

УЧЕБНЫЕ ПОСОБИЯ

**Ю.В. САЖИН
В.А. БАСОВА**

**МНОГОМЕРНЫЕ
СТАТИСТИЧЕСКИЕ МЕТОДЫ**



Москва

Ю.В. САЖИН
В.А. БАСОВА

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

УЧЕБНОЕ ПОСОБИЕ

*Рекомендовано Учебно-методическим объединением
по статистике в качестве учебного пособия
для студентов по специальности «Статистика»*



Москва 2002

ББК 60.6в6

С 14

Рецензенты:

кафедра макроэкономической, международной статистики
и национального счетоводства Московского
государственного университета экономики,
статистики и информатики;
доктор экономических наук, профессор В.С. Мхитарян.

Сажин Ю.В., Басова В.А.

С 14 Многомерные методы анализа: Учеб. пособие. – М.: Компания Спутник+, 2002. – 163 с.

ISBN 5-93406-253-0

В учебном пособии рассматриваются многомерные генеральная и выборочная совокупности, основные теоретические положения кластерного, дисперсионного, компонентного и факторного анализа. Изложены некоторые вопросы оценивания и сравнения параметров многомерных генеральных совокупностей. Пособие содержит задачи и методические указания к выполнению контрольной работы, тесты для осуществления самоконтроля усвоения знаний, данные для выполнения лабораторной работы на ЭВМ.

Предназначено для студентов по специальности «Статистика» дневного и заочного отделений. Может быть полезно студентам вузов и колледжей всех экономических специальностей.

ББК 60.6в6

ISBN 5-93406-253-0

© Сажин Ю.В., Басова В.А., 2002



МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

УЧЕБНОЕ ПОСОБИЕ

ББК 60.6в6

С 14

Рецензенты:

кафедра макроэкономической, международной статистики
и национального счетоводства Московского
государственного университета экономики,
статистики и информатики;
доктор экономических наук, профессор В.С. Мхитарян.

Сажин Ю.В., Басова В.А.

С 14 Многомерные методы анализа: Учеб. пособие. – М.: Ком-
пания Спутник+, 2002. – 163 с.

ISBN 5-93406-253-0

В учебном пособии рассматриваются многомерные генеральная и выборочная совокупности, основные теоретические положения кластерного, дисперсионного, компонентного и факторного анализа. Изложены некоторые вопросы оценивания и сравнения параметров многомерных генеральных совокупностей. Пособие содержит задачи и методические указания к выполнению контрольной работы, тесты для осуществления самоконтроля усвоения знаний, данные для выполнения лабораторной работы на ЭВМ.

Предназначено для студентов по специальности «Статистика» дневного и заочного отделений. Может быть полезно студентам вузов и колледжей всех экономических специальностей.

ББК 60.6в6

ISBN 5-93406-253-0

© Сажин Ю.В., Басова В.А., 2002



МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

УЧЕБНОЕ ПОСОБИЕ

СОДЕРЖАНИЕ

Введение	6
1. МНОГОМЕРНАЯ ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТИ	7
1.1. Распределение генеральной совокупности	7
1.2. Некоторые характеристики генеральной совокупности	8
1.3. Многомерная нормально распределенная генеральная совокупность	11
2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ	14
2.1. Точечные оценки параметров многомерной генеральной совокупности	15
2.2. Доверительные области	17
2.3. Доверительная область для вектора математических ожиданий	18
3. ГИПОТЕЗЫ О ПАРАМЕТРАХ МНОГОМЕРНОЙ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ	21
3.1. Сравнение вектора генеральных средних со стандар- том	21
3.2. Сравнение двух генеральных совокупностей	22
4. ОСНОВЫ КЛАСТЕРНОГО АНАЛИЗА	26
4.1. Расстояние между объектами (кластерами) и мера близости	31
4.2. Функционалы качества разбиения	35
4.3. Иерархические кластер-процедуры	37
5. ДИСПЕРСИОННЫЙ АНАЛИЗ	39
5.1. Сущность метода	39
5.2. Правила организации дисперсионного комплекса	45
5.3. Однофакторный дисперсионный комплекс	48
5.4. Двухфакторный дисперсионный комплекс	49
6. КОМПОНЕНТНЫЙ АНАЛИЗ	55

6.1. Основные понятия факторного и компонентного ана- лиза	55
6.2. Метод главных компонент	58
7. ФАКТОРНЫЙ АНАЛИЗ	67
7.1. Факторный анализ как статистический метод	67
7.2. Основные техники факторного анализа	71
7.3. Модель факторного анализа	74
7.4. Основные методы определения оценок общностей	78
7.5. Вращение пространства общих факторов	83
ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ	89
ПРАКТИЧЕСКАЯ ЧАСТЬ	110
1. Задачи по кластерному анализу	111
2. Задачи по дисперсионному анализу	115
3. Задачи по компонентному анализу	121
4. Задачи по факторному анализу	127
ТЕСТЫ	131
ЛАБОРАТОРНАЯ РАБОТА	151
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	153

ВВЕДЕНИЕ

На современном этапе развития науки признанным средством изучения социально-экономических процессов и явлений становятся многомерные статистические методы, многие из которых реализованы на ПЭВМ. Овладение мощным аппаратом современного многомерного анализа – одна из составляющих университетского экономического образования. Многомерный экономико-статистический анализ опирается на широкий спектр методов. Изложить все мыслимые подходы к исследованию в достаточно кратком учебном пособии не представляется возможным. Поэтому ограничимся рассмотрением тех методов, которые уже получили наибольшее распространение в практике исследовательской работы, а также в практической деятельности специалистов разных профилей.

Кластерный анализ является одним из методов многомерной классификации, которые предназначены разделять рассматриваемые генеральные совокупности объектов, субъектов или явлений на группы, в определенном смысле однородные.

Сущность дисперсионного анализа заключается в изучении степени влияния одного или нескольких факторных признаков на результирующий признак, что в некотором роде аналогично задачам корреляционного анализа.

Наличие множества исходных признаков, характеризующих экономические процессы, ставит целью отбор наиболее существенных, чтобы в дальнейшем изучать меньший набор показателей. Достижение такой цели обеспечивается методами снижения размерности, к которым относятся факторный и компонентный анализы.

Учебное пособие состоит из нескольких разделов: изложения теоретических основ многомерных статистических методов, примеров решения типовых задач, задач для самостоятельного решения, тестов и данных для выполнения лабораторной работы на ПЭВМ, что делает его удобным и полезным как для самостоятельной работы по изучению многомерного статистического анализа, так и для организации учебного процесса преподавателем в соответствии с современными требованиями к последнему.

1. МНОГОМЕРНАЯ ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТИ

1.1. Распределение генеральной совокупности

Закономерности, которым подчиняется в математической статистике исследуемая переменная (случайная величина), полностью определяются комплексом условий ее наблюдения. Математически эти закономерности задаются соответствующими законами распределения вероятностей. Однако при проведении статистических исследований более удобным и распространенным является понятие генеральной совокупности. Таким образом, понятия «генеральная совокупность», «случайная величина» и «закон распределения вероятностей», соответствующие данному комплексу условий, можно считать в определенном смысле синонимами.

Генеральной совокупностью называется множество всех мыслимых наблюдений, которые могли бы быть произведены при данном комплексе условий.

Поскольку в определении идет речь о мысленно возможных наблюдениях, то генеральная совокупность есть понятие абстрактное и ее не следует смешивать с реальными совокупностями, подлежащими статистическому исследованию. Так, обследовав даже все предприятия отрасли, мы можем рассматривать их как представителей гипотетически возможной более широкой совокупности предприятий, которые могли бы функционировать в рамках комплекса условий.

Генеральная совокупность может быть как конечной, так и бесконечной. Конечная совокупность имеет место, например, при обследовании школ города. Бесконечная генеральная совокупность наблюдается, например, в научных исследованиях, когда нас интересует средний результат большого числа экспериментов.

В простейшем случае генеральная совокупность есть одномерная случайная величина X с функцией распределения

$$F(x) = P(X \leq x),$$

которая определяется вероятностью того, что X примет значение, меньшее фиксированного действительного числа x .

В многомерном статистическом анализе изучаются генеральные совокупности с точки зрения нескольких признаков (обычно более 2-х). Рассматриваемое множество признаков обозначается вектором X , имеющим k компонент, каждая из которых характеризует соответствующий признак $x_j, j = 1, 2, \dots, k$.

Таким образом, объектом исследования в многомерном статистическом анализе является случайный вектор X в k -мерном евклидовом пространстве, система k случайных (одномерных) величин, k -мерная случайная величина $(x_1, x_2, \dots, x_k)$.

Функцией распределения случайного вектора X называется детерминированная неотрицательная величина, определяемая по формуле:

$$F(x) = P(X < x_1, X < x_2, \dots, X < x_k) = P(x < X),$$

где x - k -мерный вектор фиксированных действительных чисел $x = (x_1, x_2, \dots, x_k)^T$.

Например, в простейшем случае, когда X одномерная случайная величина, графическое изображение функции распределения представляет собой некоторую неубывающую кривую, значения которой изменяются от 0 до 1. В случае дискретной случайной величины $F(x)$ увеличивается скачками всякий раз, когда x при своем изменении проходит через одно из возможных значений x_j случайной величины X . Между двумя соседними значениями функция $F(x)$ постоянна, поэтому ее графическое изображение для дискретной случайной величины имеет вид ступенчатой кривой, а для непрерывной случайной величины - плавной монотонно возрастающей кривой.

1.2. Некоторые характеристики генеральной совокупности

Закон распределения полностью характеризует генеральную совокупность, то есть последняя может быть задана функцией распределения или плотностью распределения. Однако такая ис-

черпывающая характеристика генеральной совокупности довольно сложна и не требуется для решения многих практических задач, в которых достаточно знать лишь числовые характеристики законов распределения. Например, для грубой характеристики одномерной случайной величины можно ограничиться ее средним значением и величиной разброса возможных значений.

Математическим ожиданием или генеральной средней дискретной случайной величины X называется сумма произведений возможных значений этой величины на соответствующие вероятности:

$$Mx = \sum_{i=1}^N x_i p_i,$$

где p_i - вероятность появления i -го возможного значения x_i случайной величины x ($i = 1, \dots, N$).

Для непрерывной генеральной совокупности X с плотностью $p(x)$ математическое ожидание равно

$$Mx = \int_{-\infty}^{\infty} x p(x) dx,$$

Начальным ν_k и центральным μ_k моментами k -го порядка случайной величины x называются математические ожидания соответственно величин x^k и $(x - Mx)^k$, то есть $Mx_k^k = \nu_k, M(x - Mx)^k = \mu_k$.

На практике бывает достаточно ограничиваться моментами первого и второго порядка. Отсюда следует, что начальный момент первого порядка есть математическое ожидание случайной величины $\nu_1 = Mx$. Дисперсия случайной величины есть центральный момент второго порядка $Dx = M(x - Mx)^2$. Средним квадратическим отклонением называется положительный квадратный корень из ее дисперсии $\sigma = \sqrt{Dx}$.

Приведем названия, обозначения и формулы еще некоторых моментов: $Mx_j^l = M(x_j)^l$ - начальный момент l -го порядка, ма-

тематическое ожидание 1-ой степени j-ой компоненты вектора X;

$$\sigma_{jj} = Mx_j^2 - M(x_j)^2 = M(x_j - Mx_j)^2 = \sigma_j^2,$$

где σ_{jj} - центральный момент второго порядка, дисперсия j-ой компоненты вектора X, $j = 1, 2, \dots, k$;

$$\sigma_{ij} = M((x_i - Mx_i)(x_j - Mx_j)) = M(x_i x_j) - Mx_i Mx_j.$$

где σ_{ij} - центральный момент второго порядка, коэффициент ковариации i-ой и j-ой компонент вектора X, $i, j = 1, 2, \dots, k$ при $i \neq j$.

Ковариационная матрица определяется как математическое ожидание произведения центрированного случайного вектора на этот же транспонированный вектор $\Sigma = M[(x - Mx)(x - Mx)^T]$.

Матрица размерности $k \times k$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} \\ \dots & \dots & \dots & \dots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{pmatrix}$$

называется ковариационной матрицей случайного вектора X. Она симметрическая и неотрицательно определена.

Коэффициент ковариации нормированных случайных величин называется коэффициентом корреляции, или коэффициентом парной корреляции

$$\rho_{ij} = M\left(\frac{(x_i - Mx_i)(x_j - Mx_j)}{\sigma_i \sigma_j}\right) = \frac{\sigma_{ij}}{\sigma_i \sigma_j},$$

где σ_i, σ_j - средние квадратические отклонения случайных величин x_i и x_j .

Квадрат коэффициента корреляции называют коэффициентом детерминации. Так как для любых случайных величин x_i и x_j

$$M\left(\frac{x_i - Mx_i}{\sigma_i} \pm \frac{x_j - Mx_j}{\sigma_j}\right)^2 = 2 \pm 2\rho_{ij} \geq 0,$$

то $-1 \leq \rho_{ij} \leq 1$.

Причем, при $\rho_{ij} = \pm 1$ x_i и x_j связаны линейной функциональной связью. Если же x_i и x_j независимы, то $\rho_{ij} = 0$.

Матрица размерности $k \times k$

$$R = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix}$$

называется корреляционной. Она симметрическая и неотрицательно определена.

1.3. Многомерная нормально распределенная генеральная совокупность

В статистических курсах, где изучались вариационные ряды, мы пользовались данными, полученными в результате измерений, наблюдений, то есть эмпирическим материалом. Их графическое изображение, например, на координатной плоскости принято называть эмпирической кривой распределения. В каждом отдельном случае эмпирическую кривую (в общем случае, гиперплоскость) можно рассматривать как некоторое приближение к соответствующей кривой распределения случайной величины. Степень приближения будет возрастать по мере увеличения числа наблюдений или измерений. Следовательно, имея некоторый эмпирический ряд распределения, мы стремимся описать его при помощи математической модели - закона распреде-

ления, анализируя образование частот в эмпирическом материале. В статистической практике получили распространение, например, биномиальный закон распределения, закон Пуассона и т.д. Нередко при рассмотрении различных моделей статистического анализа предполагается нормальное распределение всех или некоторых признаков генеральной совокупности. Говорят, что непрерывная k -мерная случайная величина распределена нормально, если плотность распределения имеет вид

$$p(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left\{ -1/2 (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

где $\mu = (\mu_1, \mu_2, \dots, \mu_k)^T$ - k -мерный вектор математических ожиданий;

Σ^{-1} - матрица, обратная ковариационной матрице Σ размерности $k \times k$;

$|\Sigma|$ - определитель матрицы Σ .

Таким образом, многомерный нормальный закон распределения определяется вектором математических ожиданий μ и ковариационной матрицей Σ , то есть $k + \frac{k(k+1)}{2}$ параметрами генеральной совокупности.

Покажем, что при $k = 1$ имеет место одномерный нормальный закон распределения. При $k = 1$ $\Sigma = \sigma_{11} = \sigma^2$. Тогда $|\Sigma| = \sigma^2$, а обратная матрица $\Sigma^{-1} = 1/\sigma^2$. Подставляя найденные значения в выражение плотности многомерного распределения имеем

$$p(x) = 1/\sqrt{2\pi\sigma^2} \exp \left\{ -(x - \mu)^2 / 2\sigma^2 \right\}$$

Полученная плотность распределения одномерного нормального закона зависит от 2-х параметров: математического ожидания μ и среднего квадратического отклонения σ .

Графическое изображение этой плотности распределения дает кривую нормального распределения, которая симметрично располагается относительно вертикальной прямой $x = \mu$, поэтому μ называют центром распределения. Рассматривая влияние сред-

него квадратического отклонения на форму кривой, заметим, что чем меньше параметр σ , тем кривая более вытянута около центра распределения. А так как площадь, ограниченная кривой остается равной 1, то вытягивание вверх должно компенсироваться сжатием около центра распределения и более быстрым приближением кривой к оси абсцисс. При большом среднем квадратическом отклонении σ кривая становится более плоской, растягиваясь вдоль оси абсцисс.

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Задачи математической статистики фактически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам выборки. Выборкой из генеральной совокупности называют результаты ограниченного ряда наблюдений $x_1, x_2, \dots, x_n$, где n - объем выборки. Достоверность выводов, получаемых в результате статистической обработки данных, во многом зависит от успешного решения вопроса представительности выборки, то есть полноты и адекватности представления свойств анализируемой генеральной совокупности.

Необходимость выборочного обследования при решении практических задач может быть связана со следующими причинами:

1) генеральная совокупность настолько многочисленна, что проведение обследования всех элементов совокупности слишком трудоемко. С такой ситуацией приходится встречаться при контроле качества продукции крупносерийного и массового производства;

2) при бесконечно большой генеральной совокупности, когда даже весьма большое множество наблюдений не исчерпывает всей совокупности. Например, при разработке статистически обоснованных временных нормативов на изготовление изделия. В этом случае выводы, полученные по результатам конечного числа наблюдений, распространяются на всю генеральную совокупность, охватывающую все детали, которые могут быть изготовлены на оборудовании данного типа на протяжении ряда последующих лет;

3) когда в процессе проведения испытания происходит разрушение отбираемых образцов (например, испытания предела прочности).

2.1. Точечные оценки параметров многомерной генеральной совокупности

Выборку объема n из k -мерной генеральной совокупности X можно представить в виде матрицы данных

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

строки которой рассматриваются как n независимых реализаций k -мерного случайного вектора. Таким образом, элементы x_{ij} матрицы X можно рассматривать либо как случайные (одномерные) величины, независимые по i , либо как конкретные наблюдаемые значения - координаты n точек в k -мерном евклидовом пространстве.

Приведем точечные оценки моментов генеральной совокупности, которые получили наибольшее практическое применение.

