МАТЕМАТИЧЕСКИЕ МЕТОДЫ

в психологии

Барнаул 2001

Оглавление

1.1 Введение.

В настоящее время математика стала языком, на котором говорят между собой представители различных научных дисциплин, а также удобным инструментом описания тех или иных явлений в различных отраслях человеческой деятельности. Несомненным достоинством этого языка является то, что он не терпит недомолвок и двусмысленностей. Поэтому, чтобы изъясняться на нем, исследователю приходится еще раз переосмыслить свои результаты, придать им стройность и завершенную форму.

Область применения математики постепенно расширяется. Все большее количество дисциплин превращается в поставщиков новых интересных задач для математики. Психология присоединилась к числу этих дисциплин в первых рядах - в начале XX века. Она не только использовала результаты наиболее бурно развивающейся отрасли математики - математической статистики,- но и сама способствовала возникновению ее новых разделов, в первую очередь таких, как факторный и дискриминантный анализ.

В настоящем курсе рассмотрены только самые первые ступени длинной и крутой лестницы, которую нужно преодолеть на пути к уверенному применению математических методов. За подробными сведениями, конкретными алгоритмами и другими методами исследования результатов эксперимента отсылаю читателя к литературе, список которой приведен в конце.

1.2 Что нужно знать, чтобы успешно понимать все, написанное ниже?

В этом разделе приведен список сведений из курса высшей математики, которые необходимо освежить в памяти для дальнейшего чтения. Основные определения и факты здесь также приводятся.

1.2.1 Теория матриц.

Матрица $n \times m$ - прямоугольная таблица с n строками и m столбцами, обозначаемая прописной буквой. Мы будем писать $[A]_{n,m}$ если хотим подчеркнуть, что матрица имеет именно такой размер. Если $n = m$, то матрица называется квадратной, а число n - ее порядком. Элементы матрицы A обозначаем строчными буквами с индексами, указывающими положение этого элемента в таблице. Через A^t условимся обозначать транспонированную матрицу, т.е. $[A^t]_{m,n}$, и для любых индексов i, j справедливо $a^t_{i,j} = a_{j,i}$. Матрица A называется симметричной, если $A = A^t$. В частности, для симметричной матрицы $n = m$. Элементы квадратной матрицы $a_{i,i}$ называются элементами главной диагонали.

Если $n = 1$, то матрицу $[A]_{1,n}$ называют вектором - строкой, если же $m = 1$, то $[A]_{m,1}$ - вектором-столбцом. Числа m, n в этом случае называют размерностями векторов. Векторы условимся обозначать строчными буквами со стрелкой вверху, понимая под $\vec{a}$ вектор-столбец. Тогда вектор-строка запишется как $\vec{a}^t$.

Для квадратных матриц A порядка n введем понятие определителя $\mid A \mid$. В частности, для $n = 2$,

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{1,1} & a_{2,2} \end{vmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$$

а для определителя Δ_{n+1} порядка $n+1$ справедливо

$$\Delta_{n+1} = \sum_{j=1}^{n+1} (-1)^{j+1} a_{1,j} \Delta_n^{1,j} ,$$

где $\Delta_n^{1,j}$ - определитель порядка n, полученный из Δ_{n+1} вычеркиванием первой строки и j-го столбца (разложение по первой строке).

Для умножения матрицы (в частности, вектора) на число необходимо каждый элемент матрицы умножить на это число. Определяются также сумма (поэлементная) матриц и их произведение по следующему правилу. Пусть $[A]_{n,m}$, $[B]_{m,l}$, $C = AB$. Тогда $[C]_{n,l}$, причем

$$c_{i,j} = \sum_{k=1}^{m} a_{i,k} b_{k,j}, \quad i = 1, ..., n, \ j = 1, ...l.$$

Например, если $\vec{a}$ вектор размерности n, то $A\vec{a}$ - вектор размерности m. Можно проверить, что если A, B - квадратные матрицы, то

$$\mid AB \mid = \mid A \mid \mid B \mid, \quad \mid A^t \mid = \mid A \mid .$$

Матрица $[I]_{n,n}$ называется единичной порядка n, если по ее главной диагонали стоят единицы, а остальные элементы равны нулю. Очевидно, что для любой квадратной матрицы $[A]_{n,n}$ справедливо $AI = IA = A$. Матрица A^{-1} называется обратной для матрицы A, если

$$AA^{-1} = A^{-1}A = I.$$

Пусть A - квадратная матрица. Если для вектора $\vec{a}$ найдется такое число λ, что $A\vec{a} = \lambda\vec{a}$, то вектор $\vec{a}$ называется собственным вектором матрицы, отвечающим собственному числу λ. Оказывается, собственные числа матрицы можно найти, решив уравнение

$$\mid A - \lambda I \mid = 0.$$

Собственный вектор, отвечающий данному собственному числу, определен неоднозначно с точностью до умножение на любое число α, так как

$$A(\alpha\vec{a}) = \alpha(A\vec{a}) = \lambda(\alpha\vec{a}),$$

т.е. $\alpha\vec{a}$ также является таким собственным вектором. Поэтому можно поставить задачу найти собственный вектор, отвечающий заданному собственному числу, и имеющий единичную длину. Для решения этой задачи достаточно найти любой собственный вектор $\vec{a}$ и поделить его на

$$\|\vec{a}\| \;=\; \sqrt{\sum_{j=1}^{n} a_j^2} \;-$$

длину вектора $\vec{a}$.

Известно, что у симметричной матрицы порядка n обязательно n собственных чисел.

1.2.2 Геометрическая интерпретация.

Пары и тройки чисел (двумерные и трехмерные векторы) можно представлять себе точками плоскости или, соответственно, трехмерного пространства. Векторы большей размерности обычно отождествляют с точками пространства с числом измерений, равным размерности вектора. Поскольку геометрическая интуиция, связанная с числом измерений, большим 3, отсутствует, то чаще всего этот способ - единственная возможность работать с такими пространствами.

Умножая вектор на матрицу, мы вновь получаем вектор, а значит квадратная $n \times n$ матрица может рассматриваться как преобразование n-мерного пространства, а матрица $n \times m$ - как преобразование m-мерного пространства в n-мерное.

Известно, что в случае, когда матрица A ортогональна, т.е. $A^{-1} = A^t$ и n равно 2 или 3, то умножение на A соответствует повороту. Поэтому и в многомерном случае умножение на ортогональную матрицу можно интерпретировать как поворот. Вообще, умножение на любую матрицу можно рассматривать как поворот и растяжение (неодинаковое по разным направлениям). Из определения собственного вектора следует, что в направлении собственного вектора (и только в таких направлениях) действие матрицы является чистым растяжением.

В многомерном случае угол между векторами $\vec{a}$, $\vec{b}$ определяется как

$$\varphi \;=\; \arccos \frac{<\vec{a},\, \vec{b}>}{\|\vec{a}\|\, \|\vec{b}\|},$$

где $<\vec{a},\, \vec{b}> = \sum_k a_k b_k$ - скалярное произведение векторов. В частности, $\vec{a}$, $\vec{b}$ перпендикулярны, если $<\vec{a},\, \vec{b}> = 0$. Выпишем также формулу

$$<\vec{a},\, \vec{b}> = \vec{a}^t\vec{b},$$

выполненную по определению произведения матриц и в силу соглашения рассматривать только векторы-столбцы.

Заметим, наконец, что $C = \vec{a}\vec{b}^t$ - $n \times n$ - матрица с элементами $c_{i,j} = a_i b_j, \quad i, j = 1, ..., n.$

1.2.3 Теория вероятностей.

В теории вероятностей рассматриваются случайные события и случайные величины. Дискретная случайная величина X может характеризоваться своим рядом распределения:

X	x_1	...	x_n
	p_1	...	p_n

Здесь $x_1, ..., x_n$ - значения случайной величины, $p_1, ..., p_n$ - вероятности этих значений. Можно задавать ее также функцией распределения

$$F(x) \;=\; P(X < t) \;=\; \sum_{j:x_j<t} p_j \;.$$

Математическим ожиданием такой случайной величины (или ее средним значением) называется число

$$\mathbf{M}X \;=\; \sum_{k=1}^{n} x_k p_k \;.$$

Дисперсия случайной величины - это числовая характеристика разброса ее значений вокруг среднего. По определению,

$$\mathbf{D}X \;=\; \mathbf{M}(X - \mathbf{M}X)^2 \;.$$

Известно, что всегда

$$\mathbf{M}(X+Y) \;=\; \mathbf{M}X + \mathbf{M}Y, \quad \mathbf{M}(\alpha X) \;=\; \alpha \mathbf{M}X, \quad \mathbf{D}(\alpha X) \;=\; \alpha^2 \mathbf{D}X,$$

и если X, Y независимы, то

$$\mathbf{M}XY \;=\; \mathbf{M}X\;\mathbf{M}Y, \quad \mathbf{D}(X+Y) \;=\; \mathbf{D}X \;+\; \mathbf{D}Y.$$

Число $\sigma \;=\; \sqrt{\mathbf{D}X}$ называют средним квадратическим отклонением X. Известно, что в интервале $(\mathbf{M}X - 3\sigma,\; \mathbf{M}X + 3\sigma)$ лежит не менее $8/9$ всех значений X (неравенство трех сигм).

Величина

$$\rho(X,Y) \;=\; \frac{\mathbf{cov}(X,Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}} \;,$$

где $\mathbf{cov}(X,Y) \;=\; \mathbf{M}XY \;-\; \mathbf{M}X\mathbf{M}Y \;=\; \mathbf{M}(X - \mathbf{M}X)(Y - \mathbf{M}Y)$ - ковариация между X и Y, называется коэффициентом корреляции между этими величинами. Значения ρ лежат между -1 и 1. Крайние его значения соответствуют линейной зависимости между X и Y. Если $\rho(X,Y) \;=\; 0$, то величины называют некоррелированными. В частности, независимые случайные величины являются некоррелированными. Для ρ выделяют обычно три зоны: $\mid \rho(X,Y) \mid \;< 1/3$ соответствует слабой связи (зависимости) между X и Y, $1/3 \;\leq \mid \rho(X,Y) \mid \;< \;2/3$ умеренной связи и $\mid \rho(X,Y) \mid \;\geq \;2/3$ сильной связи. Если $\rho > 0$ говорят о положительной, иначе об отрицательной связи.

