; 2,67).
12. $2+0{,}928\ln x$.
13. 3) Да; 6) Нет; 9) Нет; 12) Нет; 15) Да; 18) Да.

6. Взаимосвязь нескольких признаков

1. 0,5.
2. а) 8453 руб.; б) 13,2%; в) 20200 руб.
3. 3) Нет; 6) Нет; 9) Да; 12) Да; 15) Нет; 18) Да.

7. Динамика: основные понятия

1. 3) Нет; 6) Да; 9) Да; 12) Нет; 15) Нет; 18) Нет; 21) Да.

8. Динамика: тренд

1. а) $19 - 4t - 2t^2$; б) убыток 11 млн. руб.
2. а) $c = 0{,}083(\overline{yt^2} - 4\bar{y})$; б) $45 + 4t - t^2$; в) 13 млн. руб.
3. а) 10,15,20,25,30.
4. $7{,}74\exp(-0{,}69\times t)$.
5. $1/(1{,}34e^{-0{,}91t} + 1)$.
6. $4{,}51T^{0{,}56}$.
7. 3) Да; 6) Нет; 9) Да; 12) Да. 15) Нет. 18) Нет.

9. Динамика: колеблемость

1. а) 40+1,7t; б) в 2007 г. – дно, в) 0,5.
2. а) -0,12; б) 3-й год - дно (-1,4), 4-й год - пик (0,5).
3. а) $11{,}4-1{,}2t$; $24{,}8-1{,}5t$; б) 0,82; в) дополн.; г) 0,5; д) яблоки и груши осенью.
4. а) 5,4 и 5,9; б) 0,741 и 1,356; в) 0,778 и 1,33; г) 4,2 и 7,85; д) 17,2; е) 88%.

10. Индексы

1. а) 0,19; б) 22; в) 52,4%; г) 5,2.
2. 265,1 и 603, 4.
3. 45,2 и 165,5.
4. а) 0,878; б) 0,815.
5. 0,824.

11.Стохастические модели

1. (18, 68, 14), С
2.а) 9, 45, 24; б) -22; в) 27%, 0%, 72%.