Оценка начального момента m -го порядка l -ой компоненты случайного вектора x вычисляется по формуле

$$\bar{x}_l^m = \frac{1}{n} \sum_{i=1}^n x_{il}^m, \quad l=1, k,$$

при $m=1$ имеем среднюю арифметическую

$$\bar{x}_l = \frac{1}{n} \sum_{i=1}^n x_{il}, \quad l=1, k.$$

Оценка ковариационной матрицы Σ случайного вектора X (матрицы выборочных дисперсий и коэффициентов ковариации) определяется как

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1k} \\ s_{21} & s_{22} & \dots & s_{2k} \\ \dots & \dots & \dots & \dots \\ s_{k1} & s_{k2} & \dots & s_{kk} \end{pmatrix}$$

$$s_{lj} = \frac{1}{n} \sum_{i=1}^n (x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j) \quad l, j = 1, k.$$

Причем $s_{11} = s_1^2$ - выборочная дисперсия 1-ой компоненты случайного вектора X ,

s_{lj} - выборочный коэффициент ковариации компонент 1-ой и j -ой вектора X .

Вместо S употребляют также несмещенную оценку матрицы Σ :

$$\hat{S} = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} & \dots & \hat{s}_{1k} \\ \hat{s}_{21} & \hat{s}_{22} & \dots & \hat{s}_{2k} \\ \dots & \dots & \dots & \dots \\ \hat{s}_{k1} & \hat{s}_{k2} & \dots & \hat{s}_{kk} \end{pmatrix} = \frac{n}{n-1} S,$$

$$\hat{s}_{lj} = \frac{1}{n-1} \sum_{i=1}^n (x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j) \quad l, j = 1, k.$$

При алгоритмизации задач многомерного анализа полезным может оказаться вычисление оценок параметров генеральной совокупности с использованием операций над матрицами.

Оценку корреляционной матрицы

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}$$

можно получить по формуле

$$r_{lj} = \frac{s_{lj}}{s_l s_j},$$

где r_{lj} - оценка парного коэффициента корреляции между 1-ой и j -ой компонентами X ($l, j = 1, k$).

Или

$$R = \frac{1}{n} Z^T Z,$$

где

$$Z = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1k} \\ z_{21} & z_{22} & \dots & z_{2k} \\ \dots & \dots & \dots & \dots \\ z_{n1} & z_{n2} & \dots & z_{nk} \end{pmatrix}, \quad z_{lj} = \frac{x_{ij} - \bar{x}_j}{s_j}.$$

2.2. Доверительные области

При малом объеме выборки точечные оценки могут достаточно далеко отклоняться от оцениваемых параметров и поэтому вводится понятие интервальной оценки параметра генеральной совокупности.

Определения и понятия интервального оценивания можно перенести на случай векторного параметра $\theta = (\theta_1, \theta_2, \dots, \theta_k)^T$ с за-

мной доверительного интервала доверительной областью в соответствующем k -мерном пространстве.

Доверительной областью вектора параметров θ генеральной совокупности называется случайная область, полностью определяемая результатами наблюдений, которая с близкой к 1 доверительной вероятностью (надежностью) γ содержит неизвестное значение вектора θ .

Очевидно, что существует бесконечное множество доверительных областей, соответствующих одному и тому же значению γ . Обычно стараются определить доверительные области, имеющие минимальные размеры при данной надежности γ . Часто этому условию удовлетворяют области, симметрические относительно вектора оценок $\hat{\theta}$ параметров θ .

Основную трудность в построении доверительной области представляет определение законов распределения подходящих статистик. В настоящее время эти вопросы хорошо разработаны только для нормального распределения наблюдаемых случайных величин.

2.3. Доверительная область для вектора математических ожиданий

Пусть по результатам n наблюдений из генеральной совокупности X с k -мерным нормальным распределением $N_k(\mu, \Sigma)$ найдены вектор средних $\bar{x}$ и несмещенная оценка матрицы ковариаций S . Требуется найти с надежностью γ доверительную область для k -мерного вектора генеральных средних μ .

Предположим, что ковариационная матрица известна. Найдем такую подходящую статистику, чтобы ее распределение было известным и по которой однозначно можно определить доверительную область.

Напомним, что для одномерной нормально распределенной генеральной совокупности доверительный интервал для μ определяется из формул

$$\left| t \right| < \Phi^{-1}(\gamma) t = \frac{\bar{x} - \mu}{\sigma} \sqrt{n},$$

где статистика t подчиняется стандартному закону $N_1(0, 1)$. Последнее равенство можно переписать в виде

$$t^2 = n(\bar{x} - \mu)^T \left(\sigma^2 \right)^{-1} (\bar{x} - \mu).$$

Оно обобщается на случай k -мерной совокупности следующим образом:

$$\chi^2 = n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu).$$

Известно, что статистика t^2 при $k=1$ имеет хи-квадрат распределение с числом степеней свободы равным 1. Статистика χ^2 для любого k также распределена по хи-квадрат с числом степеней свободы $\nu=k$.

Таким образом, с надежностью γ можно утверждать, что вектор μ покрывается доверительной областью, задаваемой неравенством

$$n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \leq \chi^2_{1-\gamma}.$$

Пусть теперь ковариационная матрица неизвестна. Чтобы при $k=1$ построить доверительный интервал для μ , используют статистику

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

которая имеет t -распределение Стьюдента с $n-1$ степенями свободы. Равенство можно переписать в эквивалентной форме

$$t^2 = n(\bar{x} - \mu)^T \left(\frac{s^2}{n} \right)^{-1} (\bar{x} - \mu).$$

По аналогии строится статистика T^2 Хотеллинга, которую используют при построении доверительной области для вектора средних μ k -мерной генеральной совокупности:

$$T^2 = n(\bar{x} - \mu)^T \hat{S}^{-1} (\bar{x} - \mu),$$

где $\hat{S}^{-1}$ - матрица, обратная матрице несмещенных оценок ковариационной матрицы $\hat{S}$.

Учитывая, что F и T^2 распределения связаны соотношением

$$T^2_{\alpha, k, n-k} = \frac{k(n-1)}{n-k} F_{\alpha, k, n-k},$$

получим уравнение поверхности, ограничивающей доверительную область k генеральных средних с надежностью γ

$$n(\bar{x} - \mu)^T \hat{S}^{-1} (\bar{x} - \mu) = \frac{k(n-1)}{n-k} F_{\alpha, k, n-k}.$$

где $F_{\alpha, k, n-k}$ - точка F -распределения, соответствующая уровню значимости α и числам степеней свободы k и $n-k$.

Два последних уравнения определяют k -мерный эллипсоид (эллипс при $k = 2$) с центром $\bar{x}$, так как их левая часть представляет собой положительно определенную квадратичную форму относительно μ .

3. ГИПОТЕЗЫ О ПАРАМЕТРАХ МНОГОМЕРНОЙ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

3.1. Сравнение вектора генеральных средних со стандартом

Рассматривается k -мерная генеральная совокупность распределения $N_k(\mu, \Sigma)$, $|\Sigma| \neq 0$. По выборке объема n из нее определим вектор средних арифметических и несмещенная оценка ковариационной матрицы.

Если ковариационная матрица известна, то для проверки гипотезы о равенстве вектора генеральных средних стандартному (заданному) значению, $H_0: \mu = \mu_0$ против альтернативы $H_1: \mu \neq \mu_0$ употребляется статистика

$$\chi^2 = n(\bar{x} - \mu_0)^T \Sigma^{-1} (\bar{x} - \mu_0),$$

имеющая распределение хи-квадрат Пирсона с числом степеней свободы $v = k$ при справедливости гипотезы H_0 .

Если же ковариационная матрица Σ неизвестна, то можно воспользоваться статистикой Хотеллинга

$$T^2 = n(\bar{x} - \mu_0)^T \hat{S}^{-1} (\bar{x} - \mu_0).$$

Ранее уже было сказано, что при истинности H_0 имеет место равенство

$$T^2_{\alpha, k, n-k} = \frac{k(n-1)}{n-k} F_{\alpha, k, n-k}.$$

Поэтому для критической области

$$T^2 > T^2_{hp} = T^2_{\alpha, k, n-k}$$

можно вычислить T^2_{hp} с помощью таблиц F-распределения Фишера-Снедекора.

Следует отметить, что применяемые статистики являются обобщениями соответствующих статистик для одномерной случайной величины. Это следует из того факта, что квадрат нормированной нормально распределенной одномерной случайной величины имеет распределение хи-квадрат со степенью свободы равной 1, а квадрат случайной величины, имеющей t-распределение Стьюдента, распределен по F-распределению с числом степеней свободы 1 и n-1.

3.2. Сравнение двух генеральных совокупностей

Будем называть две генеральные совокупности однородными, если кроме одних и тех же признаков эти совокупности имеют одинаковые законы распределения вероятностей.

Рассмотрим две нормально распределенные совокупности X и Y. Их распределения полностью задаются параметрами μ_x , Σ_x и μ_y , Σ_y . Следовательно, для проверки однородности этих совокупностей достаточно сравнить их ковариационные матрицы Σ_x и Σ_y . Затем, в случае принятия гипотезы о равенстве этих ковариационных матриц, сравнить генеральные средние совокупностей μ_x и μ_y .

Для сравнения матриц генеральных коэффициентов ковариации проверяется гипотеза $H_0: \Sigma_x = \Sigma_y$ против альтернативы $H_1: \Sigma_x \neq \Sigma_y$ с уровнем значимости α на основе выборок из совокупностей соответственно объемов n_x и n_y .

В качестве статистики критерия проверки берется случайная величина

$$W = ba,$$

где

$$b = 1 - \left(\frac{1}{n_x - 1} + \frac{1}{n_y - 1} - \frac{1}{n_x + n_y - 2} \right) \frac{2k^2 + 3k - 1}{6(k+1)};$$

$$a = (n_x + n_y - 2) \ln \left| \hat{S}_{xy} \right| - \left[(n_x - 1) \ln \left| \hat{S}_x \right| + (n_y - 1) \ln \left| \hat{S}_y \right| \right];$$

$\hat{S}_x$ - несмещенная оценка ковариационной матрицы с элементами

$$\hat{s}_{x_{ml}} = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_{im} - \bar{x}_m)(x_{il} - \bar{x}_l), m, l = 1, k;$$

$\hat{S}_y$ - несмещенная оценка ковариационной матрицы с элементами

$$\hat{s}_{y_{ml}} = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (y_{im} - \bar{y}_m)(y_{il} - \bar{y}_l), m, l = 1, k;$$

$$\hat{S}_{xy} = \frac{1}{n_x + n_y - 2} \left[(n_x - 1) \hat{S}_x + (n_y - 1) \hat{S}_y \right]$$

несмещенная оценка одной и той же ковариационной матрицы $\Sigma_x = \Sigma_y$.

При справедливости гипотезы H_0 , достаточно больших объемах выборок n_x и n_y и достаточно малой величине

$$c = \frac{k(k+1)}{48b^2} \left\{ (k-1)(k+2) \left[\frac{1}{(n_x-1)^2} + \frac{1}{(n_y-1)^2} + \frac{1}{(n_x-1)^2} + \frac{1}{(n_y-1)^2} \right] - 6(1-b)^2 \right\}$$

статистика W аппроксимируется распределением хи-квадрат с числом степеней свободы $\nu = \frac{k(k+1)}{2}$.

Таким образом, критическая область имеет вид

$$W > W_{hp} = \chi^2_{\alpha; \frac{k(k+1)}{2}}$$

Если $W_{набл} = ba$ попадает в критическую область ($W_{набл} > W_{кр}$), то гипотеза $H_0: \Sigma_x = \Sigma_y$ отвергается с вероятностью ошибки α . Тогда считается доказанным, что ковариационные матрицы Σ_x и Σ_y неодинаковы и, следовательно, генеральные совокупности неоднородны.

Если $W_{набл}$ не попало в критическую область ($W_{набл} \leq W_{кр}$), то гипотеза $H_0: \Sigma_x = \Sigma_y$ не отвергается. Так как других выборок не производится, то гипотеза принимается, то есть считается, что ковариационные матрицы Σ_x и Σ_y одинаковы. При таком условии можно сравнивать генеральные средние, то есть проверять гипотезу $H_0: \mu_x = \mu_y$ против альтернативы $H_1: \mu_x \neq \mu_y$ с уровнем значимости α .

Для проверки применяется статистика Хотеллинга вида

$$T^2 = \frac{n_x n_y}{n_x + n_y} (x - y)^T S_{xy}^{-1} (x - y).$$

Если гипотеза справедлива, то статистики T^2 и F связаны формулой

$$T^2_{hp} = \frac{k(n_x + n_y - 2)}{n_x + n_y - k - 1} F_{\alpha; k; n_x + n_y - k - 1},$$

где $F_{\alpha; k; n_x + n_y - k - 1}$ находится по таблицам распределения Фишера-Снедекора с числом степеней свободы в числителе $\nu_1 = k$ и в знаменателе $\nu_2 = n_x + n_y - k - 1$.

Критическая область имеет вид $T^2 > T^2_{кр}$.

Если гипотеза $H_0: \mu_x = \mu_y$ отвергается с вероятностью ошибки α , то считается доказанной неоднородность генеральных совокупностей X и Y . Если же она не отвергается, то, принимая ее, мы считаем, что генеральные совокупности однородны с надежностью $\gamma = 1 - \alpha$.

4. ОСНОВЫ КЛАСТЕРНОГО АНАЛИЗА

Кластерный анализ - это совокупность методов, позволяющих классифицировать многомерные наблюдения, каждое из которых описывается набором исходных переменных $X_1, X_2, \dots, X_n$. Целью кластерного анализа является образование групп схожих между собой объектов, которые принято называть кластерами. Слово кластер (cluster) английского происхождения, переводится как сгусток, пучок, группа. Родственные понятия, используемые в литературе, - класс, таксон, сгущение.

В отличие от комбинационных группировок кластерный анализ приводит к разбиению на группы с учетом всех группировочных признаков одновременно. Например, если каждый наблюдаемый объект характеризуется двумя признаками X_1 и X_2 , то при выполнении комбинационной группировки вся совокупность объектов будет разбита на группы по X_1 , а затем внутри каждой выделенной группы по X_2 . Такой подход получил название монотетического. Определить принадлежность каждого объекта к той или иной группе можно, последовательно сравнивая его значения X_1 и X_2 с границами выделенных групп. Образование группы всегда связано в этом случае с указанием ее границ по каждому группировочному признаку отдельно.

В кластерном анализе используется иной принцип образования групп, так называемый политетический подход. Все группировочные признаки одновременно участвуют в группировке, то есть они учитываются все сразу при отнесении наблюдения в ту или иную группу. При этом, как правило, не указаны четкие границы каждой группы, а также неизвестно заранее, сколько же групп целесообразно выделить в исследуемой совокупности.

Кластерный анализ - одно из направлений статистического исследования. Особо важное место он занимает в тех отраслях науки, которые связаны с изучением массовых процессов и явлений. Необходимость развития методов кластерного анализа и их использования продиктована прежде всего тем, что они помогают построить научно обоснованные классификации, выявить внутренние связи между единицами наблюдаемой совокупности.

Кроме того, методы кластерного анализа могут использоваться с целью сжатия информации, что является важным фактором в условиях постоянного увеличения и усложнения потоков статистических данных.

Первые публикации по кластерному анализу появились в конце 30-х годов нашего столетия, но активное развитие этих методов и их широкое использование началось в конце 60-х - начале 70-х годов. В дальнейшем это направление многомерного анализа очень интенсивно развивалось. Появились новые методы, новые модификации уже известных алгоритмов, существенно расширилась область применения кластерного анализа. Если первоначально методы многомерной классификации использовались в психологии, археологии, биологии, то сейчас они стали активно применяться в социологии, экономике, статистике, в исторических исследованиях. Особенно расширилось их использование в связи с появлением и развитием ЭВМ. Это связано прежде всего с трудоемкостью обработки больших массивов информации (вычисление и обращение матриц больших размерностей).

Методы кластерного анализа позволяют решать следующие задачи:

- проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи, как правило, приводит к углублению знаний о совокупности классифицируемых объектов;
 - проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, то есть поиск существующей структуры;
 - построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привести в нее структуру.
- Методы кластерного анализа можно разделить на две большие группы: агломеративные (объединяющие) и дивизимные (разделяющие). Агломеративные методы последовательно объединяют отдельные объекты в группы (кластеры), а дивизимные методы расчленяют группы на отдельные объекты. В свою очередь каждый метод как объединяющего, так и разделяющего типа

может быть реализован при помощи различных алгоритмов. В частности наиболее доступным и поэтому наиболее популярным можно считать иерархический агломеративный кластерный анализ. Следует заметить, что как агломеративные, так и дивизимные алгоритмы трудоемки, и их сложно использовать для больших исследуемых совокупностей. Результаты реализации таких алгоритмов для большего удобства их визуального анализа принято изображать графически в виде дендрограммы (дерева).

В кластерном анализе существуют также методы, которые трудно отнести к первой или ко второй группе. Это так называемые итеративные методы. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.). Итеративные методы в большей степени, чем иерархические, требуют от пользователя интуиции при выборе типа классификационных процедур и задания начальных условий разбиения, так как большинство этих методов очень чувствительны к изменению задаваемых параметров. Например, выбранное случайным образом число кластеров может не только сильно увеличить трудоемкость процесса классификации, но и привести к образованию «размытых» или мало наполняемых кластеров. Поэтому целесообразно сначала провести классификацию по одному из иерархических методов или на основании экспертных оценок, а затем уже подбирать начальное разбиение и статистический критерий для работы итерационного алгоритма.

В отличие от агломеративных и дивизимных методов итеративные алгоритмы могут привести к образованию пересекающихся кластеров, когда один объект может одновременно принадлежать нескольким кластерам.

Один из таких методов – *метод k-средних* – принадлежит к группе итеративных методов эталонного типа. Само название метода было предложено Дж. Мак-Куином в 1967 г.