Если значения случайной величины целиком заполняют какой-либо отрезок, то она называется абсолютно непрерывной, и ее характеризуют плотностью распределения. Наиболее часто в приложениях встречается так называемое нормальное распределение с параметрами a и σ^2, причем первый параметр представляет из себя математическое ожидание, а второй дисперсию соответствующей случайной величины. Его плотность имеет вид

$$p(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\,\mathbf{exp}\left\{-\frac{(x-a)^2}{2\sigma^2}\right\}.$$

При $a=0,\ \sigma=1$ такое распределение называют стандартным нормальным. Выпишем плотность этого распределения.

$$\varphi(x) \;=\; \frac{1}{\sqrt{2\pi}}\exp(-x^2/2).$$

Функция

$$\Phi(t) \;=\; \int\limits_{-\infty}^{t} \varphi(x)\;dx$$

называется функцией стандартного нормального распределения. Известно, что если X имеет нормальное распределение с параметрами a, σ^2, то $Y = (X-a)/\sigma$ имеет стандартное нормальное распределение, т.е., например,

$$P(X<t) \;=\; P\left(\frac{X-a}{\sigma} \;<\; \frac{t-a}{\sigma}\right) \;=\; \Phi\left(\frac{t-a}{\sigma}\right).$$

Нам будет нужен также многомерный случай. Случайным n-мерным вектором называется вектор $\vec{x}$, все координаты которого $x_1, ..., x_n$ - случайные величины. Через $\mathbf{M}\vec{x}$ обозначим вектор, каждая из координат которого равна математическому ожиданию соответствующей координаты $\vec{x}$. Аналогом понятия дисперсии служит ковариационная матрица C, причем

$$c_{i,j} \;=\; \mathbf{cov}(x_i, x_j), \quad i,j = 1, ..., n.$$

В частности, $c_{j,j} \;=\; \mathbf{D}x_j$. С учетом матричных соотношений, рассмотренных в конце первого подпункта,

$$C \;=\; \mathbf{M}(\vec{x} \;-\mathbf{M}\vec{x})(\vec{x} \;-\; \mathbf{M}\vec{x})^t.$$

1.3 Данные наблюдений и их виды. Понятие выборки.

Все наши выводы так или иначе основываются на наблюдениях. Результаты наблюдений для выявления статистических закономерностей, как правило, формируются в повторных экспериментах. Простейший и наиболее распространенный вариант - измерение некоторых числовых характеристик. В

этом случае говорят, что мы имеем в нашем распоряжении выборку объема n, т.е. набор n независимых наблюдений над какой-то случайной величиной. Одно наблюдение называют элементом выборки или выборочным значением. Например, в качестве наблюдаемой случайной величины могут выступать результаты какого-либо тестирования (в баллах) или процент верно опознанных изображений (фонем). Близко к этому расположены данные оценки близости в шкалах "максимально похожи - максимально различны", для которых также нетрудно установить числовые значения.

В практике социальных дисциплин встречаются и нечисловые данные. Отметим, что полностью нерегулярные данные представить себе довольно трудно, и на практике они не встречаются. Всегда имеются некоторые группы (категории) в которые можно отнести наблюдаемые характеристики. Такими группами (категориями) могут служить, например, темпераменты испытуемых (4 категории), данные о географическом происхождении наблюдений, об их времени и т.п. (см. пример с посетителями кафе ниже). В этом случае мы говорим о категоризованных данных.

Интересным примером является также изучение результатов ранжирования (расположения в порядке убывания значимости) ряда факторов независимыми экспертами. Рассмотрим пример. Пяти студентам, пользующимся общественным транспортом, предложили пронумеровать в порядке убывания значимости следующие факторы: Ч - частота следования транспорта, З - степень его заполненности пассажирами, О - оборудование салона (комфортность сидений, кондиционер и т.п.), Д - исправность дверей и окон, К - настроение и доброжелательность кондуктора, С - освещение салона, Ц - стоимость проезда. Самому важному с точки зрения опрашиваемого фактору он присваивает номер 1, следующему по важности - 2 и т.д. Если студент не может или не хочет упорядочивать несколько факторов (они для него равноценны), то он присваивает им равные номера (ранги). При этом сумма всех присвоенных рангов должна быть равна $1+2+...+7 = 28$. Например, если студент уверен, что самые важные факторы - Ц и Ч, но он не может их различить, то каждому из них присваивается ранг $(1+2)/2 = 1{,}5$. Данные соответствующего опроса приведены в таблице. Такая таблица называется матрицей экспертных оценок.

Встречаются также и данные, имеющие смешанный характер. Для примера рассмотрим следующие результаты наблюдений за 12 посетителями кафе. Ниже x_1 - сумма, истраченная посетителем, x_2 - время в минутах, проведенное в кафе, x_3, x_4, x_5 - закуска, основное блюдо и напиток, выбранные посетителем. Здесь x_1, x_2 - числовые переменные, x_3, x_4, x_5 - нечисловые категоризованные, x_3 имеет 3 градации, x_4, x_5 по 4 градации.

Основные методы обработки данных разработаны для случая числовых данных (выборок). Поэтому важное значение приобретают методы придания нечисловым данным числовых значений (оцифровка). Такие методы мы обсудим ниже. Тем не менее, в силу общего происхождения (из наблюдений), условимся все наши данные далее называть выборочными данными.

Данные ранжирования
пятью экспертами семи факторов по убыванию значимости

эксперты факторы	1	2	3	4	5
Ч	1	2	2	1	3
З	3	1	2	5	2
О	5	3	6,5	5	5
Д	7	4	6,5	5	4
К	6	7	5	5	7
С	4	6	2	5	6
Ц	2	5	4	2	1

Двенадцать посетителей кафе

Посетитель	x_1	x_2	x_3	x_4	x_5
1	100	63	1	4	1
2	85	63	1	2	1
3	65	45	1	2	2
4	65	45	2	2	2
5	110	95	2	3	3
6	120	95	2	3	3
7	125	135	2	3	4
8	170	95	2	1	3
9	180	135	2	1	4
10	95	63	3	4	1
11	105	95	3	3	3
12	175	135	3	4	4

1.4 Первичная обработка и группировка данных. Грубые ошибки наблюдений.

Если выборка X имеет небольшой объем n, то мы можем непосредственно приступить к расчету выборочных характеристик наблюдаемой величины. Число

$$\bar{X} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

называется выборочным средним или выборочным математическим ожиданием, а

$$S^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2 \;-\; \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \;-$$

выборочной дисперсией. Эти характеристики не следует путать с математическим ожиданием и дисперсией наблюдаемой случайной величины x. $\bar{X}$ и S^2 - это оценки $\mathbf{M}x$, $\mathbf{D}x$ по результатам наблюдений и равны последним (теоретическим) характеристикам лишь приближенно. В частности, значение S^2 в основном оказывается меньше теоретической дисперсии (имеется систематическая ошибка). Чтобы улучшить оценку, при малых n имеет смысл применять т.н. исправленную (правильное название - несмещенную) оценку дисперсии:

$$S_0^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{X})^2.$$

В прикладных исследованиях применяются также такие характеристики, как асимметрия и эксцесс:

$$\mathbf{As}X \;=\; \frac{1}{nS^3}\sum_{i=1}^{n}(x_i-\bar{X})^3, \quad \mathbf{Ex}X \;=\; \frac{1}{nS^4}\sum_{i=1}^{n}(x_i-\bar{X})^4 \;-\; 3.$$

Здесь S - корень квадратный из выборочной дисперсии S^2. Эти величины характеризуют соответственно смещение пика плотности влево (асимметрия положительна) или вправо (асимметрия отрицательна) относительно середины интервала, и остроту (эксцесс положителен) или пологость (эксцесс отрицателен) этой плотности. Случай нулевого эксцесса соответствует нормальной кривой.

Величины

$$X_{(1)} \;=\; \min\{x_1,...,x_n\}, \quad X_{(n)} \;=\; \max\{x_1,...,x_n\}$$

характеризуют размах выборки

$$T \;=\; X_{(n)} \;-\; X_{(1)}.$$

Довольно часто бывает так, что выборка содержит повторяющиеся значения или имеется много близких по величине элементов. В этом случае всю

имеющуюся в выборке информацию удобно хранить в сгруппированном виде:

Значение	x_1	x_2	...	x_k
Количество таких значений (повторности)	n_1	n_2	...	n_k

Очевидно, здесь $\sum_{j=1}^{k} n_j = n$. Теперь выписанные выше формулы примут вид

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \qquad S^2 = \tfrac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{X})^2,$$

$$\mathbf{As}X = \frac{1}{nS^3}\sum_{i=1}^{k} n_i (x_i - \bar{X})^3, \quad \mathbf{Ex}X = \tfrac{1}{nS^4}\sum_{i=1}^{k} n_i (x_i - \bar{X})^4 - 3$$

и для $k \ll n$ вычисления существенно упрощаются. Поэтому возникает желание "образовать"повторяющиеся элементы даже если совпадающих элементов в выборке нет. Этот процесс носит название группировки данных.

Далее даются некоторые рекомендации по группировке несгруппированных числовых данных. Группировку можно производить и иначе, но те требования, которые обязательно должны быть выполнены при проведении любой группировки будут отдельно оговорены.

- Определим размах выборки и первоначальное число групп - интервалов. Если это число заранее никак не было определено, рекомендуется пользоваться формулой Стерджеса

  $$r = [\log_2 n] + 1,$$

  где [.] - целая часть, т.е. наибольшее целое, не превосходящее данное. При этом, очевидно, можно использовать вместо формулы Стерджеса следующую таблицу:

n	8-15	16-31	32-63	64-127	128-255	256-511	512-1023
r	4	5	6	7	8	9	10

  Еще раз отметим, что этот выбор r достаточно произволен, и впоследствии число групп может меняться.

- Определим нижнюю границу группировки. Это может быть либо $-\infty$ либо 0, либо $X_{(1)} - \varepsilon$, где ε - достаточно малое число. Наличие здесь его обуславливается одним из двух основных принципов группировки, обязательным для соблюдения при любом способе:

Границы групп не должны совпадать с выборочными значениями

После выбора нижней границы z_0 строим остальные по формулам

$$z_1 = X_{(1)} + T/r,$$
$$z_{j+1} = z_j + T/r,\ 1 \le j \le r-2,$$
$$z_r = X_{(n)} + \varepsilon.$$

Вместо последней формулы можно использовать $z_r = +\infty$. Если некоторые из построенных границ попали на выборочные значения - двигаем границы (на ε) влево или вправо до тех пор, пока это не будет устранено. Итак, построены группы $\Delta_j = [z_{j-1},\ z_j],\ j = 1, ..., r$.

- Вычислим n_j - количества элементов выборки, попавших в $\Delta_j,\ j = 1, ..., r$. Тут появляется второй основной принцип группировки:

Для всех групп $3 \le n_j \le 19$.

Если хотя бы в одной из групп это условие нарушено - необходимо передвинуть границы интервалов или объединить "слишком пустые", или разбить "слишком наполненные"на более мелкие интервалы (совсем не обязательно одинаковой длины). При этом r может измениться.

Выделенные условия являются основными, и можно проводить группировку "на глазок", ориентируясь лишь на них. После того, как мы добились их выполнения, группировка закончена, и мы заменяем нашу выборку таблицей

X	$\tilde{x}_1$	$\tilde{x}_2$	...	$\tilde{x}_r$
	n_1	n_2	...	n_r

Здесь $\tilde{x}_j$ - середина интервала $\Delta_j,\ j = 1, ..., r$.

По сгруппированной выборке можно определить моду, медиану, построить гистограмму и полигон распределения наблюдаемой величины. Мода - это наиболее часто встречающееся значение, т.е. то из $\tilde{x}_j$, для которого n_j максимально. Медианой называется то из выборочных значений, левее и правее которого расположено поровну элементов выборки. Гистограмма - это столбчатая диаграмма, у которой на интервале Δ_i значение равно $n_i/n,\ i = 1, ..., r$. Полигоном называется ломаная линия с узлами в точках с координатами $(\tilde{x}_i,\ n_i/n)$.

Рассмотрим числовой пример на группировку данных. В опыте по изучению амплитудно-частотной характеристики колебаний руки оператора получены следующие амплитудные характеристики установившихся колебаний в мм ($n = 100$).

64	72	60	67	63	65	60	75	51	80
65	62	73	62	71	63	55	56	64	61
65	69	69	65	68	58	62	52	68	72
66	62	67	60	68	60	60	58	57	60
64	59	64	65	60	63	59	60	58	62
63	55	61	45	46	64	72	70	70	63
63	41	62	60	69	71	58	60	64	70
73	52	59	54	64	65	70	65	58	52
56	55	60	54	59	71	63	55	55	58
66	62	82	54	74	58	55	62	75	62

Здесь $X_{(1)} = 41$, $X_{(n)} = 82$. Размах выборки $T = 41$. Первоначальное рекомендованное число интервалов $r = 7$. Длина типичного интервала $T/r \approx 5,86$. Результаты первичной группировки:

	интервал	n_i
1	40,90-46,76	3
2	46,76-52,61	4
3	52,61-58,47	19
4	58,47-64,33	39
5	64,33-70,19	22
6	70,19-76,04	11
7	76,04-82,10	2

Объединим первые два и последние два интервала. Получим грубую группировку:

Δ_i	40,90-52.61	52,61-58,47	58,47-64,33	64,33-70,19	70,19-82,10
$\tilde{x}_i$	46,76	55,54	61,40	67,26	76,04
n_i	7	19	39	22	13

Заметим, что в "переполненных"интервалах значения распределены следующим образом:

59-61	62	63-64	65	66-70
16	9	14	7	15

Таким образом, можно разбить интервал (58,41, 70,19) на 5 интервалов. Получим

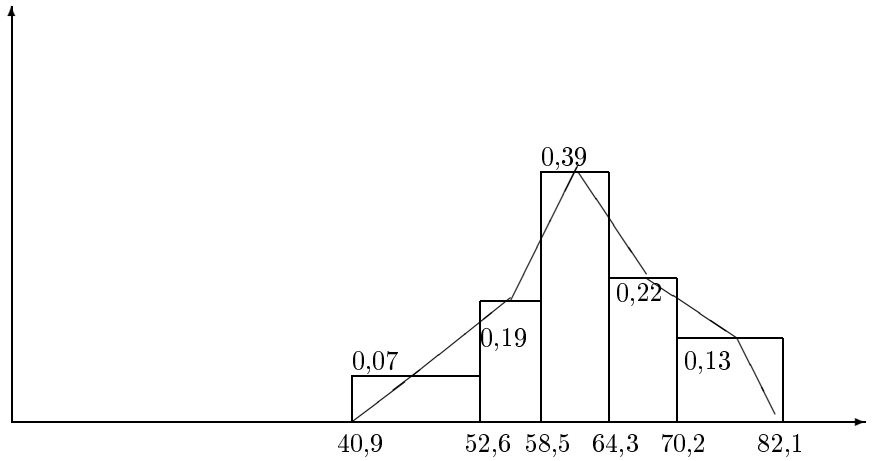
i	Δ_i	$\tilde{x}_i$	n_i
1	40,90-52,61	46,76	7
2	52,61-58,47	55,54	19
3	58,47-61,50	59,98	16
4	61,50-62,50	62,00	9
5	62,50-64,50	63,50	14
6	64,50-65,50	64,00	7
7	65,50-70,19	67,85	15
8	70,19-82,10	76,04	13

Окончательно получено 8 интервалов. Чтобы проследить влияние разбиения на выборочные характеристики, приведем значения четырех средних: по полной выборке $\bar{X} = 62,52$; с использованием первичной группировки - 62,22; после усечения краев (грубая группировка) - 62,45; по окончательной группировке - 62,44. Мы видим, что окончательная группировка незначительно сказалась на среднем, а число интервалов увеличилось. Поэтому вычисления дисперсии, асимметрии и эксцесса проведем с использованием грубой группировки. Заполним таблицу

i	1	2	3	4	5
n_i	7	19	39	22	13
$\tilde{x}_i$	46,76	55,54	61,40	67,26	76,04
$\tilde{x}_i - \bar{X}$	-15,46	-6,68	-0,82	5,04	13,82
$(\tilde{x}_i - \bar{X})^2$	239,01	44,62	0,67	25,40	190,99
$(\tilde{x}_i - \bar{X})^3$	-3695,12	-298,08	-0,55	128,02	2639,51
$(\tilde{x}_i - \bar{X})^4$	57126,54	1191,16	0,45	645,24	36478,09

Вычисления дают $S^2 = 55,90$, $S = \sqrt{S^2} = 7,48$, **As**X = 0,13 (пик слегка влево), **Ex**X = 2,91 - 3 = -0,09 (пик скорее острый, чем пологий).

Гистограмма и полигон



Бывает, что среди наблюдаемых значений присутствуют такие, которые сильно отличаются от остальных. Как правило, это крайние по величине наблюдения. Эти наблюдения (если они действительно резко выбиваются из общего ряда наблюдений) называют грубыми ошибками наблюдения. Их желательно исключить из обрабатываемой выборки. Существует много способов (критериев) определения, является ли данное наблюдение грубой ошибкой. Эти способы иногда называют методами цензурирования. Один из таких методов - исключение тех значений, которые оказались в единственном числе при осуществлении группировки выборки, да еще отделены от остальных пустыми интервалами. Другой состоит в том, что отбрасыванию подлежит то значение, которое существенно изменяет $\bar{X}$ (см. ниже).

Мы приводим следующий критерий: рассчитывается

$$t \;=\; \frac{\max \mid x_i - \bar{X} \mid}{S}$$

и сравнивается со значением t_n, приводимым ниже в таблице. Если $t > t_n$, то выделяющееся значение нужно отбросить.

n	t_n	n	t_n
5	1,972	30	3,291
10	2,616	35	3,364
15	2,905	40	3,424
20	3,079	45	3,474
25	3,200	50	3,518

Следует иметь ввиду, что для уверенного пользования этим критерием нужно, чтобы наблюдения имели нормальный (в смысле распределения) характер. Соответствующий критерий для проверки этого будет дан ниже.

1.5 Доверительные интервалы. Таблицы некоторых распределений.

Как уже отмечалось, $\bar{X}, S^2$ являются лишь оценками математического ожидания и дисперсии, причем их совпадение с теоретическими характеристиками практически исключено (имеет нулевую вероятность). Иногда бывает удобно указать интервал, внутрь которого теоретическая, недоступная непосредственному измерению характеристика попадает с достаточно большой, близкой к единице, вероятностью. Такой интервал называют доверительным. Более точно, если θ - неизвестный параметр, а ε - достаточно малое число, то интервал $[\theta^-, \theta^+]$ называется доверительным интервалом для θ уровня $1-\varepsilon$, если неравенство $\theta^- \le \theta \le \theta^+$ выполнено с вероятностью $1-\varepsilon$, или в $100(1-\varepsilon)$ процентах случаев.

Приведем без обоснования формулы для вычисления границ доверительных интервалов для математических ожиданий и дисперсий.

1.5.1 Построение доверительного интервала для математического ожидания, если дисперсия σ^2 заранее известна. Таблица стандартного нормального распределения.

Нужные границы рассчитываются по формулам

$$a^- \;=\; \bar{X} \;-\; \frac{\sigma t_\varepsilon}{\sqrt{n}}, \quad a^+ \;=\; \bar{X} \;+\; \frac{\sigma t_\varepsilon}{\sqrt{n}},$$

где t_ε ищется по таблице функции стандартного нормального распределения Φ из соотношения

$$\Phi(-t_\varepsilon) \;=\; \frac{\varepsilon}{2}.$$

Наиболее употребимые t_ε

$$\Phi(0,00) = 0,500 \quad \Phi(-1,28) = 0,100 \quad \Phi(-1,64) = 0,050$$
$$\Phi(-2,32) = 0,010 \quad \Phi(-2,58) = 0,005 \quad \Phi(-2,98) = 0,001$$

Ниже приводится таблица $\Phi(x)$, $x < 0$. Для положительных значений x применяем формулу $\Phi(x) = 1 - \Phi(-x)$,

Функция стандартного нормального распределения

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \mathbf{e}^{-\frac{t^2}{2}}\, dt$$

x	-3,00	-2,80	-2,60	-2,40	-2,20	-2,00	-1,80	-1,60
$\Phi(x)$	0,001	0,003	0,005	0,008	0,014	0,023	0,036	0.055
x	-1,40	-1,20	-1.00	-0,80	-0,60	-0,40	-0,20	-0,10
$\Phi(x)$	0,081	0,115	0,159	0,212	0,274	0,345	0,421	0,460

1.5.2 Построение доверительного интервала для математического ожидания, если дисперсия неизвестна. Распределение Стьюдента.

В этом случае дисперсия заменяется на корень квадратный из ее оценки S^2, и меняется характер распределения.

$$a^- = \bar{X} - \frac{S\tau_{\varepsilon,n}}{\sqrt{n-1}}, \quad a^+ = \bar{X} + \frac{S\tau_{\varepsilon,n}}{\sqrt{n-1}},$$

где $\tau_{\varepsilon,n}$ - двусторонняя критическая точка распределения Стьюдента с $n-1$ степенью свободы. Ее значения находятся из таблицы.

Двусторонние критические точки
(коэффициенты) Стьюдента $\tau_{\varepsilon,n}$

n	ε		n	ε	
	0,1	0,01		0,1	0,01
5	2,02	4,03	25	2,06	2,79
10	2,23	3,17	30	2,04	2,75
15	2,13	2,95	50	2,01	2,68
20	2,09	2,85	∞	1,64	2,57

1.5.3 Построение доверительного интервала для дисперсии. Таблицы распределения хи-квадрат.