В отличие от иерархических процедур метод k-средних не требует вычисления и хранения матриц расстояний между объектами. Алгоритм этого метода предполагает использование только

исходных значений переменных. Для начала процедуры классификации должны быть заданы k случайно выбранных объектов, которые будут служить эталонами, то есть центрами кластеров. Считается, что алгоритмы эталонного типа достаточно удобны и быстродейственны. В этом случае важную роль играет выбор начальных условий, которые влияют на длительность процесса классификации и на его результаты.

Еще одним из итеративных методов классификации является метод поиска сгущений. Он не требует начального задания числа кластеров.

В теории и на практике существует несколько различных модификаций этого метода. Каждая модификация отличается задаваемым начальным состоянием и критериями завершения классификации. Например, суть итеративного алгоритма типа «фо-рель» заключается в применении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков с целью поиска локальных сгущений точек. Схема данного алгоритма в общем виде выглядит следующим образом.

Метод поиска сгущения требует вычисления матрицы расстояний между объектами. Затем выбирается объект, который является первоначальным центром первого кластера. Выбор такого объекта может быть произвольным, а может основываться на предварительном анализе точек и их окрестностей. При использовании второго подхода можно значительно сократить число итераций, приводящих к распределению всех точек по кластерам.

Выбранная точка принимается за центр гиперсферы заданного радиуса R . Определяется совокупность точек, попавших внутрь этой сферы, и для них вычисляются координаты центра (вектор средних значений признаков). Далее вновь рассматривается гиперсфера того же радиуса, но с новым центром, и для совокупности попавших в нее точек опять рассчитывается вектор средних значений, он принимается за новый центр сферы и т.д. Когда очередной пересчет координат центра сферы приводит к такому же результату, как и на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер

и из дальнейшего процесса кластеризации исключаются. Для всех оставшихся точек процедуры повторяются, то есть опять выбирается произвольный объект, который является первоначальным центром сферы радиуса R , и т.д. Существуют доказательства сходимости этого алгоритма для любых начальных точек и для совокупностей различного объема. Таким образом, работа алгоритма завершается за конечное число шагов и все точки оказываются распределенными по кластерам. Число образовавшихся кластеров заранее неизвестно и сильно зависит от выбора радиуса сферы.

Обычной формой представления исходных данных в задачах кластерного анализа служит прямоугольная матрица:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

каждая строка которой представляет результат измерений k рассматриваемых признаков на одном из n обследованных объектов. В конкретных ситуациях может представлять интерес как группировка объектов, так и группировка признаков. В тех случаях, когда разница между этими двумя задачами не существенна, например, при описании некоторых алгоритмов, будем пользоваться только термином «объект», включая в это понятие и «признак».

Матрица X не является единственным способом представления данных в задачах кластерного анализа. Иногда исходная информация задана в виде квадратной матрицы:

$$R = (r_{ij}), i, j = 1, 2, \dots, k.$$

элемент r_{ij} которой определяет степень близости i -го объекта к j -му объекту.

Большинство алгоритмов кластерного анализа полностью исходит из матрицы расстояний (или близостей), либо требует

вычисления отдельных ее элементов, поэтому, если данные представлены в форме X , то первым этапом решения задачи поиска кластеров будет выбор способа вычисления расстояний, или близости, между объектами или признаками.

Относительно проще решается вопрос об определении близости между признаками. Как правило, кластерный анализ признаков преследует те же цели, что и факторный анализ - выделение групп связанных между собой признаков, отражающих определенную сторону изучаемых объектов. Мерами близости в этом случае служат различные статистические коэффициенты связи.

4.1. Расстояние между объектами (кластерами) и мера близости

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов. В общем случае, понятие однородности объектов задается либо введением правила вычисления расстояний $\rho(x_i, x_j)$ между любой парой исследуемых объектов $(x_1, x_2, \dots, x_n)$, либо заданием некоторой функции $\tau(x_i, x_j)$, характеризующей степень близости i -го и j -го объектов.

Выбор метрики или меры близости является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы при данном алгоритме разбиения. В каждом конкретном случае этот выбор должен производиться по-своему, в зависимости от целей исследования, физической и статистической природы вектора наблюдений X , априорных сведений о характере вероятностного распределения X .

Рассмотрим наиболее широко используемые в задачах кластерного анализа расстояния и меры близости.

Обычное Евклидово расстояние

$$\rho_f(x_i, x_j) = \sqrt{\sum_{l=1}^k (x_{il} - x_{jl})^2},$$

где x_{il}, x_{jl} - величина l -ой компоненты у i -го (j -го) объекта ($i, j = 1, 2, \dots, k; i, j = 1, 2, \dots, n$).

Использование этого расстояния оправдано в следующих случаях:

а) наблюдения берутся из генеральной совокупности имеющей многомерное нормальное распределение с ковариационной матрицей вида $\sigma^2 E_k$, то есть компоненты X взаимно независимы и имеют одну и ту же дисперсию, где E_k - единичная матрица;

б) компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

Естественное, с геометрической точки зрения, евклидово пространство может оказаться бессмысленным (с точки зрения содержательной интерпретации), если признаки измерены в разных единицах. Чтобы исправить положение прибегают к нормированию каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы X к нормированной матрице с элементами:

$$z_{il} = \frac{x_{il} - x_l}{s_l},$$

где z_{il} - значение l -го признака у i -го объекта;

x_l - среднее значение l -го признака;

$$s_l = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{il} - x_l)^2},$$

среднее квадратическое отклонение l -го признака.

Однако эта операция может привести к нежелательным последствиям. Если кластеры хорошо разделены по одному признаку и не разделены по другому, то после нормирования дискриминирующие возможности первого признака будут уменьшены в связи с увеличением «шумового» эффекта второго.

«Взвешенное» Евклидово расстояние

$$\rho_{ef}(x_i, x_j) = \sqrt{\sum_{l=1}^k \omega_l (x_{il} - x_{jl})^2}$$

применяется в тех случаях, когда каждой компоненте x_l вектора наблюдений X удастся приписать некоторый «вес» ω_l пропорционально степени важности признака в задаче классификации. Обычно принимают $0 \leq \omega_l \leq 1$, где $l = 1, 2, \dots, k$. Определение «веса», как правило, связано с дополнительными исследованиями, например, организацией опроса экспертов и обработкой их мнений.

Хеммингово расстояние используется как мера различия объектов, задаваемых дихотомическими признаками. Это расстояние определяется по формуле:

$$\rho_X(x_i, x_j) = \sum_{l=1}^k |x_{il} - x_{jl}|$$

и равно числу несовпадений значений соответствующих признаков в рассматриваемых i -м и j -м объектах.

Как правило, решение задач классификации многомерных данных предусматривает в качестве предварительного этапа ис-

следования реализацию методов, позволяющих выбрать из компонент наблюдаемого вектора X сравнительно небольшое число наиболее информативных, то есть уменьшить размерность наблюдаемого пространства. В ряде процедур классификации (кластер-процедур) используют понятия расстояния между группами объектов и меры близости двух групп объектов.

Пусть s_i - i -ая группа (кластер), состоящая из n_i объектов, x_i - среднее арифметическое векторных наблюдений s_i группы, то есть «центр тяжести» i -ой группы: $\rho(s_i, s_m)$ - расстояние между группами s_i и s_m .

Наиболее употребительными расстояниями и мерами близости между классами объектов являются:

- расстояние, измеряемое по принципу «ближайшего соседа»

$$\rho_{\min}(s_i, s_m) = \min_{x_i \in s_i, x_j \in s_m} \rho(x_i, x_j);$$

- расстояние, измеряемое по принципу «дальнего соседа»

$$\rho_{\max}(s_i, s_m) = \max_{x_i \in s_i, x_j \in s_m} \rho(x_i, x_j);$$

- расстояние, измеряемое по «центрам тяжести» групп

$$\rho_b(s_i, s_m) = \rho(x_i, x_m);$$

- расстояние, измеряемое по принципу «средней связи»

$$\rho_p(s_i, s_m) = \frac{1}{n_i n_m} \sum_{x_i \in s_i} \sum_{x_j \in s_m} \rho(x_i, x_j).$$

Академиком А.Н.Колмогоровым было предложено «обобщенное расстояние» между классами, которое включает в себя в качестве частных случаев все рассмотренные выше виды расстояний.

Расстояние между группами элементов особенно важно в так называемых агломеративных иерархических кластер-процедурах, так как принцип работы таких алгоритмов состоит в последовательном объединении элементов, а затем и целых групп, сначала самых близких, а затем все более и более отдаленных друг от друга. При этом расстояние между классом s_l и классом $s_{(m, q)}$, являющимся объединением двух других классов s_m и s_q , можно определить по формуле:

$$\rho_{l(m, q)} = \rho(s_l, s_{(m, q)}) = \alpha \rho_{lm} + \beta \rho_{lq} + \gamma \rho_{mq} + \delta |\rho_{lm} - \rho_{lq}|,$$

$$\text{где } \rho_{lm} = \rho(s_l, s_m); \quad \rho_{lq} = \rho(s_l, s_q)$$

и $\rho_{mq} = \rho(s_m, s_q)$ - расстояния между классами s_l , s_m и s_q ;

$\alpha, \beta, \gamma, \delta$ - числовые коэффициенты, значения которых определяют специфику процедуры, ее алгоритм.

Например, при $\alpha = \beta = -\delta = 1/2$ и $\gamma = 0$ приходим к расстоянию, построенному по принципу «ближайшего соседа». При $\alpha = \beta = \delta = 1/2$ и $\gamma = 0$ - расстояние между классами определяется по принципу «дальнего соседа», то есть как расстояние между двумя самыми дальними элементами этих классов.

$$\text{И, наконец, при } \alpha = \frac{n_m}{n_m + n_q}; \beta = \frac{n_q}{n_m + n_q}; \gamma = \delta = 0$$

рассматриваемое соотношение приводит к расстоянию ρ_{cp} между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой из другого.

4.2. Функционалы качества разбиения

При использовании различных методов кластерного анализа для одной и той же совокупности могут быть получены различ-

ные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают, во-первых, набор признаков, по которым осуществляется классификация, во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Если используются методы с эталонными алгоритмами, например, метод k-средних, то задаваемые начальные условия разбиения в значительной степени определяют конечный результат разбиения.

После завершения процедур классификации необходимо оценить полученные результаты. Для этой цели используется некоторая мера качества классификации, которую принято называть функционалом или критерием качества разбиения исследуемых объектов на кластеры. Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается экстремальное (минимальное или максимальное) значение целевой функции – функционала качества.

Рассмотрим наиболее распространенные функционалы качества разбиения.

1. Сумма квадратов расстояний до центров кластеров:

$$2. F_1 = \sum_{l=1}^k \sum_{i \in S_l} \rho^2(x_i, \bar{x}_l),$$

где l – номер кластера ($l = 1, 2, \dots, k$), $\bar{x}_l$ – центр l -го кластера, x_i – вектор значений переменных для i -го объекта, входящего в l -ый кластер.

При использовании этого критерия стремятся получить такое разбиение совокупности объектов на k кластеров, при котором значение F_1 было бы минимальным.

3. Сумма внутриклассовых расстояний между объектами:

$$F_2 = \sum_{l=1}^k \sum_{i, j \in S_l} \rho^2(ij).$$

В этом случае следует считать наилучшим такое разбиение, при котором достигается минимальное значение F_2 , то есть получены кластеры большой «плотности». Объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

4. Суммарная внутриклассовая дисперсия:

$$F_3 = \sum_{l=1}^k \sum_{j=1}^p \sigma_{lj}^2,$$

где σ_{lj}^2 – дисперсия j -ой переменной в кластере S_l .

В случае использования данного критерия разбиение, при котором сумма внутриклассовых (внутригрупповых) дисперсий будет минимальной, следует считать оптимальным.

Если оценивать качество разбиения по степени удаленности кластеров друг от друга, то можно использовать функционал F_4 – средние межклассовые расстояния:

$$F_4 = \frac{\sum_{i \in S_q} d_{ij}}{\sum_{l < q} n_l n_q} - \max.$$

4.3. Иерархические кластер-процедуры

Иерархические (древовидные) процедуры являются наиболее распространенными (в смысле реализации на ЭВМ) алгоритмами кластерного анализа. Как подчеркивалось ранее, они бывают двух типов: агломеративные и дивизимные. В агломеративных процедурах начальным является разбиение, состоящее из n одноэлементных классов, а конечным – из одного класса; в дивизимных – наоборот.

Принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов, сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства).

К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. Алгоритмы требуют вычисления на каждом шаге матрицы расстояний, а, следовательно, емкой машинной памяти и большого количества времени. В этой связи реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, нецелесообразна, а в ряде случаев и невозможна.

В качестве примера приведем агломеративный иерархический алгоритм. На первом шаге алгоритма каждое наблюдение x_i ($i = 1, 2, \dots, n$) рассматривается как отдельный кластер. В дальнейшем, на каждом шаге работы алгоритма, происходит объединение двух самых близких кластеров, и с учетом принятого расстояния по формуле пересчитывается матрица расстояний, размерность которой, очевидно, снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс.

Большинство программ, реализующих алгоритм иерархической классификации, предусматривает графическое представление результатов классификации в виде дендрограммы.

5. ДИСПЕРСИОННЫЙ АНАЛИЗ

5.1. Сущность метода

С помощью дисперсионного анализа изучается степень влияния одного или нескольких факторных признаков на результативный признак, то есть решается задача аналогичная корреляционному анализу. Отличие дисперсионного анализа состоит в том, что в ходе его изучается вариация лишь одного признака - результативного, а простейшим показателем вариации служит дисперсия, которая, как известно из правил сложения дисперсий, может быть разложена на межгрупповую и внутригрупповую.

Межгрупповая дисперсия возникает под действием какого-либо фактора, который приводит к разным величинам средних, имеющихся в отдельных группах, и отражает вариацию этих средних. Внутригрупповая дисперсия возникает под действием прочих факторов, которые можно назвать случайными; она отражает вариацию внутри групп. При разложении общей дисперсии на две ее составляющие можно выделить ту часть, которая обусловлена действием одного, затем другого, третьего и т.д. факторов в общей дисперсии, то есть получить возможность измерять их действие.

Изучение влияния факторов по их дисперсиям называется дисперсионным анализом. Суть метода состоит в разложении общей вариации изучаемого показателя на части, соответствующие раздельному и совместному влиянию факторов, и статистическом изучении этих частей с целью выяснения приемлемости гипотез о существовании этих влияний. Модели дисперсионного анализа в зависимости от числа факторов классифицируются на однофакторные, двухфакторные и т.д. По цели исследования выделяют следующие модели: детерминированная (M1) - уровни всех факторов заранее фиксированы и проверяют именно их влияние; случайная (M2) - уровни каждого фактора получены как случайная выборка из генеральной совокупности уровней фактора; смешанная (M3) - уровни одних факторов заранее фиксированы, а уровни других - случайная выборка.

В дисперсионном анализе измерение влияния факторов ведется с помощью дисперсий. При этом нужно различать дисперсию как сумму квадратов отклонений всех вариантов от средней и дисперсию степеней свободы, когда эта сумма делится на число степеней свободы.

Общая дисперсия признака (дисперсия всего комплекса) представляет собой сумму квадратов отклонений всех вариантов результативного признака от общей его средней:

$$D_y^2 = \sum \sum (y_{ij} - \bar{y})^2$$

где y_{ij} - отдельные значения результативного признака;

$\bar{y}$ - их общая средняя (средняя комплекса).

Общая дисперсия всегда больше дисперсии, показывающей влияние изучаемых факторов, поскольку в одном исследовании невозможно освободиться от действия множества других факторов, влияющих на результативный признак. При разложении общей дисперсии выделяются факторная дисперсия, идентичная межгрупповой, и случайная, или остаточная, дисперсия, вызванная не учитываемыми в данном опыте факторами, идентичная внутригрупповой.

Факторная дисперсия - это сумма квадратов отклонений частных (групповых) средних от общей средней, умноженных на число единиц в каждой группе (n_x):

$$D_x^2 = \sum (\bar{y}_j - \bar{y})^2 n_x$$

где $\bar{y}_j$ - частные (групповые) средние.

Случайная дисперсия - это сумма групповых сумм квадратов отклонений всех вариантов результативного признака в группах от средних значений признака в них:

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

Общая дисперсия представляет собой сумму факторной и случайной дисперсий:

$$D_y^2 = D_x^2 + D_z^2$$

так как в дисперсионном анализе используется свойство суммы квадратов отклонений от средней арифметической, состоящее в том, что при действии различных факторов на результативный признак сумма факторных дисперсий, измеряющая вариацию под влиянием отдельных факторов, равна общей дисперсии, характеризующей вариацию под влиянием их комплекса:

$$D_A^2 + D_B^2 + D_C^2 + \dots = D_y^2$$

где $D_A^2, D_B^2, D_C^2, \dots$ - дисперсии под воздействием факторов А, В, С и т.д.

Отношение факторной дисперсии к общей показывает долю вариации под воздействием изучаемого фактора в общей вариации, то есть степень статистического влияния факторов на результативный признак:

$$\eta_x^2 = \frac{D_x^2}{D_y^2}; \quad \eta_z^2 = \frac{D_z^2}{D_y^2}$$

где η_x^2 - влияние рассматриваемых факторов;

η_z^2 - влияние случайных величин.

Причем $\eta_x^2 + \eta_z^2 = 1$.