Формулы для соответствующих границ имеют вид

$$(\sigma^2)^- = \frac{nS^2}{T^+_{\varepsilon,n}}, \quad (\sigma^2)^+ = \frac{nS^2}{T^-_{\varepsilon,n}},$$

где $T^+_{\varepsilon,n}$ - критическая точка распределения χ^2 с $n-1$ степенью свободы уровня $\frac{\varepsilon}{2}$, а $T^-_{\varepsilon,n}$ - такая же точка уровня $1-\frac{\varepsilon}{2}$. Эти значения берутся из таблицы.

Критические точки распределения χ^2

n	ε		n	ε	
	0,95	0,05		0.95	0,05
1	$0{,}39\times10^{-5}$	3,84	19	10,12	30,14
2	0,103	5,99	20	10,85	31,41
3	0,352	7,81	24	13,85	36,42
4	0,711	9,49	25	14,61	37,65
5	1,15	11,07	29	17,71	42,56
6	1,64	12,59	30	18,49	43,77
9	3,33	16,92	34	21,66	48,60
10	3,94	18,31	35	22,47	49,80
14	6,57	23,68	49	33,93	66,34
15	7,26	25,00	50	34,76	67,50

1.6 Проверка статистических гипотез - общие принципы.

Предположим, что нам нужно принять одно из двух взаимоисключающих решений (гипотез) H_0 или H_1 относительно наблюдаемой случайной величины. Если при этом мы руководствуемся имеющейся выборкой X, то говорим, что имеет место задача проверки статистических гипотез, а соответствующее решающее правило, указывающее, которую из гипотез надо принять, называют критерием для проверки H_0 против H_1. H_0 обычно выбирается так, что для ее принятия нужны гораздо менее убедительные аргументы, чем для того, чтобы ее отвергнуть. Тогда H_0 называют основной гипотезой, а H_1 - альтернативной гипотезой или альтернативой. Часто H_1 вообще не формулируется, имея ввиду, что это решение совпадает с отрицанием H_0.

В силу широкого спектра применения таких задач имеется обширная терминология, связанная с критериями. Для их определения введем обозначения.

	Статус объекта		
	"случай"(H_0)	"не случай"(H_1)	Всего
H_0 принимается	a	b	$a+b$
H_0 отвергается	e	f	$e+f$
Всего	$a+e$	$b+f$	n

Здесь a - число тех изученных объектов, для которых принимается H_0, и она действительно справедлива, b - число тех объектов, для которых мы приняли основную гипотезу, а она в действительности не верна и т.д. Определим некоторые характеристики критерия.

- Вероятность ошибки первого рода - $\frac{e}{a+e}$.
- Вероятность ошибки второго рода - $\frac{b}{b+f}$.
- Специфичность, или мощность критерия - $\frac{f}{b+f}$.
- Чувствительность - $\frac{a}{a+e}$.
- Частота случаев - $\frac{a}{n}$.
- Относительный риск - $\frac{a}{a+e} \ : \ \frac{e}{e+f}$.
- Доля ложноотрицательных - $\frac{e}{e+f}$.
- Доля ложноположительных - $\frac{b}{a+b}$.

Среди этих характеристик только три независимых, остальные могут быть вычислены через них. В прикладных работах чаще всего используется частота случаев, чувствительность и специфичность.

Среди всех критериев для проверки конкретных гипотез выделяется группа так называемых критериев согласия. В основе их действия лежит следующее простое соображение. Предположим, что гипотеза H_0 справедлива и посмотрим, согласуется ли это предположение с характером выборочных данных. Если некоторая, соответствующим образом подобранная, мера отклонения ожидаемых с точки зрения справедливости H_0 от наблюдаемого в эксперименте принимает небольшие (не превосходящие некоторого критического) значения, то основная гипотеза принимается, если большие - отвергается, и принимается альтернативная гипотеза. При этом в качестве критического берется то значение, которое при H_0 превосходится выбранной мерой с малой вероятностью (0,1, 0,01 и т.п.). Рассмотрим примеры некоторых из статистических критериев.

1.6.1 Проверка равенства средних значений двух выборок .

Пусть есть выборки X, Y объемов n, m соответственно. Предположим, что $n \leq m$. Основная гипотеза состоит в том, что $\mathbf{M}X = \mathbf{M}Y$. Альтернатива -

математические ожидания не совпадают (обычно не формулируется). Для проверки построим

$$u_i \;=\; x_i \;-\; \sqrt{\tfrac{n}{m}}\, y_i,\; i = 1, ..., n,$$

$$\bar{U} \;=\; \frac{1}{n}\sum_{i=1}^{n} u_i, \qquad S_U^2 \;=\; \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(u_i - \bar{U})^2, \qquad T \;=\; \dfrac{\bar{X} - \bar{Y}}{S_U}\sqrt{n-1}.$$

Если $\mid T \mid$ не превосходит двусторонней критической точки распределения Стьюдента с $n-1$ степенью свободы, то гипотезу H_0 можно принять.

Как уже упоминалось выше, этот критерий можно применять для исключения грубых ошибок наблюдения. Для этого в качестве выборки X возьмем Y с удаленным "подозрительным"значением ($n = m - 1$). Если различие $\mathbf{M}X$ и $\mathbf{M}Y$ оказалось существенным (принята альтернативная гипотеза), то удаленное значение следует признать грубой ошибкой. Для пользования этим критерием не требуется проверки равенства дисперсий, но нужна нормальность распределений X, Y.

1.6.2 Проверка значимости коэффициента корреляции ρ.

Рассмотрим две выборки X, Y одинакового объема n. Основная гипотеза состоит в том, что наблюдаемые распределения некоррелированы, т.е. $\rho(X, Y) \;=\; 0$. Это предположение близко к независимости X и Y и часто им подменяется. Это совершенно законно для случая, когда обе выборки имеют совместное нормальное распределение. Для проверки рассчитаем

$$T \;=\; \frac{R\sqrt{n-2}}{\sqrt{1-R^2}},$$

где R - выборочный коэффициент корреляции:

$$R \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2}}$$

и сравним T с двусторонней критической точкой распределения Стьюдента с $n-2$ степенями свободы. Если критический уровень не превзойден - коэффициент корреляции ρ незначимо отличается от 0.

1.6.3 Проверка равенства дисперсий.

Даны две выборки - X объема n и Y объема m. Основная гипотеза состоит в том, что дисперсии двух наблюдаемых случайных величин совпадают. Рассчитаем

$$f \;=\; \frac{(m-1)nS_X^2}{(n-1)mS_Y^2},$$

где $S_X^2,\ S_Y^2$ - выборочные дисперсии X, Y соответственно. Найдем левую f^- и правую f^+ критические точки распределения Фишера с $n-1,\ m-1$ степенями свободы по соответствующей таблице (имеется практически в любом учебнике по статистическим методам). Если выполнено неравенство $f^- \le f \le f^+$, то можно считать дисперсии равными.

Отметим, что все критерии, приведенные выше, корректно работают лишь для случая выборок из нормального распределения. Проверка гипотезы о виде распределения описана в следующем разделе.

1.7 Проверка гипотезы о виде распределения. Критерий χ^2.

Пусть основная гипотеза состоит в том, что наблюдаемая величина имеет конкретное, известное распределение (например, нормальное распределение с заранее заданными параметрами $a,\ \sigma^2$). Предположим, что значения нашей величины разбиты на группы $\Delta_1, ..., \Delta_r$ и $p_1, ..., p_r$ - вероятности попадания величины, имеющей требуемое распределение, в каждую из групп соответственно. Естественно, при этом предполагается, что $\sum_{i=1}^r p_i = 1$. Обозначим, как и раньше, через $n_i,\ i = 1, ..., r$ количества элементов выборки, попавших в каждую из групп. Статистикой хи-квадрат называется

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}.$$

Если рассчитанное значение χ^2 меньше, чем критическая точка распределения хи-квадрат с $r-1$ степенью свободы, то гипотеза о виде распределения принимается.

Как видно из формулы, все вычислительные трудности сводятся к расчету чисел p_i. Если, например, проверяется гипотеза о стандартном нормальном распределении, и $\Delta_i = [z_{i-1},\ z_i]$, то $p_i = \Phi(z_i) - \Phi(z_{i-1})$. Если рассматривается гипотеза о нормальном характере распределения, и параметры его неизвестны, а именно такая гипотеза в приложениях встречается наиболее часто, то следует положить

$$p_i = \Phi\left(\frac{z_i - \bar{X}}{S}\right) - \Phi\left(\frac{z_{i-1} - \bar{X}}{S}\right) \tag{1.1}$$

и число степеней свободы распределения хи-квадрат уменьшается на число параметров, которые мы заменили их оценками (а их два). Этот результат следует из более общей теоремы Фишера, которую здесь мы обсуждать не будем.

Итак, для того, чтобы проверить гипотезу о нормальном характере выборки, нужно произвести ее группировку, используя $z_0 = -\infty,\ z_r = \infty$ и заполнить следующую таблицу:

строка	содержание	способ вычисления
1	z_j	по выборке
2	$\frac{z_j-\bar{X}}{S}$	по строке 1 и выборке
3	$\Phi\left(\frac{z_j-\bar{X}}{S}\right)$	по таблице $\Phi(x)$ и строке 2
4	p_j	по формуле (1.1) и строке 3
5	np_j	по строке 4
6	n_j	по выборке
7	$(n_j - np_j)^2$	по строкам 5, 6
8	$\frac{(n_j-np_j)^2}{np_j}$	по строкам 7, 5

Сумма последней, восьмой строки и есть значение статистики χ^2. Оно сравнивается с критической точкой распределения хи-квадрат с $r-3$ степенями свободы, где r - окончательное число групп.

Рассмотрим числовой пример, группировка для которого уже была проведена. Это данные о колебании руки оператора. Напомним, что в этой выборке $\bar{X} = 62,45$, $S = 7,48$. Границы первого и последнего интервалов заменяем бесконечными. Составляем таблицу.

1	z_j	$-\infty$	52,61	58,47	64,33	70,19	∞
2	$\frac{z_j-\bar{X}}{S}$	$-\infty$	-1,31	-0,53	0,25	1,03	∞
3	$\Phi\left(\frac{z_j-\bar{X}}{S}\right)$	0	0,090	0,300	0,580	0,850	1
4	p_j		0.090	0,210	0,280	0,270	0,150
5	np_j		9	21	28	27	15
6	n_j		7	19	39	22	13
7	$(n_j - np_j)^2$		4	4	121	25	4
8	$\frac{(n_j-np_j)^2}{np_j}$		0,444	0,190	4,321	0,926	0,267

Заметим, что строки, начиная с четвертой, не полны, поскольку интервалов на один меньше, чем их границ. Сумма последней строки $\chi^2 = 6,149$. Число степеней свободы 5-3 = 2. Критическая точка по таблице - 5,99. Поэтому формально гипотезу о нормальном характере распределения следует отвергнуть. Однако следует иметь ввиду два обстоятельства - близость расчетного и критического значений, а также то, что группировка, используемая здесь, проведена с нарушением одного из двух основных принципов - имеются "перенаселенные"интервалы. Внимательное изучение таблицы показывает, что именно один из них дает самый существенный вклад в величину χ^2. Поэтому необходим повторный расчет, по окончательной группировке из п. 4, который показывает, что гипотезу следует принять.

С помощью критерия χ^2 можно проверять и ряд других гипотез, например, гипотезу однородности, которая состоит в отсутствии существенных различий в поведении двух наблюдаемых величин. Рассмотрим эту задачу.

Пусть даны две выборки: X объема n и Y объема m. Основная гипотеза состоит в том, что распределения двух наблюдаемых величин совпадают, т.е. фактически наблюдается одна величина. Для проверки объединим две

выборки в одну объема $n+m$, осуществим ее группировку и заполним таблицу, называемую таблицей сопряженности.

интервал	1	...	r	всего
число элементов X	n_1	...	n_r	n
число элементов Y	m_1	...	m_r	m
всего	n_1+m_1	...	n_r+m_r	$n+m$

Затем рассчитываем

$$\chi^2 \;=\; (n+m)\left(\sum_{i=1}^{r}\frac{n_i^2}{n(n_i+m_i)} \;+\sum_{i=1}^{r}\frac{m_i^2}{m(n_i+m_i)}-1\right)$$

и сравниваем с критической точкой распределения хи-квадрат с $r-1$ степенью свободы. Если соответствующее критическое значение не превзойдено, гипотезу об однородности можно принять.

Рассмотрим числовой пример. Группа студентов-филологов (X) и группа студентов-математиков (Y) получила следующие оценки при тестировании в баллах

сумма баллов	10	11	12	13	14	15	16	17	18	Всего
число X	3	4	2	7	9	11	8	5	1	50
число Y	5	3	5	5	5	2	10	2	3	40
всего	8	7	7	12	14	13	18	7	4	90

Расчеты показывают, что $\chi^2 \;=\; 90(1,117-1) \;=\; 10,53$. По таблице хи-квадрат с 8 степенями свободы находим критическую точку 15,57. Таким образом, гипотезу о несущественном влиянии направления образования на результаты теста можно принять.

2.1 Экспертные оценки.

Для установления наличия и степени согласованности мнений двух людей (экспертов) в задачах ранжирования (установления порядка по значимости) некоторых факторов используется коэффициент ранговой корреляции Спирмена, который можно рассматривать, как меру близости мнений одного эксперта к мнениям другого:

$$R(X,Y) \;=\; 1 \;-\; \frac{6\sum_{i=1}^{n} d_i^2}{n^3-n-T_X-T_Y}.$$

Этот коэффициент рассчитывается для пары экспертов X,Y. Здесь d_i- разность рангов, присвоенных экспертами i-му фактору, n - число оцениваемых факторов, $T_X,\ T_Y$ вычисляются для каждого из экспертов X,Y по формуле

$$T \;=\; \frac{1}{2}\sum(t_j^3-t_j),$$

где сумма берется по всем группам связанных (равных) рангов у выбранного эксперта, при этом t_j - число факторов с равными рангами, составляющих j-ю группу. Если у эксперта связанных рангов нет, то для него $T = 0$. Чем больше R, тем сильнее согласованы мнения экспертов.

Тех экспертов, мнения которых наиболее согласованы, можно объединить в группу, и рассматривать эту группу в качестве "коллективного эксперта". Суммы рангов всех экспертов группы для данного фактора называется его коллективным рангом. Ранги, присвоенные факторам в соответствии с возрастанием их коллективных рангов, называются групповыми рангами.

Рассмотрим пример с оцениванием студентами состояния общественного транспорта. У первого, второго и пятого экспертов нет связанных рангов, а значит $T_1 = T_2 = T_5 = 0$. У третьего имеется одна группа связанных факторов, содержащая $t_1 = 2$ фактора (ранги по 6,5). Т.о. $T_3 = (2^3 - 2)/2 = 3$. У четвертого $t_1 = 5$ связанных факторов, поэтому $T_4 = (5^3 - 5)/2 = 60$. Если бы у эксперта имелось две группы связанных факторов, содержащих, скажем, 3 и 2 фактора, то $T = ((3^3 - 3) + (2^3 - 2))/2 = 15$ для такого эксперта. Произведем, например, расчет

$$R(1,2) = 1 - 6\times$$
$$\times \frac{(1-2)^2+(3-1)^2+(5-3)^2+(7-4)^2+(6-7)^2+(4-6)^2+(2-5)^2}{7^3-7-0-0} \approx 0,41.$$

Остальные коэффициенты вычисляются аналогично. Заполним таблицу.

Коэффициенты ранговой корреляции
в задаче о студентах и транспорте

Эксперты	1	2	3	4	5
1	1	0,41	0,75	0,78	0,64
2	-	1	0,24	0,22	0,61
3	-	-	1	0,40	0,43
4	-	-	-	1	0,56
5	-	-	-	-	1

Нижняя часть таблицы не заполнена, т.к. $R(i,j) = R(j,i)$ для всех пар индексов, т.е. таблица симметрична. Объединим в группу первого и четвертого эксперта, т.к. максимальная согласованность мнений наблюдается именно между ними. Можно было бы к ним добавить третьего, но он, несмотря на его хорошую согласованность с первым, имеет не очень хорошее согласование с четвертым. Оставшаяся часть таблицы содержит максимальное число $R(2,5)$ - можно сформировать еще одну группу, состоящую из второго и пятого экспертов. Итак,

- 1-я группа: 1, 4 эксперты;
- 2-я группа: 2, 5 эксперты;
- 3-я группа - 3-й эксперт.

Коллективные ранги в группах:

группы	ф а к т о р ы						
	ч	з	о	д	к	с	ц
1	2	8	10	12	11	9	4
2	5	3	8	8	14	12	6
3	2	2	6,5	6,5	5	2	4

Заполним таблицу групповых рангов:

группы	ф а к т о р ы						
	ч	з	о	д	к	с	ц
1	1	3	5	7	6	4	2
2	2	1	4,5	4,5	7	6	3
3	2	2	6,5	6,5	5	2	4

Теперь, считая каждую группу новым (групповым) экспертом, снова можно вычислить числа $T_i,\ i = 1,2,3$ уже для групп и, например, оценить согласованность мнений разных групп. Заметим, наконец, что значимость отличия R от нуля можно проверять, пользуясь критерием для коэффициента корреляции.

2.2 Регрессионный анализ.

Если имеются основания считать, что случайная величина X зависит от другой случайной величины Y (например, коэффициент корреляции между ними значимо отличается от нуля), то можно попытаться восстановить формулу, выражающую эту зависимость. Естественно, речь может идти только о приближенной формуле из какого-то параметрического класса формул, обычно о линейной зависимости вида

$$Y = \alpha X + \beta, \quad X = kY + b. \tag{2.2}$$

Параметры α, β, k, b подбираются, как правило, из условий

$$S(\alpha,\beta) = \sum_{i=1}^{n}(Y_i - \alpha X_i - \beta)^2 \to \min_{\alpha,\beta} \tag{2.3}$$

и

$$T(k,b) = \sum_{i=1}^{n}(X_i - kY_i - b)^2 \to \min_{k,b} \tag{2.4}$$

Такой метод подбора коэффициентов называют методом наименьших квадратов. Найденные из условий (2.3 , 2.4) уравнения (2.2) могут быть записаны так:

$$Y = \rho\frac{S_Y}{S_X}(X - \bar{X}) + \bar{Y}, \quad X = \rho\frac{S_X}{S_Y}(Y - \bar{Y}) + \bar{X},$$

где ρ- выборочный коэффициент корреляции, S_X, S_Y - корни квадратные из выборочных дисперсий X, Y соответственно. Эти два уравнения называются уравнениями прямых регрессии.

Если мы хотим увидеть характер зависимости между X и Y наглядно, имеет смысл построить так называемое поле корреляции. Для этого осуществим группировку выборки X на r групп $A_1, ..., A_r$, а выборку Y разобьем на s групп $B_1, ..., B_s$. Обозначим через $\Delta_{i,j}$ прямоугольник, проекции которого на оси Ox и Oy совпадают с A_i и B_j соответственно. Двумерный вектор с координатами X, Y попадает, таким образом, в $\Delta_{i,j}$ тогда и только тогда, когда $X \in A_i$ и $Y \in B_j$. Пусть $n_1, ..., n_r$ количества элементов X в $A_1, ..., A_r$, а $m_1, ..., m_s$ - элементов Y в $B_1, ..., B_s$ соответственно. Определим групповые средние

$$\bar{Y}_{(i)} = \frac{1}{n_i}\sum_{\{k:X_k\in A_i\}} Y_k, \quad i = 1, ..., r,$$
$$\bar{X}_{(j)} = \frac{1}{m_j}\sum_{\{k:Y_k\in B_j\}} X_k, \quad j = 1, ..., s.$$

Далее построим на плоскости Oxy ломаные с узлами в точках с координатами $(\tilde{X}_i, \bar{Y}_{(i)})$, $i = 1, ..., r$ и $(\bar{X}_{(j)}, \tilde{Y}_j)$, $j = 1, ..., s$. Напомним, что $\tilde{X}_i$ - это середина отрезка A_i, а $\tilde{Y}_j$ - середина B_j. Построенные ломаные называют эмпирическими линиями регрессии, и их можно рассматривать как приближенные графики зависимостей Y от X и X от Y.

Результаты этого раздела позволяют приближенно предсказывать, какое значение будет принимать одна из величин, если значение другой нам известно. Конечно же, речь идет не о точном значении, а лишь об оценке.

2.3 Дисперсионный анализ.

Рассмотрим задачу выявления и оценки степени влияния некоторого фактора A на изменчивость случайной величины X. Требуется выяснить, является ли влияние фактора на эту величину существенным. Фактор A при этом обычно считается нечисловым категоризованным или числовым, принимающим небольшое число значений. Его градации принято называть уровнями.