Дисперсия как показатель вариации зависит от числа единиц в группе. Для определения степени влияния факторов, найденных по выборочным данным, нужен показатель, свободный от подобного влияния и позволяющий сравнивать группы с разной численностью. Такими показателями служат оценки

дисперсии, вычисляемые со степень свободы вариации:
оценка общей дисперсии

$$s_y^2 = \frac{D_y^2}{k_y} = \frac{\sum \sum (y_{ij} - \bar{y})^2}{k_y};$$

оценка факторной дисперсии

$$s_x^2 = \frac{D_x^2}{k_x} = \frac{\sum (\bar{y}_j - \bar{y})^2 n_x}{k_x};$$

оценка случайной дисперсии

$$s_z^2 = \frac{D_z^2}{k_z} = \frac{\sum \sum (y_{ij} - \bar{y}_j)^2}{k_z},$$

где k_y - число степеней свободы по комплексу, $k_y = n-1$;

k_x - число степеней свободы для факторной дисперсии
 $k_x = r-1$;

k_z - число степеней свободы случайной дисперсии
 $k_z = n-r$.

При этом $k_y = k_x + k_z$.

Для определения достоверности вывода, сделанного по выборочному обследованию исчисляется отношение оценок факторной дисперсии к случайной, каждая из которых рассчитана со степенями свободы, то есть $F = s_x^2 : s_z^2$. Если F равно или больше стандартной величины, которая определяется с разной

вероятностью по таблицам, то вывод на основании корреляционного отношения достоверен. F табличное, в свою очередь, зависит от числа степеней свободы, большей и меньшей из оценок факторной и случайной дисперсий.

В дисперсионном анализе при изучении влияния одного фактора решается однофакторный комплекс, при изучении влияния двух факторов - двухфакторный комплекс и т.д.

При дисперсионном анализе комплексов решаются следующие вопросы: имеют ли влияние изучаемые факторы на результативный признак; какова роль отдельного фактора и их сочетания; можно ли наблюдаемое в выборке влияние данного фактора (или многих факторов) перенести на всю генеральную совокупность в соответствующих условиях или определить достоверность вывода о влиянии факторов.

Представим общую схему дисперсионного анализа:

- 1) нахождение общей дисперсии комплекса;
- 2) определение факторной дисперсии;
- 3) вычисление корреляционного отношения, показывающего степень статистического влияния фактора (факторов);
- 4) нахождение случайной дисперсии;
- 5) проверка равенства суммы факторной и случайной дисперсий общей дисперсии: $D_y^2 = D_x^2 + D_z^2$;
- 6) определение числа степеней свободы вариации для дисперсии комплекса, факторной и случайной дисперсий (проследить, чтобы $k_y = k_x + k_z$);

7) вычисление оценок этих дисперсий со степенями свободы;

8) расчет величины F ;

9) определение табличных значений F по приложениям;

10) сравнение табличных и расчетных значений F . Вывод о достоверности заключения о влиянии факторов. При этом необходимо, чтобы при данных числах степеней свободы k_x и k_z и при заданной вероятности того, что различие вычисленных оценок дисперсий неслучайно, отношение вычисленных

дисперсий $s_x^2:s_z^2$ было бы равно или больше такого же отношения, найденного по таблицам.

При решении двух-трех факторных комплексов D_x^2 представляет собой дисперсию суммарного действия для изучаемых факторов. Она разлагается на частные факторные дисперсии и дисперсии по сочетанию факторов. Если первые получаются как суммы квадратов отклонений частных средних от общей средней, взвешенных по численностям групп, то вторые можно рассчитать как разность между факторной дисперсией и суммой частных дисперсий. Например, при изучении действия двух факторов А и В получим:

$$D_A^2 = \sum (\bar{y}_A - \bar{y})^2 n_A; D_B^2 = \sum (\bar{y}_B - \bar{y})^2 n_B;$$

$$D_{AB}^2 = D_x^2 - D_A^2 - D_B^2.$$

Величина D_x^2 находится как сумма квадратов отклонений средних групп, выделенных по всем факторам, от общей средней:

$$D_x^2 = \sum (\bar{y}_i - \bar{y})^2 n_x.$$

Действие каждого фактора определяется соответствующим отношением:

$$\eta_A^2 = \frac{D_A^2}{D_y^2}; \quad \eta_B^2 = \frac{D_B^2}{D_y^2}; \quad \eta_{AB}^2 = \frac{D_{AB}^2}{D_y^2}.$$

Дисперсионный анализ позволяет определить также достоверность вывода о действии каждого фактора по отношению дисперсии со степенью свободы по этому фактору к случайной дисперсии со степенью свободы по всему комплексу и на основании сравнения этого отношения с табличным по описанной уже схеме.

5.2. Правила организации дисперсионного комплекса

Чтобы получить достоверные результаты при дисперсионном анализе, нужно выполнять правила организации дисперсионного комплекса. Необходимо также знать способы его расчета.

Выбор факторов. Если организуется однофакторный комплекс, теоретически можно изучить влияние любого признака на результативный, лишь была бы возможность разделить значения результативного признака на группы по факторному признаку. Это могут быть признаки того же явления. Например, при изучении объема продукции предприятий стоимость их основных фондов, численность персонала и др.

Если организуется двух- или трехфакторный комплекс, то к отбору этих факторов предъявляется еще требование полной независимости друг от друга. Для таких комплексов нельзя брать, например, изучение влияния основных фондов и машино-вооруженности, поскольку эти факторы взаимосвязаны. Независимыми факторами при изучении продукции предприятий будут, например, стоимость основных фондов и квалификация работников.

Группировка по факторам. Проведение дисперсионного анализа требует разделения совокупности единиц, обладающих определенной величиной результативного признака, на группы по факторному признаку. Следовательно, факторный признак выступает здесь в качестве основания группировки.

Как в однофакторном, так и в многофакторном комплексе группировка по факторному признаку может быть количественной и качественной. Например, можно выделять предприятия крупные, средние, мелкие; выполняющие нормы выработки и не выполняющие их. При группировке единиц по факторному признаку важно установить границы интервалов, так как это существенно влияет на результаты дисперсионного анализа.

На результате дисперсионного анализа влияния факторов значительно сказывается также фон, на котором происходит это изучение.

Отбор единиц для построения комплекса. Точность выводов по решению дисперсионного комплекса зависит от правильности отбора единиц для его составления. Они должны правильно отражать численные значения изучаемого признака и быть верно расставлены по числу единиц в группах.

Чтобы единицы правильно отражали численные значения результативного признака, они должны быть отобраны по принципу случайной выборки. Нарушение данного принципа приводит к нерепрезентативным результатам. Если единицы дисперсионного комплекса отобраны по случайной выборке, такая организация его называется рандомизацией, а сами комплексы — рандомизированными.

По количеству единиц в группах комплексы могут быть составлены равномерно, пропорционально и неравномерно, поэтому и сами комплексы разделяют на равномерные, пропорциональные и неравномерные. В равномерных и пропорциональных комплексах сумма частных дисперсий по факторам и их сочетанию равна общей факторной дисперсии:

$$D_x^2 = D_A^2 + D_B^2 + D_{AB}^2,$$

что облегчает анализ.

Практическая организация дисперсионного комплекса приводит к построению непропорциональных комплексов. Это усложняет расчеты.

В пропорциональных комплексах численности единиц в группах находятся в одинаковой пропорции. Иногда пропорциональный комплекс организуется на базе непропорционального с использованием чисел пропорциональности. Для этого надо перемножить числа пропорциональности по факторам в каждой группе, а затем фактические данные разделить на полученные произведения и взять наименьшую из исчисленных величин. Для получения частот пропорционального комплекса следует найденную таким путем величину умножить на числа пропорциональности.

Равномерные и пропорциональные комплексы называются ортогональными.

Преобразование величин результативного признака. Поскольку в дисперсионном анализе все расчеты ведутся с величинами результативного признака, для их выражения удобнее пользоваться малозначными целыми числами. Для этого многозначные и дробные числа можно преобразовать.

Расчет показателей η^2 и F по преобразованным величинам не требует никаких поправок. Расчет же D^2 , S^2 и $\bar{y}$ требует действий, обратных преобразованиям.

Можно произвести следующие преобразования:

1) все значения результативного признака можно разделить и умножить на одну и ту же величину k. Это делается в случае деления всех значений на одно и то же число без остатка, при переходе к новым единицам измерения и когда значения выражены дробными числами и можно получить целые числа, удобные для расчетов. Поправка к конечным характеристикам при делении и умножении всех значений признака состоит в обратных действиях, которые произведены перед этим;

2) из всех вариантов можно вычесть одно и то же число. Это делается в том случае, когда варианты велики сами по себе и расстояния между ними малы. При расчетах дисперсионных комплексов поправку требуется вносить только для средней. Значения D^2 , η^2 , S^2 и F никаких поправок не требуют;

3) сначала из всех вариантов вычесть одну и ту же величину A, а затем полученные числа разделить на другую величину k. После двойного преобразования для получения полных характеристик D^2 , S^2 сокращенные характеристики требуется умножить на k^2 , а для получения $\bar{y}$ сокращенную характеристику следует умножить на k и к произведению прибавить A. Величины η^2 и F не требуют поправок и могут быть использованы в преобразованном виде.

При расчете дисперсионных комплексов можно успешно использовать преобразование величин результативного признака, а также формулы, отличные от рассмотренных при опи-

сании сущности метода. Тогда при тех же результатах значительно сокращается вычислительная работа.

Преобразованные значения результативного признака позволяют легко использовать формулы для расчета дисперсий основанных на применении свойств дисперсии.

Если в дисперсионном комплексе вычислить вспомогательные величины $h = \frac{(\sum y)^2}{n_x}$, $H = \frac{(\sum y)^2}{n}$, где h - по структуре

то же, что и H , но только исчисленное по каждой группе комплекса, то дисперсии будут выражаться следующими формулами:

$$D^2_y = \sum y^2 - H; D^2_z = \sum y^2 - \sum h; D^2_x = \sum h - H.$$

Эти формулы дают тот же результат, что и формулы, приведенные при изложении сущности метода, но значительно сокращают и упрощают вычислительную работу.

5.3. Однофакторный дисперсионный комплекс

Если методами дисперсионного анализа изучается действие одного фактора на результативный признак, то строится однофакторный дисперсионный комплекс, в котором производятся необходимые расчеты. Однофакторный дисперсионный комплекс всегда ортогонален и не требует пропорциональности, поскольку нет других факторов, с которыми можно было бы сравнивать численности его группы. Однако организации однофакторного дисперсионного комплекса должна удовлетворять требованию рандомизации, чтобы отбор единиц для построения комплекса был свободен от тенденциозности. Решение однофакторного комплекса позволяет применить те преобразования величин результативного признака, которые описаны выше. Общая схема дисперсионного анализа однофакторного комплекса представлена в табл. 1.

Таблица 1

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Оценка	F-критерий	
				Модель 1	Модель 2
Факторный признак	D_x^2	$k_x = r - 1$	$S_x^2 = D_x^2 / k_x$	$F = S_x^2 / S_z^2$	$F = S_x^2 / S_z^2$
Случайные факторы	D_z^2	$k_z = n - r$	$S_z^2 = D_z^2 / k_z$	-	-
Общий	D_y^2	$k_y = n - 1$	$S_y^2 = D_y^2 / k_y$	-	-

Степень влияния:

$$\eta_x^2 = \frac{D_x^2}{D_y^2}; \eta_z^2 = \frac{D_z^2}{D_y^2}; \eta_y^2 = \eta_x^2 + \eta_z^2 = 1.$$

5.4. Двухфакторный дисперсионный комплекс

Изучение действия двух факторов в одном комплексе требует их независимости. Кроме того, комплекс может быть пропорциональным, когда соблюдается определенное соотношение частот, и непропорциональным, когда это соотношение не соблюдается.

Рассмотрим пропорциональные комплексы. При изучении действия более чем одного фактора требуется учитывать влияние суммарного действия факторов и каждого фактора в отдельности, а также их сочетаний. Суммарное действие всех факторов находится, как и при решении однофакторного комплекса.

Установив силу и достоверность влияния по рассмотренной схеме, следует выяснить, как велико влияние каждого фактора отдельно при неизменном действии другого: каково

действие сочетания факторов. Чтобы решить эти вопросы для двух факторов, строят статистический комплекс (табл. 2):

Таблица 2

Группы по признаку А	А ₁		А ₂	
Группы по признаку В	В ₁	В ₂	В ₁	В ₂
Варианты результативного признака	У ₁ У ₂ У ₃	У ₄ У ₅ У ₆	У ₇ У ₈ У ₉	У ₁₀ У ₁₁ У ₁₂

Решение двухфакторного статистического комплекса начинается с определения общей дисперсии как суммы квадратов отклонений всех вариантов от общей средней. Затем исчисляются частные средние по всем группам, суммы квадратов отклонений отдельных вариантов от групповых средних. Сумма этих сумм и покажет величину случайной дисперсии. После этого определяется дисперсия суммарного действия обоих факторов как сумма квадратов отклонений групповых средних от общей средней. Причем квадраты отклонений по каждой группе взвешиваются по числу единиц в группе, чтобы можно было сравнивать дисперсии, полученные для разного числа слагаемых. Далее следуют операции самого двухфакторного комплекса. Это операции по определению частных факторных дисперсий и дисперсии по сочетанию факторов. Частная дисперсия фактора А определяется вариацией частных средних по группам этого фактора.

В приведенной схеме только две группы единиц по признаку А (А₁ и А₂). Для каждой группы есть одинаковая комбинация групп другого фактора В (В₁ и В₂). Обе группы фактора А имеют свои средние. Сумма квадратов отклонений этих средних от общей средней, взвешенных по численностям групп, и составит частную дисперсию по фактору А. Точно так же частная дисперсия фактора В определяется вариацией частных средних по группам этого фактора. Частная дисперсия по фактору В выразится как сумма квадратов отклонений частных средних от общей средней, взвешенных по численности групп.

Дисперсия по сочетанию факторов составляет часть общей дисперсии результативного признака, которая вызывается раз-

личным действием одного фактора при разных группировках другого. Ввиду этого суммарное действие всех факторов складывается из действия каждого фактора в отдельности и своеобразного действия их сочетания.

Общая дисперсия, таким образом, в равномерных комплексах выглядит так:

$$D_y^2 = D_A^2 + D_B^2 + D_{AB}^2 + D_z^2.$$

В более развернутом виде формула расчета общей дисперсии двухфакторного дисперсионного комплекса выглядит следующим образом:

$$\sum_{k=1}^m \sum_{j=1}^r \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y})^2 = \sum_{j=1}^r (\bar{y}_j - \bar{y})^2 n_j + \sum_{k=1}^m (\bar{y}_k - \bar{y})^2 n_k + \sum_{k=1}^m \sum_{j=1}^r (\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2 n_{jk} + \sum_{k=1}^m \sum_{j=1}^r \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{jk})^2,$$

где y_{ijk} - i-ое значение результативного признака в j-ой группе по фактору А и k-ой по фактору В, $j=1, r; k=1, m$;

$\bar{y}_{jk}$ - среднее значение результативного признака у в группе, образованной комбинацией j-го значения фактора А и k-го значения фактора В;

$\bar{y}_j$ - среднее значение результативного признака у в j-ой группе по фактору А;

$\bar{y}_k$ - среднее значение результативного признака у в k-ой группе по фактору В;

n_{jk} - число наблюдений результативного признака у в группе, образованной комбинацией j-го значения фактора А и k-го значения фактора В;

n_j - число наблюдений результативного признака у в j-ой группе по фактору А;

n_k - число наблюдений результативного признака у в k-ой группе по фактору В;

n - общее число наблюдений,

$$n = \sum_{j=1}^r n_j = \sum_{k=1}^m n_k = \sum_{k=1}^m \sum_{j=1}^r n_{jk}.$$

Степень влияния факторов и их сочетания определяется, как и в однофакторном комплексе, делением частных дисперсий на общую:

$$\eta_A^2 = \frac{D_A^2}{D_y^2}; \eta_B^2 = \frac{D_B^2}{D_y^2}; \eta_{AB}^2 = \frac{D_{AB}^2}{D_y^2}; \eta_x^2 = \frac{D_x^2}{D_y^2}; \eta_z^2 = \frac{D_z^2}{D_y^2}.$$

Необходима проверка:

$$\eta_x^2 = \eta_A^2 + \eta_B^2 + \eta_{AB}^2; \eta_y^2 = \eta_x^2 + \eta_z^2 = 1.$$

Для определения степени влияния факторов и их сочетания, то есть для проверки основных гипотез, вычисляются оценки сумм квадратов отклонений со степенями свободы: S_y^2 , S_A^2 , S_B^2 , S_{AB}^2 , S_x^2 , S_z^2 . При расчете числа степеней свободы следует учитывать, что это число для дисперсии любого фактора равно числу групп по данному фактору минус 1, а дисперсия по сочетанию факторов имеет число степеней свободы, равное произведению числа степеней свободы факторов сочетания. Для дисперсии суммарного действия число степеней свободы равно произведению числа групп по обоим факторам без единицы.

В двухфакторном дисперсионном анализе испытываемые гипотезы формулируются следующим образом:

$$1) H_0: \mu_1 = \mu_2 = \dots = \mu_r; \quad 2) H_0: \mu_1 = \mu_2 = \dots = \mu_k;$$

$$3) H_0: \mu_{11} = \mu_{12} = \dots = \mu_{rk}.$$

Таким образом, для проверки первой гипотезы, то есть для проверки существенности влияния фактора А, необходимо сравнить наблюдаемое значение статистики $F = S_A^2 / S_z^2$ с найденным по таблице распределения Фишера-Снедекора $F_{\alpha, r-1, n-mr}$. Если $F_{набл} > F_{кр} = F_{табл}$, то H_0 отвергается и влияние фактора А считается статистически значимым на уровне α .

Вторая гипотеза испытывается на основе сравнения $F = S_B^2 / S_z^2$ с найденным по таблице $F_{кр} = F_{\alpha, m-1, n-mr}$. Третья же - на основе сравнения $F = S_{AB}^2 / S_z^2$ с $F_{кр} = F_{\alpha, (r-1)(m-1), n-mr}$. Во всех случаях, если $F_{набл} > F_{кр}$, H_0 отклоняется.