Пусть нам заранее известна дисперсия $\mathbf{D}_0 X$ величины X, и мы имеем выборку значений X под действием фактора A. Очевидно, что если A не влияет на изменчивость, то вычисленная по выборке дисперсия $\mathbf{D}$ не значимо отличается от $\mathbf{D}_0 X$. Если же $\mathbf{D}X > \mathbf{D}_0 X$, то следует признать существенный характер влияния. Вообще говоря,

$$\mathbf{D}X = \mathbf{D}_0 X + \mathbf{D}_A X,$$

где через $\mathbf{D}_A X$ обозначена часть дисперсии, объясняемая влиянием фактора A. Если же исследуемых факторов несколько, то

$$\mathbf{D}X = \mathbf{D}_0 X + \mathbf{D}_A X + \mathbf{D}_B X + \mathbf{D}_{A,B} X + ...$$

Идея оценки влияния факторов основана на изучении доли дисперсии, которая объясняется через изучаемый фактор.

Основные предположения, необходимые для применения описываемого далее аппарата:

1. выборки нормально распределены;
2. изучаемый фактор влияет на среднее значение X;
3. дисперсии X в каждой из градаций факторов однородны, т.е. их отличия незначимы.

Основная задача при решении общей проблемы дисперсионного анализа - выделить небольшое число факторов, существенно влияющих на изменчивость наблюдаемой величины, а затем оценить влияние каждого из них. Первый этап задачи - оценка общего влияния группы факторов, второй - частного (парциального) влияния каждого из факторов группы.

Выделяются следующие разновидности дисперсионного анализа: по числу факторов влияния; по количеству градаций (уровней) изменения изучаемых факторов, например, 3- или 2-х уровневый; по наличию или отсутствию повторных (параллельных) испытаний. Различают также полный и дробный, т.е. содержащий пропущенные уровни при испытаниях, дисперсионный анализ.

Рассмотрим подробнее полный однофакторный дисперсионный анализ с параллельными испытаниями. Исходные данные собраны в таблицу. Предполагается, что на каждом из m уровней фактора поставлено по n параллельных испытаний, в которых наблюдалось значение величины X. Ее значения, наблюденные в i-м испытании в предположении, что фактор находился на j-м уровне, обозначены $x_{i,j}$.

Испытание	Уровни		
	A_1	...	A_m
1	$x_{1,1}$	...	$x_{1,m}$
...	...	...	...
n	$x_{n,1}$	...	$x_{n,m}$
средние	$\bar{x}_1$	...	$\bar{x}_m$

Объем выборки здесь, таким образом, равен $N = nm$. Обозначим

$$\bar{\bar{X}} \;=\; \frac{1}{m}\sum_{j=1}^{m}\bar{x}_j \;=\; \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}x_{i,j}.$$

Нетрудно заметить, что

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{i,j}-\bar{\bar{X}})^2 \;=\; \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{i,j}-\bar{x}_j)^2 \;+\; n\sum_{j=1}^{m}(\bar{x}_j-\bar{\bar{X}})^2.$$

Обозначим левую часть этого равенства через Q, двойную сумму в правой части через Q_0, а второе слагаемое справа без множителя n через Q_A. При этом Q интерпретируется как общая изменчивость X, Q_0 как сумма изменчивостей внутри уровней, а Q_A как изменчивость X между уровнями, т.е. при переходе от уровня к уровню.

Таким образом, можно записать

$$\begin{aligned} Q &= \textstyle\sum_{j=1}^{m}\sum_{i=1}^{n}x_{i,j}^2 \;-\; \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n}x_{i,j}\right)^2;\\ Q_0 &= \textstyle\sum_{j=1}^{m}\sum_{i=1}^{n}x_{i,j}^2 \;-\; \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}x_{i,j}\right)^2;\\ Q_A &= \textstyle\frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}x_{i,j}\right)^2 \;-\; \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n}x_{i,j}\right)^2. \end{aligned}$$

Определим

$$D_A \;=\; \frac{Q_A}{N-1}, \quad D_0 \;=\; \frac{Q_0}{N-1}, \quad F \;=\; \frac{nD_A+D_0}{D_0}.$$

Теперь сравним рассчитанное по выборочным данным значение F с критической точкой распределения Фишера с $m-1$, $m(n-1)$ степенями свободы. Если критическое значение не превзойдено, следует принять гипотезу об отсутствии значимого влияния фактора A на величину X.

Пример. Сравнить три метода преподавания (использование различного наглядного материала для обучения). Результаты тестирования в трех группах учеников по 15 человек в баллах, а также некоторые промежуточные расчетные характеристики и их обозначения приводятся в таблице. Заметим, что в этом примере $N = 45$.

Вычислим по таблице $z \;=\; (S_1+S_2+S_3)/15 \;=\; 7539,0$. Теперь мы готовы привести основные результаты дисперсионного анализа:

- Факторная вариативность $Q_A \;=\; z \;-\; \frac{t^2}{N} \;=\; 89,2$, $D_A = 2,03$.

Сравнение трех методов преподавания

учащийся	метод 1	метод 2	метод 3	
1	9	15	18	
2	11	16	14	
3	10	15	17	
4	12	10	9	
5	7	13	14	
6	11	14	17	
7	12	15	16	
8	10	7	15	
9	13	13	16	
10	11	15	8	
11	13	15	14	
12	11	14	10	
13	10	11	16	
14	12	15	15	
15	13	10	17	всего
суммы S_j	165	198	216	$t = 579$
суммы квадратов S_j^2	1853 27225	2706 39204	3242 46656	$u = 7801$ $v = 113085$

- Случайная вариативность $Q_0 = u - z = 262,0,\ D_0 = 5,95.$
- Общая вариативность $Q = u - \frac{t^2}{N} = 351,2.$
- Значение критерия $F = 6,12.$

По таблице распределения Фишера с 2, 42 степенями свободы, находим критическую точку уровня 0,01. Она равна 5,15. Поскольку расчетное значение критерия больше, то следует признать существенность влияния метода применения наглядных материалов на успеваемость учащихся. Более того, определяя долю Q_A в Q, видим, что примерно треть изменчивости баллов тестирования может быть объяснена через этот фактор.

2.4 Проблема отбора наиболее информативных показателей.

Пусть при изучении n объектов у каждого из них наблюдается большое количество p показателей. Рассмотрим задачу уменьшения числа p с наименьшими потерями информации, содержащихся в наших наблюдениях. При этом исследователем могут ставиться следующие цели:

1. большая наглядность (визуализация) данных;
2. лаконизм получаемой модели, обозримость и простота зависимостей;

3. сжатие объемов хранимой информации о наблюдениях.

Конечно же, возможны иные цели или комбинация перечисленных. Меньшее количество признаков q (как правило, $q \ll p$) может выбираться из уже имеющихся p или строиться вновь, как комбинации наблюдаемых показателей. Возможны разные варианты требований к новым показателям, например:

- наибольшая информативность (в разных смыслах);
- взаимная независимость показателей или их некоррелированность;
- наименьшее искажение геометрической структуры данных и т.д.

В зависимости от вида требований задается критерий оптимальности для предлагаемой системы признаков и строится алгоритм оптимального построения. При этом имеется три основных типа предпосылок к успешному решению поставленной задачи:

1. дублирование информации (сильная связь между показателями);
2. неинформативность некоторых из показателей (их незначительная изменчивость при переходе от объекта к объекту);
3. возможность агрегирования, т.е. объединения нескольких показателей в один без существенного ущерба для информативности.

Поставим задачу снижения размерности формально. Пусть $x^{(1)}, ...x^{(p)}$ - наблюдаемые показатели, $X = (x^{(1)}, ..., x^{(p)})$, $Z = Z(X)$ - q-мерная векторная функция, $q \ll p$, $Z(X) = (Z^{(1)}(X), ..., Z^{(q)}(X))$, $I_q(Z(X))$ - мера информативности или критерий оптимальности. Этот критерий определяется сущностью решаемой задачи. Варианты I будут приведены ниже. Предположим также, что задан класс F допустимых преобразований Z. Тогда задача ставится так: построить такое $\tilde{Z}$ из класса F, что

$$I_q(\tilde{Z}(X)) = \max_{Z \in F} I_q(Z(X)).$$

Тот или иной выбор I, F приводит к методу главных компонент, факторному анализу, методу экстремальной группировки признаков, многомерному шкалированию, дисперсионному или регрессионному анализу. Далее следует краткий обзор этих методов, некоторые из которых подробнее рассмотрены в последующих разделах.

2.4.1 Метод главных компонент.

Здесь F - класс линейных преобразований следующего вида:

$$Z^{(j)} = \sum_{k=1}^{p} c_{j,k}(x^{(k)} - \bar{x}^{(k)}), \quad j = 1, ..., q,$$

причем

$$\sum_{k=1}^{p} c_{j.k}^2 \;=\; 1,\; j=1,...,q, \quad \sum_{k=1}^{p} c_{j,k}c_{i,k} \;=\; 0,\; j,k=1,..,q,\; j\neq k.$$

Критерием оптимальности будет

$$I_q(Z) \;=\; \frac{\sum_{j=1}^{q} \mathbf{D}Z^{(j)}}{\sum_{j=1}^{p} \mathbf{D}x^{(j)}}.$$

Итак, мы ищем такую нормированную комбинацию исходных показателей, изменчивость которой которая объясняет максимально возможную долю изменчивости всего набора исходных показателей.

2.4.2 Экстремальная группировка признаков.

Поставим задачу разбить $x^{(1)},...,x^{(p)}$ на заданное число групп $S_1,...,S_q$ так, чтобы внутри одной группы показатели были коррелированы относительно сильно, а между группами наблюдалась бы относительно слабая корреляция. После того, как это проделано, каждую группу заменим одним показателем Z.

Здесь F - класс нормированных линейных преобразований, выбираемых так, чтобы $\mathbf{D}Z=1$,

$$I_q(Z,S) \;=\; \sum_{j=1}^{q}\sum_{k\in S_j} \rho^2(x^{(k)},Z^{(j)}),$$

где $\rho(.,.)$ - коэффициент корреляции. При этом набор групп S также может подбираться. Простейший пример такого метода - группировка выборки (см. соответствующий раздел).

2.4.3 Многомерное шкалирование.

Предположим, что результатами наблюдений являются не состояния каких-то объектов, а характеристики их попарной близости $\rho_{i,j}$ или расстояния между ними $d_{i,j}$. Такая ситуация наблюдается, когда мы изучаем опросы, анкетирование или экспертные оценки. Задача состоит в наглядном изображении результатов в пространстве небольшого количества измерений с наименьшим геометрическим искажением структуры данных. Определим

$$\Delta(Z) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,j}^{\alpha} \mid \hat{d}_{i,j}(Z) \;-\; d_{i,j} \mid^{\beta}.$$

Здесь $\hat{d}_{i,j}$ - расстояния между объектами в новом пространстве более низкой размерности, а числа α,β выбираются исследователем ($\alpha<0$).