Приведенный выше анализ двухфакторного дисперсионного комплекса справедлив для модели М1, неслучайной, в которой уровни всех контролируемых факторов заранее зафиксированы. Для моделей М2 и М3 F-критерии при проверке тех же гипотез несколько отличны от приведенных. Общая схема дисперсионного анализа двухфакторного комплекса приведена в табл. 3.

Таблица 3

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Оценка	F-критерий		
				Модель 1	Модель 2	Модель 3*
Фактор А	D_A^2	$k_A = r - 1$	$S_A^2 = D_A^2 / k_A$	$F = S_A^2 / S_z^2$	$F = S_A^2 / S_{AB}^2$	$F = S_A^2 / S_z^2$
Фактор В	D_B^2	$k_B = m - 1$	$S_B^2 = D_B^2 / k_B$	$F = S_B^2 / S_z^2$	$F = S_B^2 / S_{AB}^2$	$F = S_B^2 / S_{AB}^2$
Взаимодействие факторов А и В	D_{AB}^2	$k_{AB} = (r-1)(m-1)$	$S_{AB}^2 = D_{AB}^2 / k_{AB}$	$F = S_{AB}^2 / S_z^2$	$F = S_{AB}^2 / S_z^2$	$F = S_{AB}^2 / S_z^2$
Исследуемые факторы	D_x^2	$k_x = rm - 1$	$S_x^2 = D_x^2 / k_x$	$F = S_x^2 / S_z^2$	$F = S_x^2 / S_z^2$	$F = S_x^2 / S_z^2$
Случайные факторы	D_z^2	$k_z = n - mr$	$S_z^2 = D_z^2 / k_z$	-	-	-
Общий	D_y^2	$k_y = n - 1$	$S_y^2 = D_y^2 / k_y$	-	-	-

*В модели М3 случайным является фактор А.

Степень влияния:

$$\eta_A^2 = \frac{D_A^2}{D_y^2}, \eta_B^2 = \frac{D_B^2}{D_y^2}, \eta_{AB}^2 = \frac{D_{AB}^2}{D_y^2}, \eta_x^2 = \frac{D_x^2}{D_y^2}, \eta_z^2 = \frac{D_z^2}{D_y^2}.$$

6. КОМПОНЕНТНЫЙ АНАЛИЗ

6.1. Основные понятия факторного и компонентного анализа

При решении задач методом главных компонент главное заключается в выделении линейных комбинаций случайных величин, имеющих максимально возможную дисперсию. Он опирается на ковариационную или корреляционную матрицу этих величин.

Методы факторного анализа позволяют решать следующие четыре задачи.

Первая задача заключается в «сжатии» информации до обобщимых размеров, то есть извлечении из исходной информации наиболее существенной части за счет перехода от системы исходных переменных к системе обобщенных факторов. При этом выявляются неявные, непосредственно не измеряемые, но объективно существующие закономерности, обусловленные действием как внутренних, так и внешних причин.

Вторая задача сводится к описанию исследуемого явления значительно меньшим числом m обобщенных факторов (главных компонент) по сравнению с числом исходных признаков. Обобщенные факторы - это новые единицы измерения свойств явления, линейно зависящие от исходных, непосредственно измеряемых признаков.

Третья задача связана с выявлением взаимосвязи наблюдаемых признаков с вновь полученными обобщенными факторами.

Четвертая задача заключается в построении уравнения регрессии на главных компонентах с целью прогнозирования изучаемого явления.

Главные компоненты могут быть также использованы при классификации наблюдений (объектов). В экономических исследованиях стремление полнее изучить исследуемое явление приводит к включению в модель все большего числа исходных пере-

менных, которые зачастую отражают одни и те же свойства объекта наблюдения. Это приводит к высокой корреляции между переменными, то есть к явлению мультиколлинеарности. При этом классические методы регрессионного анализа оказываются малоэффективными. Преимущество уравнения на главных компонентах в том, что последние не коррелированы между собой.

В основу факторного анализа и метода главных компонент положена линейная модель. Пусть получена информация, характеризующая n объектов по k признакам. Тогда модель факторного анализа можно представить в виде:

$$z_{ij} = \sum_{v=1}^m a_{jv} f_{iv} + a_j w_{ij} \quad (6.1.)$$

где $m < k$, $j=1, k, i=1, n$;

w_{ij} - значение j -го индивидуального фактора, присущего только j -му признаку у i -го объекта;

f_{iv} - значение v -го общего фактора, присутствующего более чем в одном признаке у i -го объекта;

z_{ij} - значение j -го нормированного признака, полученного из факторной модели для i -го объекта;

a_{jv} - весовой коэффициент v -го общего фактора на j -ой переменной;

a_j - весовой коэффициент j -го индивидуального фактора на j -ой переменной.

Теперь рассмотрим модель метода главных компонент:

$$z_{ij} = \sum_{v=1}^k a_{jv} f_{iv} \quad (6.2.)$$

где $j, v=1, k, i=1, n$.

По структуре она напоминает модель факторного анализа, но имеется существенная разница. Во-первых, отсутствуют специфические факторы. Во-вторых, число главных компонент k здесь соответствует числу признаков k . Значит, в модели главных компонент исчерпывается вся дисперсия исследуемого процесса.

Главные компоненты являются характеристическими векторами ковариационной матрицы. Множество главных компонент представляет собой удобную систему координат, а их вклад в общую дисперсию характеризует статистические свойства главных компонент. Из общего числа главных компонент для исследования, как правило, оставляют m ($m < k$) наиболее весомых, то есть вносящих максимальный вклад в объясняемую часть общей дисперсии. Опыт исследований показывает, что $m \approx (0,1: 0,25)k$. Для экономической интерпретации полученных результатов самыми наглядными являются случаи, когда $m = 1, 2, 3, 4$.

Таким образом, несмотря на то, что в методе главных компонент надо для точного воспроизведения корреляции и дисперсий между переменными найти все компоненты, большая доля дисперсии объясняется небольшим числом главных компонент. Кроме того, можно по признакам описать факторы, а по факторам (главным компонентам) описать признаки.

Наконец, следует отметить линейность и аддитивность метода. Центроидный метод, например, несет в себе только гипотезу линейности. Если она неверна, то результаты могут быть использованы только для первого приближения. В настоящее время часто используется центроидный метод для получения приближенных оценок, которые затем уточняются методами максимума правдоподобия.

Рассмотрим два аспекта проблемы неопределенности факторного решения. Во-первых, если имеется заданное число признаков и коэффициентов корреляции между ними, то может быть найдено бесконечное множество ортогональных систем факторов, которые адекватно описывают выборочные коэффициенты корреляции. С другой стороны, в факторном анализе неоднозначность факторных нагрузок a_{jv} связана с тем, что опре-

деляется m -мерное пространство, но не определяется базис в нем. Следовательно, не определяется положение факторов в этом m -мерном пространстве. С этим приходится сталкиваться при поиске решения, удовлетворяющего модели, или когда в интересах лучшей интерпретации методом вращения поворачиваются попарно оси координат при неподвижных точках, отображающих признаки. Нужно отметить, что большинство методов вычисления в факторном анализе дает неоднозначные решения для факторных нагрузок.

6.2. Метод главных компонент

Рассмотрим подробнее алгоритм компонентного анализа. На основании матрицы исходных данных X , размерности $n \times k$ вычисляются вектор средних, вектор оценок средних квадратических отклонений и матрица нормированных значений Z с элементами

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Рассчитывается матрица парных коэффициентов корреляции R .

В матричной форме модель метода главных компонент имеет вид

$$Z = FA^T, \quad (6.3.)$$

где F - матрица значений главных компонент размерности $n \times k$;

$$F = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1k} \\ f_{21} & f_{22} & \dots & f_{2k} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nk} \end{pmatrix};$$

A - матрица факторных нагрузок размерности $k \times k$,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \dots & \dots & \dots & \dots \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{pmatrix}.$$

Матрица F описывает n наблюдений в пространстве k главных компонент. При этом элементы матрицы F нормированы, то есть

$$\bar{f}_v = \frac{1}{n} \sum_{i=1}^n f_{iv} = 0, \quad s_{f_v}^2 = \frac{1}{n} \sum_{i=1}^n f_{iv}^2 = 1,$$

а главные компоненты некоррелированы между собой. Из этого следует, что

$$\frac{1}{n} F^T F = E, \quad (6.4.)$$

где E - единичная матрица размерности $k \times k$.

С целью интерпретации элементов матрицы A рассмотрим выражение для парного коэффициента корреляции между z_j - переменной и, например, f_1 - компонентой. Так как z_j и f_1 нормированы, будем иметь

$$r_{z_j f_1} = \frac{1}{n} \sum_{i=1}^n z_{ij} f_{i1} = \frac{1}{n} \sum_{i=1}^n \left(\sum_{v=1}^k a_{jv} f_{iv} \right) f_{i1} = a_{j1} \frac{1}{n} \sum_{i=1}^n f_{i1}^2 + \sum_{v=2}^k a_{jv} \left(\frac{1}{n} \sum_{i=1}^n f_{i1} f_{iv} \right), \quad (6.5.)$$

Принимая во внимание, что

$$\frac{1}{n} \sum_{i=1}^n f_{iv} f_{iv} = \begin{cases} 1, v=v^{\circ}, \\ 0, v \neq v^{\circ}, \end{cases} \quad v, v^{\circ} = 1, 2, \dots, k. \quad (6.6.)$$

окончательно получим $r_{zj} f_v = a_{jv}$.

Рассуждая аналогично, можно записать в общем виде

$$r_{zj} f_v = a_{jv} \quad (6.7.)$$

для всех $j, v = 1, k$.

Таким образом, элемент a_{jv} матрицы факторных нагрузок А характеризует тесноту линейной связи между z_j -й исходной переменной и v -й главной компонентой, то есть $-1 \leq a_{jv} \leq +1$.

Рассмотрим теперь выражение для дисперсии z_j -й нормированной переменной:

$$\begin{aligned} s_{zj}^2 &= \frac{1}{n} \sum_{i=1}^n z_{ij}^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{v=1}^k a_{jv} f_{iv} \right)^2 = \\ &= \frac{1}{n} \sum_{i=1}^n \left[\sum_{v=1}^k a_{jv}^2 f_{iv}^2 + 2 \sum_{v \neq v^{\circ}} a_{jv} a_{jv^{\circ}} f_{iv} f_{iv^{\circ}} \right] = \\ &= \sum_{v=1}^k a_{jv}^2 \left(\frac{1}{n} \sum_{i=1}^n f_{iv}^2 \right) + 2 \sum_{v \neq v^{\circ}} a_{jv} a_{jv^{\circ}} \left(\frac{1}{n} \sum_{i=1}^n f_{iv} f_{iv^{\circ}} \right), \end{aligned}$$

где $v, v^{\circ} = 1, k$.

Учитывая (6.1) окончательно получим:

$$s_{zj}^2 = \sum_{v=1}^k a_{jv}^2 = 1. \quad (6.8.)$$

По условию переменные z_j нормированы и $s_{zj}^2 = 1$. Таким образом, дисперсия z_j -й переменной представлена своими составляющими, определяющими долю вклада в нее всех k главных компонент.

Полный вклад v -й главной компоненты в дисперсию всех k исходных признаков вычисляется по формуле

$$\lambda_k = \sum_{j=1}^k a_{jv}^2. \quad (6.9.)$$

Одно из основополагающих условий метода главных компонент связано с представлением корреляционной матрицы R через матрицу факторных нагрузок A :

$$\begin{aligned} R &= \frac{1}{n} Z^T Z = \frac{1}{n} (F A^T)^T F A^T = A \left(\frac{1}{n} F^T F \right) A^T, \\ R &= A A^T. \end{aligned} \quad (6.10.)$$

Перейдем теперь непосредственно к отысканию собственных значений и собственных векторов корреляционной матрицы R . Из линейной алгебры известно, что для любой симметрической матрицы R всегда существует такая ортогональная матрица U , что выполняется условие

$$U^T R U = \Lambda. \quad (6.11.)$$

где Λ - диагональная матрица собственных значений размерности $k \times k$;

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_k \end{pmatrix};$$

U - ортогональная матрица собственных векторов размерности $k \times k$,

$$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1k} \\ u_{21} & u_{22} & \dots & u_{2k} \\ \dots & \dots & \dots & \dots \\ u_{k1} & u_{k2} & \dots & u_{kk} \end{pmatrix}$$

Так как матрица Λ положительно определена, то есть ее главные миноры положительны, то все собственные значения положительны - $\lambda_v > 0$ для всех $v=1, k$. В компонентном анализе элементы матрицы Λ ранжированы: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_v \geq \dots \geq \lambda_k > 0$. Как будет показано ниже, собственное значение λ_v характеризует вклад v -й главной компоненты в суммарную дисперсию исходного признакового пространства. Таким образом, первая главная компонента вносит наибольший вклад в суммарную дисперсию, а последняя k -я - наименьший. В ортогональной матрице U собственных векторов v -й столбец является собственным вектором, соответствующим λ_v -му собственному значению.

Собственные значения $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_v \geq \dots \geq \lambda_k$ находятся как корни характеристического уравнения

$$|R - \Lambda E| = 0. \quad (6.12.)$$

Собственный вектор V_v , соответствующий собственному значению λ_v корреляционной матрицы R, определяется как отличное от нуля решение уравнения, которое следует из условия (6.11.):

$$(R - \lambda_v E) V_v = 0. \quad (6.13.)$$

Нормированный собственный вектор равен:

$$U_v = \frac{V_v}{\sqrt{V_v^T V_v}}.$$

Из условия ортогональности матрицы U следует, что $U^{-1} = U^T$, но тогда по определению матрицы R и Λ подобны, так как они удовлетворяют условию

$$U^{-1} R U = \Lambda.$$

Так как следы, то есть суммы диагональных элементов, у подобных матриц равны, то

$$tr \Lambda = tr(U^{-1} R U) = tr(R(U U^{-1})) = tr R.$$

Напомним из линейной алгебры, что умножение матрицы U на обратную матрицу U^{-1} дает единичную матрицу E. Следы матричных произведений $(U^{-1} R U)$ и $R(U U^{-1})$ также равны.

Учитывая, что сумма диагональных элементов матрицы R равна k, получим

$$tr \Lambda = tr R = k.$$

Таким образом,

$$\sum_{v=1}^k \lambda_v = k. \quad (6.14.)$$

Представим матрицу факторных нагрузок A в виде

$$A = U\Lambda^{1/2}. \quad (6.15.)$$

Представим v -й столбец матрицы A :

$$A_v = U_v \lambda_v^{1/2},$$

где U_v - собственный вектор матрицы R , соответствующий собственному значению λ_v .

Найдем норму вектора

$$A_v \left| A_v \right|^2 = A_v^T A_v = \lambda_v^{1/2} U_v^T U_v \lambda_v^{1/2} = \lambda_v. \quad (6.16.)$$

Здесь учитывалось, что вектор нормированный U_v и $U_v^T U_v = 1$. Значит,

$$\lambda_v = \sum_{j=1}^k a_{jv}^2.$$

Сравнив полученный результат с (6.9.), можно сделать вывод, что собственное значение λ_v характеризует вклад v -й главной компоненты в суммарную дисперсию исходных признаков. Из (6.15.) следует:

$$A^T A = \Lambda. \quad (6.17.)$$

Согласно (6.14.) общий вклад всех главных компонент в суммарную дисперсию равен k . Тогда удельный вклад v -й главной компоненты определяется по формуле

$$\frac{\lambda_v}{k} 100\%.$$

Суммарный вклад m первых главных компонент определяется из выражения

$$\frac{\sum_{v=1}^m \lambda_v}{k} 100\%.$$

Обычно для анализа используют m первых главных компонент, суммарный вклад которых превышает 60-70%.

Матрица факторных нагрузок A используется для экономической интерпретации главных компонент, которые представляют собой линейные функции исходных признаков. Для экономической интерпретации используются лишь те f_v , для которых $|a_{jv}| > 0,5$.

Значения главных компонент для каждого i -го объекта ($i = \overline{1, n}$) задаются матрицей F . Матрицу значений главных компонент можно получить из формулы

$$Z = FA^T,$$

откуда

$$F = Z(A^T)^{-1} = ZU\Lambda^{-1/2}.$$

Уравнение регрессии на главных компонентах строится по алгоритму пошагового регрессионного анализа, где в качестве

аргументов используются главные компоненты, а не исходные показатели. К достоинствам последней модели следует отнести тот факт, что главные компоненты некоррелированы. При построении уравнений регрессии следует учитывать все главные компоненты.

7. ФАКТОРНЫЙ АНАЛИЗ

7.1. Факторный анализ как статистический метод

При анализе социально-экономических процессов и явлений исследователь сталкивается с многомерностью их описания, то есть с необходимостью учитывать в анализе большое число показателей (параметров или признаков). Многие признаки взаимосвязаны и в значительной мере дублируют друг друга. В то же время нередко признаки лишь в косвенной форме отражают наиболее существенные, но не поддающиеся непосредственному наблюдению и измерению внутренние, скрытые свойства явлений. Так, по данным о распределении многих демографических показателей и результатам опросов социолог судит о склонности населения к миграции; только большая совокупность непосредственно измеряемых признаков позволяет сопоставлять страны, районы и города по уровню жизни, предприятия — по особенностям функционирования, продукцию — по качеству и т.д.

В такого рода ситуациях представляется естественной попытка сконцентрировать информацию, выражая большое число исходных косвенных признаков через меньшее число более емких внутренних характеристик явления. При этом предполагается, что наиболее емкие характеристики окажутся одновременно наиболее существенными, определяющими. Методы факторного анализа и предназначены для такой конденсации информации. Существо методов факторного анализа состоит в переходе от описания некоторого множества изучаемых объектов, заданного большим набором косвенных непосредственно измеряемых признаков, к описанию меньшим числом максимально информативных глубинных переменных, отражающих наиболее существенные свойства явления. Такого рода переменные, называемые факторами, являются некоторыми функциями исходных признаков.

Важной отличительной особенностью факторного анализа является возможность одновременного исследования сколь

угодно большого числа взаимозависимых переменных. Здесь нет допущения о «неизменности всех прочих условий», свойственного многим другим методам статистического анализа. Для социально-экономических исследований отсутствие ограничений на число и взаимосвязь переменных особенно важно, поскольку изолировать влияние отдельных переменных на поведение всей системы оказывается трудным. Поэтому здесь с успехом применяются методы факторного анализа, позволяющие расчетным путем вычленить относительно независимые аспекты явления для дальнейших более углубленных исследований. В частности, весьма перспективным представляется использование результатов факторного анализа в моделях множественной регрессии.

Фактор представляет собой расчетную переменную, то есть некоторую новую характеристику объектов изучаемого множества. Описание фактора в терминах его связи с набором исходных признаков отыскивается в виде так называемой факторной матрицы, или матрицы факторных нагрузок A . Основой для построения матрицы факторных нагрузок служит матрица парных коэффициентов корреляции R . Она отражает степень взаимосвязи между каждой парой признаков, тогда как матрица факторных нагрузок характеризует степень связи между каждым из k рассматриваемых признаков и m факторами, выявленными в процессе анализа. При этом число m факторов выбирается исходя из двух условий: m должно быть много меньше k , а уровень потерь в информации достаточно мал.

Матрица факторных нагрузок позволяет выделить для каждого фактора группу признаков, наиболее тесно с ним связанных. Тем самым открывается возможность сопоставить факторы друг с другом, дать им содержательное толкование и наименование, то есть дать интерпретацию факторов.

В связи с этим важно отметить, что преобразование корреляционной матрицы в факторную не является однозначным. Одну и ту же корреляционную матрицу можно преобразовать в различные факторные. Для выбора факторной матрицы, на основе которой будет производиться интерпретация факторов,

разрабатываются специальные приемы. Как правило, в качестве таковой выбирается матрица, в которой исходные параметры сильно связаны с одним из факторов и слабо связаны со всеми другими.

В социально-экономических исследованиях исходные параметры несут, как правило, большую смысловую нагрузку, сами являются содержательными измеряемыми признаками. Главная цель факторного анализа в этих исследованиях заключается в агрегировании данных, направленном на выявление общих закономерностей. В социально-экономических исследованиях накоплен большой опыт (без факторного анализа) введения различных обобщенных индикаторов. В силу этого задача нахождения и интерпретация факторов является в социально-экономических исследованиях, как правило, лишь первой задачей факторного анализа. Вторая задача состоит в построении так называемой матрицы факторных весов, элементы которой определяют значения факторов для каждого из объектов изучаемого множества.

Изложенное не означает, что первая из упомянутых задач — задача построения матрицы факторных нагрузок — в социально-экономических исследованиях носит чисто вспомогательный характер. В ряде случаев именно эта задача является главной. Кроме того, построение матрицы факторных весов существенно зависит от того, на основе какой матрицы факторных нагрузок она строится. Поэтому интерпретация распределения значений факторов по объектам зависит от интерпретируемости матрицы факторных нагрузок, которая определяет, какие исходные параметры дают основной вклад в тот или иной фактор.

После того как значения факторов оказались определенными для всех объектов наблюдения, не составляет труда построить либо гистограммы распределений их значений в виде отдельных графиков, либо двумерных полей распределений на плоскостях, определяемых попарно взятыми факторами. Графики-гистограммы и двумерные поля распределений в наглядном и компактном виде представляют основные закономерности, содержащиеся в исходном эмпирическом материале, позволяют

исследователю осуществлять типологию множества изучаемых объектов, то есть выделять классы однородных (похожих) объектов. Визуальное представление результатов облегчает интерпретацию как самих факторов, так и выделенных типов объектов, позволяет учитывать при типологии большой багаж экспертных представлений, не включенных в первоначальный набор признаков.

Арсенал методов факторного анализа чрезвычайно широк. Вместе с тем единство методов факторного анализа зиждется на общей идее построения метода. Эта идея состоит в следующем. Тем или иным способом формализуется представление о том, что требуется для того, чтобы данная система исходных параметров считалась хорошо представленной (описанной, аппроксимированной) с помощью некоторой системы факторов. Для этого строится критерий, значение которого можно определить, как только задана матрица факторных нагрузок. Это значение имеет смысл степени качества представления данной системой факторов системы исходных параметров. После того как такой критерий построен, задача о нахождении искомой матрицы факторных нагрузок ставится как задача экстремизации построенного критерия.

Таким образом, факторный анализ с формальной точки зрения — это прежде всего несколько критериев качества матрицы факторных нагрузок и набор алгоритмов поиска экстремумов этих критериев. Почти все алгоритмы факторного анализа — это простейшие итерационные процедуры поиска экстремума. Различные критерии формализуют различные содержательные представления о том, что означает хорошее сжатие информации. Поэтому при практическом использовании факторного анализа очень полезным оказывается многовариантный подход к обработке материала (проведение анализа этого материала многими методами). Сопоставление результатов дает возможность всесторонне представить себе существенное, общее в проведенных преобразованиях. В частности, наличие большого сходства между результатами, полученными с помощью различных методов обработки, означает, что сжатое представление исходного мате-

риала действительно отражает существо информации, представленной в этом материале, поскольку это представление оказалось практически не зависящим от способа формализации того, что означает хорошее сжатие.

Полезно также сравнивать результаты обработки исследуемого материала различными алгоритмами, реализующими поиск экстремумов одного и того же критерия. Это связано с приближенным характером работы этих алгоритмов: на одном и том же материале разные алгоритмы могут давать различный результат. Поэтому сходство результатов работы нескольких алгоритмов можно рассматривать как выражение надежности данного преобразования.

7.2. Основные техники факторного анализа

Пусть в распоряжении исследователя есть совокупность из n наблюдений над состоянием изучаемого процесса или явления. При этом оно описано набором из k признаков. Во всех такого рода случаях информация представляется в виде таблицы чисел X размерностью $n \times k$. Строки этой таблицы соответствуют наблюдениям.

В ряде случаев, однако, изучаемое явление наблюдается как неделимое единое целое и не может быть представлено как совокупность объектов. Тогда исследователь может располагать множеством наблюдаемых состояний явления, определенных на ряд последовательных моментов времени. При этом строки таблицы данных соответствуют множеству временных интервалов: годы, кварталы, месяцы и т.д. Например, комплекс условий, определяющих экономическую эффективность предприятия, может изучаться по состоянию за ряд лет.

Столбцы таблицы данных соответствуют признакам, характеризующим явление. Набор признаков обычно включает наиболее существенные в рамках конкретного исследования и легко измеряемые характеристики объектов.

Таким образом, каждый признак из исследуемого набора задается в виде статистического ряда. Если строки матрицы —

суть множество объектов, то это ряды, характеризующие изменение явления в пространстве; если строки соответствуют временным интервалам, то признаки заданы рядами динамики.

Итак, матрица данных суммирует наблюдения по двум позициям: объекты — признаки. Такое сочетание позиций наблюдения преобладает в социально-экономических исследованиях с использованием факторного анализа. Однако возможны и другие сочетания. Так, исследователя может интересовать:

1) изменение значений признаков, описывающих один объект (его состояние) во времени, то есть от одного временного интервала к другому;

2) пространственные различия, то есть различия множества объектов по набору признаков;

3) изменение одного признака во времени и в пространстве, то есть на множестве объектов и на множестве интервалов времени.

Изучая явление с перечисленных позиций наблюдения, исследователь имеет дело с трехмерным кубом данных в пространстве: признак — объект — время.

С помощью факторного анализа обычно изучаются плоские двумерные срезы куба данных — матрицы данных. Одновременный факторный анализ трехмерного куба данных находится в данное время в стадии методических разработок.

Соответственно замыслам исследования в пространстве данных выбирается конкретный двумерный срез куба данных, который определяет матрицу данных для анализа. В литературе по приложению факторного анализа различные двумерные срезы куба данных часто определяют как разные виды факторного анализа, или разные техники. Различают R, Q, O, P, S и T-техники (или R, Q-анализы). Важно, что термин «техника» (неудачный, но ставший традиционным) означает в данном случае исключительно форму организации и представления исходного материала, то есть вид матрицы данных, но никак не технику (метод) обработки. Все названные здесь виды техник, то есть самые разнообразные виды матриц данных, могут обрабатываться одними и теми же методами факторного анализа.

Чаще всего методами факторного анализа исследуются матрицы данных, в которых изучается вариация признаков (столбцы матрицы) от объекта к объекту (строки). При этом определяются группы признаков со сходным характером вариации, составляющие основу факторов. Такой подход получил название R-техники, или R-анализа.

На том же срезе куба данных может быть поставлена другая задача, а именно: выявить группы объектов, имеющих сходный профиль по большому набору признаков, то есть классифицировать объекты. Классификация может быть осуществлена на основе результатов вышеописанного анализа признаков, с использованием выявленных факторов как ее критериев. Известен и другой, более непосредственный подход к решению этой задачи в рамках факторного анализа, его называют Q-анализом. При таком подходе матрицу данных, описанную выше, транспонируют: теперь столбцы матрицы соответствуют изучаемым объектам, а строки — признакам, по которым наблюдаются объекты. Объекты теперь рассматриваются в качестве переменных, значения которых варьируют от наблюдения к наблюдению, то есть от одного признака к другому.

Вернемся к кубу данных. Рассмотрим другую позицию наблюдения — временной срез данных (по признакам и во времени). На таком материале можно анализировать характер вариации признаков во времени и отыскивать факторы как некие обобщенные параметры, хорошо описывающие эту вариацию. Такое представление матрицы называют P-анализом. Столбцы матрицы данных соответствуют здесь признакам, строки — временным интервалам. Тот же временной срез матрицы данных в транспонированном виде (столбцы — временные интервалы, строки — признаки) может обрабатываться методами факторного анализа с целью выявления периодов времени со сходным (внутри периода) сочетанием значений признаков. Такой подход носить название Q-анализа.

Рассмотрим варианты, исследующие только один признак. При таком подходе в матрице исходных данных также фигурируют две позиции наблюдения, например, наблюдения по объек-

там и во времени. Пусть столбцы матрицы данных соответствуют объектам наблюдения, а строки — временным интервалам (S-анализ). Каждый элемент матрицы исходных данных есть значение некоего (одного и того же для всех столбцов) признака на i -м объекте в момент t . Примером такой постановки задачи может служить исследование движения цен на акции разных компаний: объекты — виды акций, строки соответствуют временным интервалам. Каждый элемент есть денежная оценка акций i -го вида в момент времени t , то есть исследуется один признак цена. Выявленные факторы соответствуют группам акций с общим рисунком движения цен во времени.

В такой постановке задача факторного анализа позволяет разрабатывать динамическую типологию объектов наблюдения, то есть выявлять группы объектов со сходным типом изменений во времени.

Матрица исходных данных, в которой столбцы соответствуют временным интервалам, а строки — объектам наблюдения, тогда как признак фиксирован (Т-анализ), служит исходным материалом для выявления периодов времени с характерным для каждого периода распределением значений исследуемого признака по объектам наблюдения.

Следует заметить, что на сегодняшний день в прикладных исследованиях с факторным анализом доминирует R-анализ. Часто в социально-экономических исследованиях встречаются примеры применения Q-анализа и редко Р-анализа.

7.3. Модель факторного анализа

Как уже подчеркивалось ранее, в факторном анализе исходят из того, что все k исходных признаков $x_1, x_2, \dots, x_k$ линейно зависят от меньшего числа $m < k$ других, непосредственно не изменяемых общих факторов $f_1, f_2, \dots, f_m$, которые в большинстве моделей рассматриваются как взаимно некоррелированные. При этом не предполагается в общем случае возможность восстановления каждого признака x_j по соответствующим значениям общих факторов.

Предполагается также, что каждый из исходных признаков x_j зависит от некоторой «специфической» остаточной компоненты w_j , которая и обуславливает статистический характер связи между x_j и общими факторами $f_1, f_2, \dots, f_m$ ($m < k$).

При проведении статистического исследования на основе факторного анализа, как правило, приходится решать задачу выявления и интерпретации общих факторов, стремясь при этом минимизировать как их число, так и степень зависимости x_j от своих специфических случайных составляющих w_j .

Методы и модели факторного анализа предназначены для сжатия информации, то есть снижения размерности исходного признакового пространства.

При разработке моделей факторного анализа приходится последовательно анализировать и решать вопросы существования и идентификации модели, статистического оценивания ее неизвестных структурных параметров, статистической проверки гипотез, связанных с природой параметров модели, а также построения статистических оценок для значений общих факторов.

Не для всякой заданной структуры связей между исходными признаками $x_1, x_2, \dots, x_k$ можно построить модель факторного анализа, то есть определить такие общие факторы, которые полностью объяснили бы имеющуюся корреляцию между различными парами признаков x_j и x_l ($j, l = 1, 2, \dots, k$).

Модель, определяющая зависимость между x_j и f_v ($v = 1, 2, \dots, m$), где ($m < k$), имеет место лишь при определенных корреляционных (ковариационных) матрицах R (Σ), и соотношении между числом исходных признаков k и общих факторов m .

Однако, если k, Σ, m допускают построение модели факторного анализа, то выполнение условия единственности (идентификации) модели требует дополнительных ограничений на матрицу факторных нагрузок A и ковариационную матрицу Σ_w остаточных специфических факторов $w_1, w_2, \dots, w_k$.

Предположим, что наблюдения $x_1, x_2, \dots, x_n$ центрированы. Переход к центрированным наблюдениям осуществляется с помощью переноса начала координат в «центр тяжести» ис-

ходного множества наблюдений. Тогда линейную модель факторного анализа можно представить в виде

$$X = FA^T + W, \quad (7.1.)$$

где X - исходная матрица центрированных наблюдений размерности $n \times k$,

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix};$$

F - матрица значений общих факторов размерности $n \times m$,

$$F = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1m} \\ f_{21} & f_{22} & \dots & f_{2m} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nm} \end{pmatrix};$$

f_{jv} - значение v -го общего фактора у j -го наблюдения;

A - матрица факторных нагрузок размерности $k \times m$,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{k1} & a_{k2} & \dots & a_{km} \end{pmatrix};$$

a_{jv} - коэффициенты линейного преобразования, нагрузка v -го общего фактора на j -ю исходную переменную, характеризует тесноту линейной связи между x_j и f_v , то есть $a_{jv} = \rho_{x_j f_v}$;

W - матрица значений остаточных специфических факторов размерности $n \times k$,

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \dots & \dots & \dots & \dots \\ w_{n1} & w_{n2} & \dots & w_{nk} \end{pmatrix};$$

w_{ij} - значение j -го специфического фактора у i -го наблюдения.

Применительно к отдельному i -му наблюдению над j -ой исходной переменной модель факторного анализа имеет вид

$$x_{ij} = \sum_{v=1}^m a_{jv} f_{iv} + w_{ij}, \quad i = \overline{1, n}, j = \overline{1, k}. \quad (7.2.)$$

Предположим, что вектор остаточных специфических факторов $w = (w_1, w_2, \dots, w_k)^T$ подчиняется k -мерному нормальному закону распределения $N_k(0, \Sigma_w)$, не зависит от F и состоит из взаимно независимых компонент. Тогда ковариационная матрица $\Sigma_w = M(WW^T)$ имеет диагональный вид с дисперсиями по главной диагонали.

Вектор общих факторов $f = (f_1, f_2, \dots, f_m)^T$ в зависимости от конкретной задачи может рассматриваться либо как вектор, имеющий m -мерный нормальный закон распределения с нулевым вектором средних $Mf = 0$ и ковариационной матрицей $\Sigma_f = M(ff^T) = E_m$, либо как вектор неизвестных неслучайных параметров, значения которых меняются от наблюдения к наблюдению. При этом условия центрированности дисперсий элементов вектора f имеют вид

$$\frac{1}{n} \sum_{i=1}^n f_i = 0; \quad \frac{1}{n} \sum_{i=1}^n f_i f_i^T = E_m.$$

где $f_i = (f_{i1}, f_{i2}, \dots, f_{im})^T$, E_m - единичная матрица размерности $m \times m$,

При обоих вариантах интерпретации вектора общих факторов f , вектор исходных показателей является нормально распределенным k -мерным случайным вектором: в первом случае как линейная комбинация двух случайных векторов f и w , а во втором варианте за счет нормальности специфических факторов w_j . При этом из выше сделанных допущений следует, что

$$Mx_j = 0; \begin{cases} \sigma_{jj} = \sum_{v=1}^m a_{jv}^2 + \sigma_{jj}^w \\ \sigma_{jl} = \sum_{v=1}^m a_{jv} a_{lv}; j \neq l (j, l = 1, k). \end{cases} \quad (7.3.)$$

или в матричной форме $MX=0; \Sigma = AA^T + \Sigma_w$.

При проведении факторного анализа все расчеты носят последовательный характер. Процедура выполнения вычислительных операций представлена на схеме 1. Вертикальными стрелками на схеме обозначены основные проблемы факторного анализа: проблема получения оценок общностей, проблема выделения факторов, проблема вращения факторов и проблема получения оценок значений факторов. Рассмотрим более подробно методы решения некоторых из вышеперечисленных проблем.

7.4. Основные методы определения оценок общностей

Существует два подхода к отысканию оценок элементов главной диагонали редуцированной матрицы парных коэффициентов корреляции. В первом подходе вначале определяют оценки общностей, а затем число факторов. Такой подход используется в методе главных факторов, центроидном методе, методе треугольной декомпозиции по матрице R^h . Во втором подходе предварительно устанавливается количество факторов

которые должны быть выделены. Затем подбираются значения общностей так, чтобы ранг матрицы R^h приближался к m . Предварительная оценка числа главных (общих) факторов используется в методах максимального правдоподобия и минимальных остатков.