Критерием оптимальности является

$$I_q(Z) \ = \ \frac{1}{1+\Delta(Z)}.$$

Это число тем меньше, чем меньше искажается "геометрическая структура данных".

2.4.4 Отбор наиболее информативных показателей в модели дискриминантного анализа.

Выше были рассмотрены так называемые задачи с автоинформативными критериями, т.е. критерии оптимальности подбирались из самой внутренней логики данных. Сейчас рассмотрим задачу с внешним критерием оптимальности. Этот критерий будем строить исходя из правильности разбиения объектов наблюдения на заранее определенные группы. Этими группами могут быть, например, врачебные диагнозы или темпераменты испытуемых или, в конце концов, примитивная классификация типа надежный - ненадежный, здоровый - больной, хороший - плохой и т.п. При этом класс преобразований F допускает лишь выбор из исходных показателей, без комбинаций и вращений.

$$Z(X) \ = \ (X^{i_1)}, ..., X^{(i_q)}).$$

Критерией оптимальности будет строится из условия максимальности отличия распределений показателей в разных группах:

$$I_q(Z) \ = \ \sum_{i,j=1}^{k} \delta(P_i(Z), P_j(Z)),$$

где k - число групп, а $\delta(.,.)$ - некоторая мера различия распределений. В простейшем случае, когда классификация проводится по величине средних, можно взять

$$\delta(P_i(Z), P_j(Z)) \ = \ \sum(\bar{X}_i \ - \ \bar{X}_j)^2.$$

2.4.5 Модель регрессии.

Задачи регрессии были уже рассмотрены выше. Отметим только, что в этих задачах в основе критерия оптимальности лежит наибольшая коррелированность Y с $aX+b$, а это значит, что перед нами вновь внешний критерий.

2.5 Метод главных компонент.

Здесь мы должны выбрать такие комбинации показателей, которые имеют наибольшую изменчивость при переходе от объекта к объекту. Часто это действительно возможно. Например, при снятии мерки с клиента портной

снимает 11 показателей, тогда как при покупке готовой одежды мы довольствуемся двумя - тремя (размер, рост, полнота). Формально: пусть X - p-мерый вектор, μ - вектор его средних, S - $p \times p$ - ковариационная матрица, критерий оптимальности задан формулой

$$I_q \;=\; \frac{\sum_{j=1}^{q} \mathbf{D} z^{(j)}}{\sum_{i=1}^{p} \mathbf{D} x^{(i)}} \;\to\; \max,$$

где $Z \;=\; L(X-\mu)$, $L \;-\; q \times p$-матрица с ортогональными строками (подбирается из условия оптимальности).

Итак, первая главная компонента - такая центрированно - нормированная комбинация координат X, которая обладает наибольшей дисперсией среди всех таких комбинаций, ..., k-я главная компонента - такая центрированно - нормированная комбинация, которая некоррелирована с $k-1$ предыдущими главными компонентами и среди всех таких комбинаций обладает наибольшей дисперсией. Зачит, элементы матрицы L для первой главной компоненты подбираем из условий

$$\begin{cases} \mathbf{D} \sum l_j^{(1)} x^{(j)} \to \max, \\ \sum (l_j^{(1)})^2 \;=\; 1, \end{cases} \qquad \text{или } \sum_{i,j} S_{i,j} l_i^{(1)} l_j^{(1)} \;\to\; \max,$$

что в матричной записи имеет вид

$$< S l^{(1)}, l^{(1)} > \to \max_{l_1}, \quad \|l_1\| = 1,$$

и аналогичных условий для остальных компонент, откуда $l^{(j)}$ - j-й собственный вектор матрицы S, имеющий единичную длину, и дисперсия j-й главной компоненты равна собственному числу λ_j. Решение этой задачи возможно всегда, т.к. S - симметричная положительно определенная матрица. При этом, если все параметры измерены в единицах одного масштаба, то

$$I_q \;=\; \frac{\lambda_1}{\lambda_1 + ... + \lambda_p},$$

иначе параметры следует предварительно нормировать.

Рассмотрим числовой пример. По данным измерений в миллиметрах длины x_1, ширины x_2 и высоты x_3 панциря 24 особей одного из видов черепах определена выборочная ковариационная матрица

$$S \;=\; \begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 168,70 & 103,29 & 66,65 \end{pmatrix}.$$

Для нахождения собственных чисел решаем кубическое характеристическое уравнение. Его корни: $\lambda_1 = 680,40,\ \lambda_2 = 6,50\ \lambda_3 = 2,86$. Соответствующие собственные векторы:

$$l_1 = \begin{pmatrix} 0,81 \\ 0,50 \\ 0,31 \end{pmatrix}, \quad l_2 = \begin{pmatrix} -0,55 \\ 0,83 \\ 0,10 \end{pmatrix}, \quad l_3 = \begin{pmatrix} -0,21 \\ -0,25 \\ 0.95 \end{pmatrix}.$$

Отсюда при $i = 1,2,3$ получаем $z^{(i)} = < l_i, X >$, где X - вектор отклонений x_j от соответствующих средних значений.

Вычислим величины вклада координат Z в изменчивость параметров. При этом

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9864,$$

т.е. более 98 процентов информации о всех трех размерах содержится в первой главной компоненте - а значит, ее и нужно использовать для классификации.

2.6 Факторный анализ.

Рассмотрим задачу объяснения изменчивости показателей $x^{(1)}, ..., x^{(p)}$ через непосредственно не наблюдаемые (латентные) факторы $f^{(1)}, ..., f^{(q)}$. Условимся также считать, что факторы f между собой некоррелированы. Целью нашего исследования будет выявление и интерпретация латентных факторов. При этом возникает также желание минимизировать число этих факторов и степень зависимости исходных показателей от той части их мзменчивости, которая не объясняется через $f^{(i)}$ - и это желание внутренне противоречиво. Можно интерпретировать изменчивость латентных факторов как причину, а изменчивость наблюдаемых показателей как следствие.

Перейдем к математической модели. Будем считать $x^{(j)}$ центрированными, обладающими нулевыми средними. Если они такими не были, то этого легко добиться, вычитая средние из значений каждого.

$$X = QF + U,$$

где $Q - p \times q$-матрица неизвестных коэффициентов при неизвестных факторах F, называемая матрицей нагрузок латентных факторов на показатели X, U - вектор остаточных компонент, необъяснимый с точки зрения вводимых факторов. Предполагается, что U имеет нормальный характер распределения, его компоненты независимы и не зависят от F. Через V обозначим ковариационную матрицу U. Принципиальное отличие от задач регрессии и дисперсионного анализа здесь состоит в том, что все, кроме X в выписанной формуле неизвестно.

Для того, чтобы понять связь поставленной задачи с методом главных компонент, предположим, что нашлись (возможно, в неограниченном числе) такие факторы, что

$$X = AF, \quad \mathbf{D}f^{(i)} = 1, \ i = 1, 2...$$

Отметим, что матрица A и вектор F в данной записи определены неоднозначно, достаточно взять $Z = CF$, тогда $X = (AC)Z$. Матрица C может быть любой, но необходимость сохранения некоррелированности новых факторов, накладывает на нее условие ортогональности. Итак, после нахождение каких-то A, F возможно *вращение*. Теперь через $F(m)$ обозначим

"урезанный"вектор (оставляем только первые m координат), а через A_m соответствующим образом "урезанную"матрицу и вместо X рассмотрим

$$\hat{X}(m) \;=\; A_m F(m).$$

Если теперь мы объявим критерием оптимальности минимальное отличие ковариационных матриц исходных показателей X и "урезанных"показателей $\hat{X}(m)$, то получим, что

$$f^{(i)} \;=\; \frac{1}{\sqrt{\lambda_i}}\tilde{f}^{(i)},$$

где $\tilde{f}^{(i)}$ - i-я главная компонента, а i -й столбец матрицы A имеет вид $\sqrt{\lambda_i}l_i$, l_i - собственный вектор ковариационной матрицы S исходных показателей, отвечающий собственному числу λ_i. Таким образом, в этом случае мы приходим к методу главных компонент. Если же взять за критерий оптимальности максимальное объяснение корреляции между исходными показателями с помощью латентных факторов, например, оценив адекватность такого объяснения через близость ковариаций между $x^{(i)}, x^{(j)}$ и $\hat{x}^{(i)}, \hat{x}^{(j)}$ соответственно, придем к задаче факторного анализа.

В исходной модели оказывается слишком много параметров для их точного определения. Поэтому обычно накладываются некоторые дополнительные условия. Например, можно искать матрицу нагрузок в виде

$$Q = \begin{pmatrix} q_{1,1} & 0 & 0 & ... & 0 \\ q_{2,1} & q_{2,2} & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ q_{q,1} & q_{q,2} & q_{q,3} & \cdots & q_{q,q} \\ ... & ... & ... & ... & ... \\ q_{p,1} & q_{p,2} & q_{p,3} & \cdots & q_{p,q} \end{pmatrix},$$

т.е. первый показатель мы объясняем только через первый фактор, второй показатель - через первый и второй и т.д. Возможны, конечно, и другие варианты условий, иногда объясняющиеся внутренней логикой решаемой задачи.

Существует несколько разработанных методов для оценивания матрицы нагрузок Q и матрицы ковариаций V остаточной компоненты. Мы остановимся только на центроидном методе, подробное описание которого также не входит в наши задачи. Опишем только геометрическую интерпретацию этого метода. Аккуратный же подсчет этим или другим методом в каждой конкретной задаче оставим на долю вычислительной техники (соответствующее программное обеспечение имеется в любом пакете прикладных статистических программ).

Отождествим $x^{(1)}, ..., x^{(p)}$ с векторами, выходящими из начала координат так, чтобы косинусы углов между i-м и j-м были бы равны коэффициентам корреляции $\rho_{i,j}$, а длины этих векторов - $\mathbf{D}x^{(i)}$. Изменим направления некоторых из этих векторов на противоположные так, чтобы как можно большее число ковариаций стали бы положительными (образуем тесный "пучок"). Обозначим через $f^{(1)}$ центральный вектор "пучка", имеющий

единичную длину. Перейдем теперь к остаточным показателям, вычитая из каждого из векторов проекцию $f^{(1)}$ на его направление:

$$x^{(i1)} \;=\; x^{(i)} \;-\; \hat{q}_{i,1} f^{(1)}.$$

Далее процесс повторяется с остаточными показателями до тех пор, пока не будет выделено нужное число показателей и определены оценки нагрузок $\hat{Q}$. Для оценивания V применяем соотношение

$$\hat{V} \;=\; S \;-\; \hat{Q}\hat{Q}^t.$$

Одной из главных задач факторного анализа является задача оценивания значений латентных факторов для каждого изучаемого объекта. Чтобы понять значимость этой задачи, отметим, что например при изучении результатов некоторого интеллектуального тестирования в роли латентных факторов обычно выступают способности тестируемой личности, а численная оценка таких способностей в той или иной шкале весьма привлекательна.