В настоящее время из перечисленных многомерных методов статистического анализа в основном используются метод главных компонент и метод главных факторов. Последний метод представляет собой факторный анализ, при котором вначале определяются оценки общностей, а потом число факторов. При этом обычно прибегают к следующим методам получения оценок общностей:

1. Метод определения оценки общности h_j^2 при помощи квадрата коэффициента множественной корреляции.

Привлекательной стороной этого метода является возможность вычисления h_j^2 при помощи обратной корреляционной матрицы R^{-1} .

$$h_j^2 = R_{j, 12 \dots k} = 1 - \frac{1}{r_{jj}},$$

где r_{jj} — диагональный элемент обратной матрицы R^{-1} .

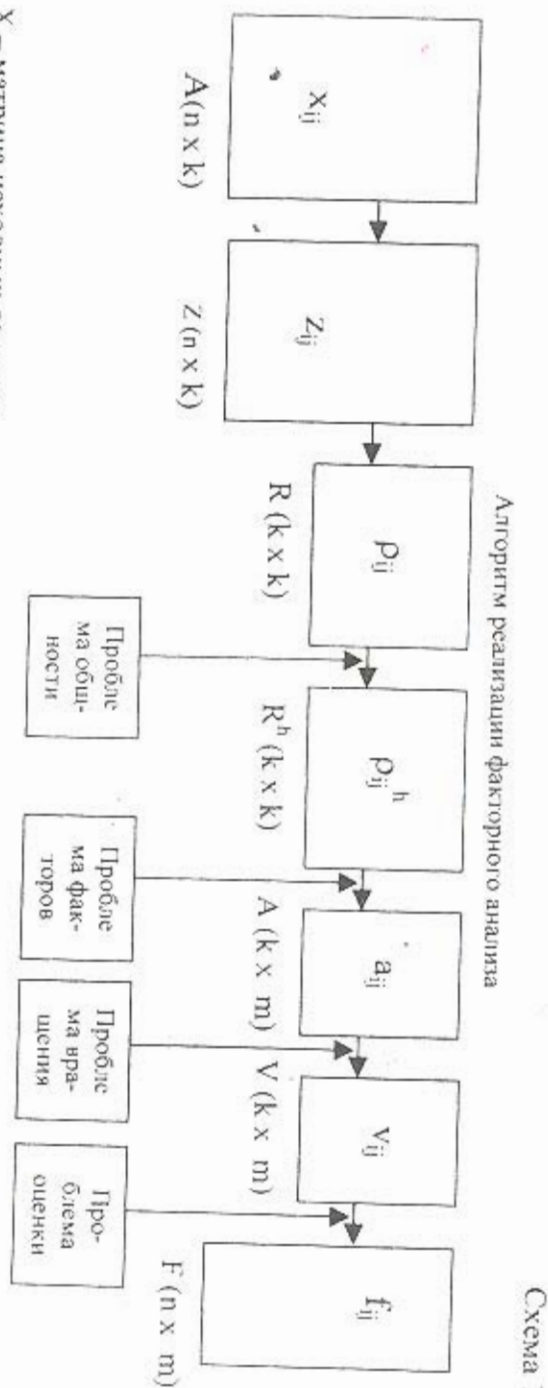
2. Метод определения оценки общности h_j^2 при помощи наибольшего коэффициента корреляции по строке (столбцу).

При большом числе исходных переменных этот метод имеет весьма широкое распространение, в частности, благодаря своей простоте. Данный метод дает значение общности, которое может оказаться выше или ниже ее истинного значения. Однако, если число признаков более 20, то конечные результаты незначительно отличаются от результатов, полученных при использовании более точных методов.

Суть метода заключается в том, что в строке матрицы R , соответствующей данному признаку, выбирается элемент с наибольшим абсолютным значением. Это наибольшее значение коэффициента корреляции и записывается на главной диагонали (разумеется, со знаком «плюс»).

3. Метод определения оценки общности h_j^2 при помощи среднего коэффициента корреляции по строке (столбцу).

Полученное значение оценки общности в данном случае



X — матрица исходных данных;
 Z — матрица нормированных данных;
 R — матрица парных коэффициентов корреляции;
 R^h — редуцированная матрица парных коэффициентов корреляции;
 A — матрица факторных нагрузок;
 V — матрица факторных нагрузок после вращения;
 F — матрица значений факторов.

должно оказаться меньше наибольшего коэффициента корреляции.

$$h_j^2 = \frac{1}{k-1} \sum_{l=1}^k r_{jl}, \quad j \neq l.$$

4. Метод триад определения оценки общности h_j^2 .

При реализации данного метода в j -ой строке (столбце) матрицы R отыскиваются два наибольших значения коэффициентов корреляции r_{jk} и r_{jl} и составляется триада

$$h_j^2 = \frac{r_{jk} \cdot r_{jl}}{r_{kl}}.$$

5. Метод первого центроидного фактора определения оценки общности h_j^2 .

Последовательность действий, соответствующих данному методу получения оценок общности следующая:

а) На главной диагонали матрицы R ставятся наибольшие коэффициенты корреляции каждой строки (столбца).

б) По новой матрице вычисляется отношение квадрата суммы элементов соответствующей строки (столбца) к сумме всех элементов матрицы

$$h_j^2 = \frac{\left(\sum_{m=1}^k r_{jm} \right)^2}{\sum_{m=1}^k \sum_{l=1}^k r_{ml}}.$$

При помощи этого метода определяется оценка общности снизу.

Следует заметить, что если при расчете получается $h_j^2 \geq 1$, то оценка общности принимается равной 1.

7.5. Вращение пространства общих факторов

Задача вращения общих факторов решается с целью улучшения их интерпретируемости. Если факторные нагрузки в структуре фактора имеют более-менее равномерное распределение, поиск названия этого фактора затрудняется из-за неясности его особенностей, структурных акцентов. И, наоборот, простая структура фактора, в которой несколько элементарных признаков очевидно доминируют над другими по своей значимости, позволяет определить его название и место в конкретном анализе легко и достаточно надежно. Покажем это на простом примере.

Предположим, имеются данные по результатам факторного анализа (метод главных факторов) до и после вращения пространства главных факторов:

Признак (X_i)	Общие факторы			
	до вращения		после вращения	
	F_1	F_2	V_1	V_2
Обеспеченность основными средствами в расчете на одного среднегодового работника (X_1)	0,463	0,796	0,920	0,087
Уровень энерговооруженности труда (X_2)	0,614	0,411	0,892	0,299
Фондоотдача (X_3)	0,496	0,545	0,410	0,754
Рентабельность продукции (X_4)	0,572	0,378	0,089	0,868
Выработка продукции на одного работника в среднем за год (X_5)	0,723	0,671	0,265	0,570
Уровень реализуемости товарной продукции (X_6)	0,581	0,511	0,372	0,461

Как видим, после вращения в структуре первого фактора наиболее существенные нагрузки сместились к значениям признаков X_1 (обеспеченность основными средствами) и X_2 (уровень энерговооруженности труда). В $\lambda_1 = 2,025$ их удельный вес теперь занимает более 80%. Этот фактор логично назвать «оснащенностью труда средствами производства». В структуре

второго фактора наиболее значимыми стали элементарные признаки: X_1 (рентабельность продукции), X_3 (фондоотдача), X_5 (выработка продукции на одного работника) и в меньшей степени X_6 (реализуемость товарной продукции). Они объясняют 95,1% ($\lambda_2 = 1,954$) дисперсии второго главного фактора и определяют его название – эффективность производственной и коммерческой деятельности.

С теоретической точки зрения вращение общих факторов есть уход от неопределенности пространственного расположения факторной координатной системы. Методы аналитической геометрии дают возможность строго устанавливать центр распределения любой m -мерной совокупности данных, но в то же время неоднозначны относительно поворота системы координат и направления координатных осей. При этом решение оптимизируется при условии, что факторные оси проходят через наиболее плотные скопления точек, так что большинство точек распределения падает на оси или располагается в непосредственной близости от начала координатных осей, приобретая наименьшие значения координат.

Вращение общих факторов может быть ортогональным, когда взаимодействие факторов исключается, и косоугольным, порождающим корреляционные связи латентных факторов. При выборе каждого из двух типов вращения требуется последовательное решение таких вопросов:

1. Какие вычислительные процедуры следует выполнить для вращения факторного пространства?
2. До каких пор (сколько раз) должна повторяться операция вращения?
3. Какой угол лучше установить для поворота?

Наиболее простым является ортогональное вращение. Оно производится умножением матрицы факторных нагрузок на некоторую ортогональную матрицу T , задающую угол поворота, размерностью $m \times m$ по числу общих факторов. Поворот может задаваться по или против часовой стрелки. Например, для матрицы факторных нагрузок A с числом общих факторов $m = 2$:

$$T = \begin{pmatrix} \cos \varphi & \sin \varphi \\ -\sin \varphi & \cos \varphi \end{pmatrix};$$

Вращение по часовой стрелке

$$T = \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}$$

Вращение против часовой стрелки

Если матрица факторных нагрузок содержит данные более чем по двум латентным факторам, строится несколько матриц преобразования T для всех парных комбинаций факторов. Так, для трехмерной матрицы A будут использоваться три матрицы преобразования T :

вращение против часовой стрелки:

$$T_{12} = \begin{pmatrix} \cos \varphi & -\sin \varphi & 0 \\ \sin \varphi & \cos \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad T_{13} = \begin{pmatrix} \cos \varphi & 0 & -\sin \varphi \\ 0 & 1 & 0 \\ \sin \varphi & 0 & \cos \varphi \end{pmatrix};$$

$$T_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \varphi & -\sin \varphi \\ 0 & \sin \varphi & \cos \varphi \end{pmatrix}$$

Полная матрица преобразования для трехмерного случая будет: $T = T_{12} \times T_{13} \times T_{23}$. Этот подход обобщается на случай, когда число общих факторов $m > 3$. Например, для четырехмерной матрицы A полная матрица преобразования T будет произведением уже шести матриц вращения, для всех пар общих факторов:

$$T = T_{12} \times T_{13} \times T_{14} \times T_{23} \times T_{24} \times T_{34} \text{ и т.д.}$$

При условии ортогонального вращения всегда выполнено условие $T^T \times T = E$.

Косоугольное вращение матрицы факторных нагрузок может проводиться поочередным вращением каждого фактора на определенный угол или одновременным вращением всех факторов посредством умножения исходной матрицы нагрузок A на неортогональную матрицу поворота, предусматривающую корреляционные связи латентных факторов.

Вопрос достаточности числа поворотов пространства факторов может решаться субъективно, построением графиков распределений наблюдаемых объектов в пространстве повернутых факторов. Это наиболее трудоемкий путь. Второй путь – применение аналитических средств, с использованием специальных критериев для оценки структуры общих факторов. Все критерии конструктивно базируются на представлении величины дисперсии факторных нагрузок как меры сложности структуры факторов. Дисперсия рассчитывается по формуле:

$$D(F) = \frac{1}{m} \sum_{k=1}^m (a_{jk}^2 - \overline{a_{jk}^2})^2,$$

где a_{jk} – элементы матрицы факторного отображения, величины факторных нагрузок;

m – число общих факторов.

Величина дисперсии будет максимальной, когда из значений квадратов нагрузок равно общности h_j^2 , а в методе главных компонент – единице, и все остальные элементы в строке нулевые. Название критериев качества структуры общих факторов определяет название самого метода вращения.

В ортогональном преобразовании факторного пространства наиболее популярны критерии:

- Квартимакс – критерий вычисляется по формуле:

$$q = \sum_{j=1}^k \frac{\sum_{k=1}^m (a_{jk})^4 - \left(\sum_{k=1}^m a_{jk}^2 \right)^2}{m^2}.$$

Метод «квартимакс» предусматривает вращение факторных осей таким образом, чтобы величины факторных нагрузок максимизировали q . Одновременно учитывается качество структуры всех участвующих в анализе общих факторов.

- Варимакс – рассчитывается V_k критерий качества структуры каждого фактора:

$$V_k = \frac{k \sum_{j=1}^k a_{jm}^4 - \left(\sum_{j=1}^k a_{jm}^2 \right)^2}{k^2}.$$

При помощи метода «варимакс» достигают максимального упрощения в описании столбцов матрицы факторного отображения. Возможно раздельное улучшение структуры факторов. Наилучшим здесь также будет максимальное значение критерия, как и в предыдущем случае.

Если в анализе используется косоугольное вращение пространства факторов, то наиболее часто пользуются критериями облимакс, квартимин, облимин.

Применение любого из критериев качества структуры общих факторов означает, что после каждого поворота факторного пространства проводятся расчет критерия и соответствующая оценка качества структуры факторов. Вращение завершается, когда критерий достигнув максимального (минимального) значения, на следующем шаге алгоритма начинает отклоняться от оптимума.

Выбор угла вращения системы факторов представляется наиболее тонким вопросом при улучшении факторной структуры. Часто его решают субъективно, например, после построения графического изображения разброса точек, отображающих расположение элементарных признаков в пространстве общих факторов. Успех решения вопроса при этом во многом будет определяться опытом самого исследователя. Могут применяться и аналитические подходы. Например, при ортогональном вращении тангенс наилучшего угла для поворота факторных осей рекомендуется находить по формуле:

$$\operatorname{tg} 4\varphi = \frac{2 \sum_{j=1}^k (2a_{ji}a_{jl})(a_{ji}^2 - a_{jl}^2)}{\sum_{j=1}^k (a_{ji}^2 - a_{jl}^2)^2 - (2a_{ji}a_{jl})^2}.$$

где k и l – пара факторов (F_k и F_l) из матрицы факторного отображения A ($k, l = 1, \dots, m$).

Вращение общих факторов дает возможность максимально сократить разброс субъективных суждений относительно названий общих факторов, неизменно присутствующий в факторном анализе. Тем самым методы факторного анализа из группы в большей степени теоретических методов статистики, ненадежных для аналитических выводов, как бы переходят в другую группу методов, гибких и позволяющих достаточно достоверно судить о явлениях и процессах.

ПРИМЕРЫ РЕШЕНИЯ ЗАДАЧ

Пример 1. Определение доверительной области для вектора генеральных средних.

По приводимым условным данным о фонде заработной платы работников централизованных бухгалтерий x_1 и товарообороте обслуживаемых аптек x_2 найти оценки математических ожиданий, дисперсий и коэффициента корреляции, доверительную область для вектора математических ожиданий с надежностью $\gamma=0.95$.

Фонд заработной платы, тыс.руб.	x_1	20	36	28	51	70	45	30	56
Товарооборот, млн.руб.	x_2	3,5	5,4	2,7	9,8	10,1	6,2	2,4	9,5

Решение.

Найдем средние арифметические $\bar{x}_1 = 42$ и $\bar{x}_2 = 6,2$ и перейдем к центрированным величинам $u_{ij} = x_{ij} - \bar{x}_j$.

$$UT = \begin{pmatrix} -22 & -6 & -14 & 9 & 28 & 3 & -12 & 14 \\ -2,7 & -0,8 & -3,5 & 3,6 & 3,9 & 0 & -3,8 & 3,3 \end{pmatrix}$$

Откуда

$$U^T U = \begin{pmatrix} -22 & -6 & -14 & 9 & 28 & 3 & -12 & 14 \\ -2,7 & -0,8 & -3,5 & 3,6 & 3,9 & 0 & -3,8 & 3,3 \end{pmatrix} \begin{pmatrix} -22 & -2,7 \\ -6 & -0,8 \\ -14 & -3,5 \\ 9 & 3,6 \\ 28 & 3,9 \\ 3 & 0 \\ -12 & -3,8 \\ 14 & 3,3 \end{pmatrix} = \begin{pmatrix} 1930 & 346,6 \\ 346,6 & 73,68 \end{pmatrix}$$

$$\hat{S}^{-1} = \frac{1}{n-1} U^T U = \frac{1}{7} \begin{pmatrix} 1930 & 346,6 \\ 346,6 & 73,68 \end{pmatrix} = \begin{pmatrix} 275,714 & 49,514 \\ 49,514 & 10,526 \end{pmatrix}.$$

Итак, оценки дисперсий и средних квадратических отклонений следующие:

$$\hat{s}_1^2 = 275,714; \hat{s}_1 = 16,605; \hat{s}_2^2 = 10,526; \hat{s}_2 = 3,244.$$

Найдем выборочный коэффициент корреляции

$$r = 49,514 / 16,605 \cdot 3,244 = 0,919.$$

Обратная матрица к $\hat{S}$ будет иметь вид

$$\hat{S}^{-1} = \begin{pmatrix} 0,023365 & -0,109909 \\ -0,109909 & 0,6120157 \end{pmatrix}.$$

Тогда согласно уравнению

$$T^2_{\alpha; k; n-k} = \frac{k(n-1)}{n-k} F_{\alpha; k; n-k}$$

запишем:

$$(42 - \mu_1 \quad 6,2 - \mu_2) \begin{pmatrix} 0,023365 & -0,109909 \\ -0,109909 & 0,6120157 \end{pmatrix} \begin{pmatrix} 42 - \mu_1 \\ 6,2 - \mu_2 \end{pmatrix} = \frac{2 \cdot 7}{8 \cdot 6} 5,14,$$

где $F_{\alpha; k; n-k} = F_{0,05; 2; 6} = 5,14$ находим по таблице для $\alpha =$

$1 - \gamma = 0,05$ и чисел степеней свободы $\nu_1 = 2$ и $\nu_2 = 6$

. После преобразований получаем уравнение эллипса

$$0,0234(42 - \mu_1)^2 + 0,0242(42 - \mu_1)(6,2 - \mu_2) + 0,612(6,2 - \mu_2)^2 = 1,499,$$

которое определяет границы доверительной области для вектора $(\mu_1, \mu_2)^T$.