Предположим, что Q, V мы уже оценили. Метод Бартлетта интерпретирует F, как коэффициенты регрессии q на x:

$$x_k^{(i)} \;=\; \sum_{j=1}^{q} q_{i,j} f_k^{(j)} \;+\; u_k^{(i)}, \quad i = 1, ..., p,\; k = 1, ..., n.$$

Их находим далее, применяя, как обычно, метод наименьших квадратов:

$$\hat{F}_k \;=\; (\hat{Q}^t \hat{V}^{-1} \hat{Q})^{-1} \hat{Q}^t X_k, \quad k = 1, ..., n.$$

Другой метод, метод Томсона, "выворачивает"описанный выше процесс наизнанку. Найдем коэффициенты $c_{i,j}$, участвующие в соотношении $\hat{F} = CX$ по методу наименьших квадратов, т.е. решим задачу на минимум:

$$\sum_{k=1}^{n} \sum_{i=1}^{q} \mathbf{M}(f_k^{(i)} \;-\; \sum_{j=1}^{p} c_{i,j} x_j^{(k)})^2 \;\to\; \min_C .$$

При этом, хотя сами $f_k^{(i)}$ неизвестны, нам достаточно знать их дисперсии и ковариации, которые легко извлекаются из соотношения

$$\mathbf{M}\left(\begin{pmatrix} X \\ F \end{pmatrix} \begin{pmatrix} X \\ F \end{pmatrix}^t \right) \;=\; \begin{pmatrix} QQ^t + V & Q \\ Q^t & I \end{pmatrix}.$$

Получаем

$$\hat{F}_k \;=\; (I \;+\; \hat{Q}^t \hat{V}^{-1} \hat{Q})^{-1} \hat{Q} \hat{V}^{-1} X_k, \quad k = 1, ..., n.$$

Рассмотрим числовой пример. После изучения оценок 220 английских школьников получена следующая корреляционная матрица оценок по гэльскому языку, английскому языку, истории, арифметике, алгебре и геометрии:

	x_1	x_2	x_3	x_4	x_5	x_6	$q_{i,1}$	$q_{i,2}$
x_1	1	0,439	0,410	0,288	0,329	0,248	0,606	0,337
x_2	0,439	1	0,351	0,354	0,320	0,329	0,611	0,197
x_3	0,410	0,351	1	0,164	0,190	0,181	0,458	0,384
x_4	0,288	0,354	0,164	1	0,595	0,570	0,683	-0,365
x_5	0,329	0,320	0,190	0,595	1	0,464	0,686	-0,335
x_6	0,248	0,329	0,181	0,570	0,464	1	0,575	-0,212

Матрица была подвергнута бифакторному анализу. В последних двух столбцах таблицы приведены нагрузки, полученные центроидным методом. Следующая задача - подсчитать значения двух латентных факторов для каждого из 220 учеников, после чего данные можно представить геометрически в виде облака из 220 точек плоскости. Метод Томсона дает

$$f_1 = 0,245x_1 + 0,208x_2 + 0,158x_3 + 0,278x_4 + 0,271x_5 + 0,157x_6,$$

$$f_2 = 0,352x_1 + 0,201x_2 + 0,309x_3 - 0,351x_4 - 0,303x_5 - 0,126x_6.$$

Простой анализ таблицы и полученных формул дает возможность интерпретировать f_1 как общую одаренность, а f_2 как гуманитарную одаренность школьника.

2.7 Многомерное шкалирование.

У нас имеется одна или несколько матриц Δ^l с элементами $\delta^l_{i,j}$, представляющие из себя оценки расстояний между i-м и j-м объектами с точки зрения l-го эксперта (возможно, просто ранги соответствующих расстояний). Исследуемые объекты располагаются в p-мерном пространстве. Задача состоит в построении геометрической комбинации точек в q-мерном пространстве ($q < p$) так, чтобы ранговый порядок расстояний между ними совпадал с определенным в Δ^l (если матрица была одна), или наилучшим образом согласовывался со всеми имевшимися матрицами. В последнем случае говорят о шкалировании индивидуальных различий. Эту задачу здесь мы более не будем упоминать.

Обычно в задачах многомерного шкалирования используют метод Торгерсона, одну из модификаций которого рассмотрим ниже. По матрице Δ рассчитаем ковариационную и корреляционную матрицы оценок расстояний. Введем в рассмотрение различия между i-м и j-м объектами

$$\delta_{i,j} = \sqrt{1 - \rho_{i,j}},$$

где $\rho_{i,j}$ - элементы корреляционной матрицы (коэффициенты корреляции). Из матрицы ковариаций выделим q нормированных главных компонент (см. соответствующий раздел). Пусть $\hat{x}^0_{i,k},\ i = 1,...p,\ k = 1,...,q$ - координаты объектов в главных компонентах. Первоначальные оценки расстояний вычисляем по формуле

$$\hat{d}^0_{i,j} = \sqrt{\sum_k (\hat{x}^0_{i,k} - \hat{x}^0_{j,k})^2}.$$

Числа $\delta^0_{i,j} = \delta_{i,j},\ \hat{d}^0_{i,j}$ образуют стартовую комбинацию метода. Характеристикой качества каждой из рассматриваемых здесь и ниже комбинаций служит так называемый стресс-критерий

$$S = \sqrt{\frac{\sum_{i,j}(\delta_{i,j} - \hat{d}_{i,j})^2}{\sum_{i,j}\hat{d}^2_{i,j}}}.$$

Рассмотрим далее итерационный процесс, начинающийся со стартовой комбинации.

Сначала расположим все пары $(i,j),\ i,j = 1,...,p$ по возрастанию различий $\delta_{i,j}$, параллельно с ними разместив отклонения $\hat{d}_{i,j}$ в следующем столбце. Равные отклонения объединяем в блоки (на первом шаге чаще всего каждый из блоков содержит ровно один объект). Первая часть этапа состоит в многократных проходах по столбцу отклонений. Если отклонения в следующем блоке меньше, чем в предыдущем, то объединяем их в один новый блок, заменяя в нем все отклонения на среднее арифметические отклонений по объединяемым блокам. Иначе не меняем ничего. Признаком окончания этой части этапа служит отсутствие изменений при очередном проходе.

Вторая часть этапа состоит в пересчете координат и отклонений по формулам

$$\hat{x}^{c+1}_{i,k} = \hat{x}^c_{i,k} - \frac{1}{p}\sum_j \left(1 - \frac{\delta^{c+1}_{i,j}}{\hat{d}^c_{i,j}}\right)(\hat{x}^c_{i,k} - \hat{x}^c_{j,k})$$

$$\hat{d}^{c+1}_{i,j} = \sqrt{\sum_k (\hat{x}^{c+1}_{i,k} - \hat{x}^{c+1}_{j,k})^2},$$

Здесь $\delta^{c+1}_{i,j}$ - различия, полученные в процессе выполнения первой части этапа, c - номер этапа.

Проверяем существенность изменения стресса S. Если это изменение несущественно (меньше выбранного заранее малого числа) - комбинация построена, иначе повторяем описанный выше этап, приняв за исходную построенную комбинацию $\delta^{c+1}_{i,j}$ (не изменяются по отношению к рассчитанным), $\hat{d}^{c+1}_{i,j}$.

2.8 Оцифровка нечисловых данных.

Пусть X - матрица данных наблюдений n p-мерных объектов в неколичественной шкале. Рассмотрим задачу присваивания "удобных"числовых меток каждому из данных для обработки методами статистического анализа. Этот же метод может быть применен для числовых переменных, принимающих небольшое количество значений. Критерий "удобности"зависит от используемого далее метода анализа. Пусть признак x имеет l_x градаций и каждой из них присвоена числовая метка $c^{(x)}_j,\ j = 1,...,l_x$, при этом выпол-

нены условия центрированности и нормированности меток:

$$\sum_{i=1}^{n} c_{r(i)}^{(x)} \;=\; 0; \quad \frac{1}{n}\sum_{i=1}^{n}(c_{r(i)}^{(x)})^2 \;=\; 1,$$

где $r(i)$ - номер градации признака x для i-го объекта.

Рассмотрим только оцифровку данных для задач сокращения размерностей. Критерий поиска оптимальных меток - максимизация

$$Q \;=\; \sum_{i<j} \rho_{i,j}^2,$$

где $\rho_{i,j}$ -коэффициент корреляции между i-м и j-м признаками после кодировки.

Пусть теперь все признаки $x^{(1)}, ..., x^{(p)}$ разбиты на две группы - X^1 из q признаков, подлежащих оцифровке и X^2 из $p-q$ признаков, которые уже оцифрованы. Соответственно разбито и Q - сумма коэффициентов корреляции внутри X^1, между X^1 и X^2:

$$Q \;=\; Q_1 \;+\; Q_{1,2} \;+\; Q_2.$$

Очевидно, Q_2 не зависит от оцифровки.

В заключение рассмотрим числовой пример, связанный с 12 посетителями кафе (см выше). В следующей ниже таблице приведены рассчитанные метки:

	x_3	x_4	x_5
1	0,89	-0,51	0,62
2	-0,46	0,61	0,53
3	-0,03	-0,49	0,51
4	-	0,38	-0,29

Сумма квадратов коэффициентов корреляции 5,5541. Отметим, что до оцифровки она была равна 2,719.

Литература

[1] Е.Ю.Артемьева, Е.М.Мартынов. Вероятностные методы в психологии.- М.:МГУ, 1975. - 206 с.

[2] С.А.Айвазян, В.М.Бухштабер, И.С.Енюков, Л.Д.Мешалкин Прикладная статистика: Классификация и снижение размерности. - М.:Финансы и статистика, 1989. - 607 с.

[3] Л.Н.Болшев, Н.В.Смирнов - Таблицы математической статистики - М.:Наука, 1983. - 416 с.

[4] Л.Ф.Бурлачук, С.М.Морозов Словарь-справочник по психологической диагностике.- Киев: Наукова думка, 1989. - 200 с.

[5] М.Дейвисон - Многомерное шкалирование: методы наглядного представления данных. - М.:Финансы и статистика, 1988.- 254 с.

[6] В.С.Дронов - Основы математики (избранные главы). - Барнаул: Изд-во АГУ, 1998 - 95 с.

[7] П.Мюллер, П.Нойман, Р.Шторм - Таблицы по математической статистике. - М.:Финансы и статистика, 1987. - 278 с.