Пример 2. Сравнение вектора генеральных средних с эталоном.

По данным годовых отчетов n промышленных предприятий с уровнем значимости $\alpha = 0,05$ проверить гипотезу о соответствии средних уровней экономических показателей работы группы предприятий (x_1 - объем валовой продукции, тыс.руб.; x_2 - производительность труда, тыс.руб./чел.; x_3 - себестоимость товарной продукции, тыс.руб.) контрольным значениям при известных значениях параметров генеральной совокупности, которая является нормально распределенной.

№	x_1	x_2	x_3
1	3	1,2	2,1
2	4	1,2	2,8
3	5	1,4	3,2
4	5	1,2	4,5
5	5	1,2	4,8
6	5	1,5	4,9
7	6	1,5	5,5
8	7	1,3	6,5
9	10	1,7	8,5
10	10	1,6	8,2
σ^2	4	0,2	4
Парная корреляция	$\rho_{12} = 0,14\sqrt{5}$	$\rho_{12} = 0,12\sqrt{5}$	$\rho_{13} = 0,9$
Контрольные значения μ	7,0	1,5	5,0

Решение.

Исходя из условий задачи требуется проверить гипотезу $H_0: \mu = \mu_0$ против $H_1: \mu \neq \mu_0$, причем $\mu_0 = (7; 1,5; 5)^T$. Здесь матрица Σ генеральных коэффициентов ковариации считается известной:

$$\sigma_1 = 2; \sigma_{12} = \sigma_{21} = \rho_{12}\sigma_1\sigma_2 = 0,14\sqrt{5} \cdot 2\sqrt{0,2} = 0,28;$$

$$\sigma_{13} = \sigma_{31} = \rho_{13}\sigma_1\sigma_3 = 0,9 \cdot 2 \cdot 2 = 3,6;$$

$$\sigma_2 = \sqrt{0,2}; \sigma_{23} = \sigma_{32} = \rho_{23}\sigma_2\sigma_3 = 0,12\sqrt{5} \cdot 2\sqrt{0,2} = 0,24; \sigma_3 = 2.$$

Таким образом,

$$\Sigma = \begin{pmatrix} 4 & 0,28 & 3,6 \\ 0,28 & 0,2 & 0,24 \\ 3,6 & 0,24 & 4 \end{pmatrix}$$

Найдем Σ^{-1} . Обратная к ковариационной матрица имеет вид

$$\Sigma^{-1} = \begin{pmatrix} 1,3551401 & -0,4672897 & -1,1915887 \\ -0,4672897 & 5,5490654 & 0,0876168 \\ -1,1915887 & 0,0876168 & 1,3171728 \end{pmatrix}$$

Критическая область задается следующим неравенством

$$n(\bar{x} - \mu_0)^T \Sigma^{-1} (\bar{x} - \mu_0) > \chi^2_{\alpha; k}.$$

Подставляя в него n , μ_0 , Σ^{-1} , указанные выше, а также $\chi^2_{\alpha; k} = \chi^2_{0,05; 3}$, получим неравенство относительно средних арифметических компонент вектора $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3)$.

Далее найдем наблюдаемое значение вектора $\bar{x}$: $\bar{x} = (6; 1,4; 5,1)^T$. Подставляя его в левую часть неравенства, получим:

$$\chi^2_{\text{набл}} = 10 \begin{pmatrix} -1 & -0,1 & 0,1 \end{pmatrix} \begin{pmatrix} 1,3551401 & -0,4672897 & -1,1915887 \\ -0,4672897 & 5,5490654 & 0,0876168 \\ -1,1915887 & 0,0876168 & 1,3171728 \end{pmatrix} \begin{pmatrix} -1 \\ -0,1 \\ 0,1 \end{pmatrix} =$$

$$= 10 \begin{pmatrix} -1,4275700 & -0,0788551 & 1,3145443 \end{pmatrix} \begin{pmatrix} -1 \\ -0,1 \\ 0,1 \end{pmatrix} = 15,669099 = 15,669.$$

Так как $\chi^2_{\text{набл}} > \chi^2_{0,05; 3}$, то есть $\bar{x}_{\text{набл}}$ принадлежит критической области, гипотеза H_0 отвергается с вероятностью ошибки 0,05. Таким образом, средние уровни экономических показателей работы группы предприятий не соответствуют контрольным цифрам. Анализ причин и следствий обнаруженного несоответствия выясняется планирующими и управляющими органами.

Пример 3. Применение кластерного анализа.

Провести классификацию n объектов, каждый из которых характеризуется двумя признаками.

Номер объекта	1	2	3	4	5	6
x_{1i}	5	6	5	10	11	10
x_{2i}	10	12	13	9	9	7

Решение.

Расположение этих точек на плоскости показано на рис. 1.

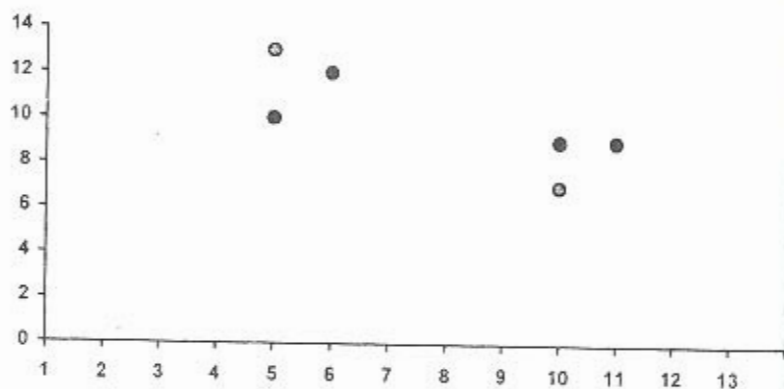
Воспользуемся агломеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами примем обычное евклидово расстояние. Тогда расстояние между объектами 1 и 2, например, равно:

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2,24.$$

а между объектами 1 и 3 -

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$$

очевидно, что, $\rho_{11} = 0$.



Аналогично находим расстояния между всеми шестью объектами и построим матрицу расстояний:

$$R_1 = (\rho(x_i, x_j)) = \begin{pmatrix} 0 & 2,24 & 3 & 5,10 & 6,08 & 5,83 \\ 2,24 & 0 & 1,41 & 5 & 5,83 & 6,40 \\ 3 & 1,41 & 0 & 6,40 & 7,21 & 7,81 \\ 5,10 & 5 & 6,40 & 0 & 1 & 2 \\ 6,08 & 5,83 & 7,21 & 1 & 0 & 2,24 \\ 5,83 & 6,40 & 7,81 & 2 & 2,24 & 0 \end{pmatrix}$$

Из матрицы расстояний следует, что объекты 4 и 5 наиболее близки ($\rho_{45} = 1,00$) и поэтому объединяются в один кластер. После объединения имеем пять кластеров.

Номер кластера	1	2	3	4	5
Состав кластера	1	2	3	4 5	6

Расстояние между кластерами будем находить по принципу «ближайшего соседа», пользуясь формулой пересчета. Так, расстояние между объектом S_1 и кластером $S_{(45)}$, равно:

$$\begin{aligned} \rho_{1(45)} &= \rho(S_1, S_{(45)}) = \frac{1}{2} \rho_{14} + \frac{1}{2} \rho_{15} - \frac{1}{2} |\rho_{14} - \rho_{15}| = \\ &= \frac{1}{2} (5,10 + 6,08) - \frac{1}{2} |5,10 - 6,08| = 5,10. \end{aligned}$$

Хорошо видно, что расстояние $\rho_{1(45)}$ равно расстоянию от объекта 1 до ближайшего к нему объекта, входящего в кластер $S_{(45)}$, то есть $\rho_{1(45)} = \rho_{14} = 5,10$. Тогда матрица расстояний равна:

$$R_2 = \begin{pmatrix} 0 & 2,24 & 3 & 5,10 & 5,83 \\ 2,24 & 0 & 1,41 & 5 & 6,40 \\ 3 & 1,41 & 0 & 6,40 & 7,81 \\ 5,10 & 5 & 6,40 & 0 & 2 \\ 5,83 & 6,40 & 7,81 & 2 & 0 \end{pmatrix}$$

Объединим объекты 2 и 3, имеющие наименьшее расстояние $\rho_{23} = 1,41$ и получим четыре кластера.

Номер кластера	1	2	3	4
Состав кластера	1	2 3	4 5	6

Вновь найдем матрицу расстояний. Для этого рассчитаем расстояние до кластера $S_{(23)}$. Воспользуемся матрицей расстояний R_2 .

Например, расстояние между кластерами $S_{(45)}$ и $S_{(23)}$ равно:

$$\rho_{(45),(23)} = \frac{1}{2} \rho_{(45)2} + \frac{1}{2} \rho_{(45)3} - \frac{1}{2} |\rho_{(45)2} - \rho_{(45)3}| = \frac{5}{2} + \frac{6,40}{2} - \frac{1,40}{2} = 5.$$

Проведя аналогичные расчеты, получим:

$$R_3 = \begin{pmatrix} 0 & 2,24 & 5,10 & 5,83 \\ 2,24 & 0 & 5 & 6,40 \\ 5,10 & 5 & 0 & 2 \\ 5,83 & 6,40 & 2 & 0 \end{pmatrix}.$$

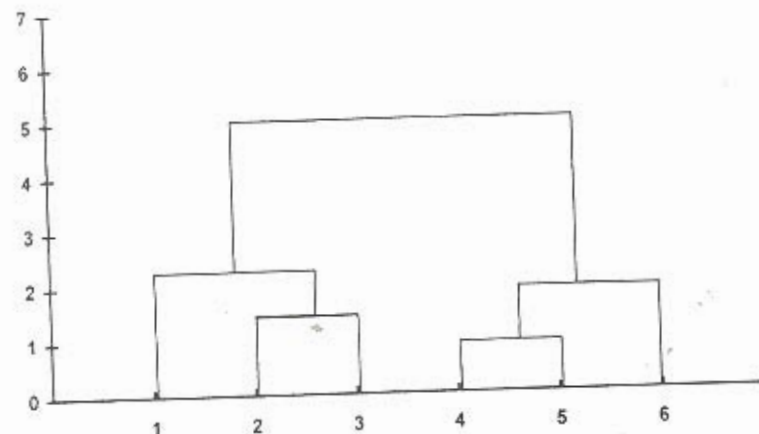
Объединяем кластеры $S_{(45)}$ и S_6 , расстояние между которыми, согласно матрице R_3 , наименьшее: $\rho_{(45)6} = 2$. В результате этого получим три кластера: S_1 , $S_{(23)}$, $S_{(456)}$. Матрица расстояний будет иметь вид:

$$R_4 = \begin{pmatrix} 0 & 2,24 & 5,10 \\ 2,24 & 0 & 5 \\ 5,10 & 5 & 0 \end{pmatrix}.$$

Объединим теперь кластеры S_1 и $S_{(23)}$, расстояние между которыми равно $\rho_{1(23)} = 2,24$. В результате получим два кластера: $S_{(123)}$ и $S_{(456)}$, расстояние между которыми, найденное по принципу «ближайшего соседа», равно:

$$\rho_{(123)(456)} = 5,0.$$

Результаты иерархической классификации объектов представлены на рис. 2 в виде дендрограммы.



В задаче предпочтение следует отдать предпоследнему этапу классификации, когда все объекты объединены в два кластера: $S_{(123)}$ и $S_{(456)}$, что наглядно видно на рис. 2.

Пример 4. Проведение анализа однофакторного дисперсионного комплекса.

Необходимо выяснить, отличаются ли четыре способа рекламирования товара по влиянию на объем его продажи. Для этого в каждом из случайно отобранных четырех однотипных городов (в них использовались различные способы рекламы) были собраны сведения об объемах продажи товара (в денежных единицах) в четырех магазинах и вычислены соответствующие выборочные характеристики.

Способ рекламы	1	2	3	4
Объем продаж	140	150	148	150
	144	149	149	155
	142	152	146	154
	145	150	147	152
n_i	4	4	4	4
$\bar{y}_i$	142,75	150,25	147,5	152,75
s_i^2	4,92	1,58	1,67	4,92

Можно ли на 5%-ом уровне значимости считать влияние доказанным?

Решение.

Здесь фактором А является способ рекламы; зафиксированы четыре его уровня, и выясняется, различаются ли по своему влиянию именно эти уровни, - это модель М1 однофакторного анализа.

Проверим ключевую гипотезу дисперсионного анализа H_0 . Для этого вычислим: $D_x^2 = 220,19$, $D_z^2 = 39,27$, $D_y^2 = 259,46$; убедившись в справедливости равенства $D_y^2 = D_x^2 + D_z^2$, найдем оценки факторной и случайной дисперсий, для которых определим степени свободы $k_x = r - 1 = 4 - 1 = 3$, $k_z = n - r = 16 - 4 = 12$; $k_y = n - 1 = 16 - 1 = 15$; $S_x^2 = 220,19/3 = 73,397$; $S_z^2 = 39,27/12 = 3,273$; $S_y^2 = 259,46/15 = 17,297$.

Из отношения факторной дисперсии и общей дисперсии следует, что влиянием способа рекламы объясняется 84,9% вариации объема продаж ($\eta^2 = D_x^2 / D_y^2 = 220,19/259,46 = 0,849$).

Проверяемую гипотезу отклоняем, то есть считаем, что не все четыре способа рекламирования продукции одинаково эффективны, в силу того, что

$$F_{\text{кр}} = \frac{S_x^2 / 3}{S_z^2 / 12} = \frac{73,397 / 3}{3,273 / 12} = 22,43,$$

а соответствующее табличное значение $F_{\text{кр}} = F_{\alpha, r-1, n-r} = F_{0,05,3,12} = 3,49$, то есть $F_{\text{набл}} > F_{\text{кр}}$.

Изменим условие задачи. Предположим, что способы рекламирования товара не заранее фиксированы, а выбраны случайным образом из всего набора способов. Тогда выяснение вопроса о том, сравнимы ли по эффективности все способы рекламирования (именно все, а не только четыре, которые попали в выборку), сводится к проверке основной гипотезы модели М2 однофакторного дисперсионного анализа. Критерий ее проверки такой же, как и в модели М1. Так как условие принятия ги-

потезы не выполняется, гипотезу забраковываем, по крайней мере до получения дополнительных данных: считаем, что способы рекламирования товаров (во всем наборе этих способов) отличаются по эффективности.

Пример 5. Проведение анализа двухфакторного дисперсионного комплекса.

В двухфакторном комплексе проведено исследование влияния возраста (фактор В) и стажа работников (фактор А) на производительность труда (результативный признак у). При $\alpha = 0,05$ проверить индивидуальное влияние факторов, их совместное влияние в условиях моделей М1, М2, М3 двухфакторного дисперсионного комплекса. В приведенных данных фактор А имеет уровни: A_1 : 1 - 4 лет, A_2 : 4 - 7 лет, A_3 : 7 - 10 лет, A_4 : более 10 лет; а фактор В: B_1 : 25 - 35 лет, B_2 : 35 - 45 лет, B_3 : 45 - 55 лет.

В \ А	A_1	A_2	A_3	A_4
B_1	20 20 22 19 20	31 32 34 30 32	39 40 41 35 35	42 40 41 41 40
B_2	23 19 20 25 20	29 30 31 20 31	36 40 41 42 45	28 31 35 40 36
B_3	21 23 20 18 19	25 26 19 25 26	24 25 24 25 24	31 24 20 25 32

Решение.

Рассчитаем внутригрупповые средние по факторам А и В, а также средние в группах, образованных комбинацией уровней этих факторов.

$$\bar{y}_j: \bar{y}_1 = 20,6; \bar{y}_2 = 28,067; \bar{y}_3 = 34,4; \bar{y}_4 = 33,73;$$

$$\bar{y}_k: \bar{y}_1 = 32,7; \bar{y}_2 = 31,1; \bar{y}_3 = 23,8;$$

$$\bar{y}_{jk}: \bar{y}_{11} = 20,2; \bar{y}_{12} = 21,4; \bar{y}_{13} = 20,2;$$

$$\bar{y}_{21} = 31,8; \bar{y}_{22} = 28,2; \bar{y}_{23} = 24,2;$$

$$\bar{y}_{31} = 38,0; \bar{y}_{32} = 40,8; \bar{y}_{33} = 24,4;$$

$$\bar{y}_{41} = 40,8; \bar{y}_{42} = 34,0; \bar{y}_{43} = 26,4;$$

$$\bar{y} = 29,2.$$

Логическим продолжением расчетов является вычисление сумм квадратов отклонений по факторам, их взаимодействию, случайным факторам:

$$D_A^2 = \sum_{j=1}^4 (\bar{y}_j - \bar{y})^2 n_j = \left\{ (20,6 - 29,2)^2 + (28,067 - 29,2)^2 + \dots \right\} \times 15 = 1842,56;$$

$$D_B^2 = \sum_{k=1}^3 (\bar{y}_k - \bar{y})^2 n_k = \left\{ (32,7 - 29,2)^2 + (31,1 - 29,2)^2 + (23,8 - 29,2)^2 \right\} \times 20 = 900,4;$$

$$D_{AB}^2 = \sum_{k=1}^3 \sum_{j=1}^4 (\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2 n_{jk} = \left\{ (20,2 - 20,6 - 32,7 + 29,2)^2 + \dots \right\} \times 5 = 537,94;$$

$$D_z^2 = \sum \sum \sum (\bar{y}_{ijk} - \bar{y}_{jk})^2 = 441,2;$$

$$D_y^2 = \sum \sum \sum (\bar{y}_{ijk} - \bar{y})^2 = (20 - 29,2)^2 + (20 - 29,2)^2 + \dots = 3721,6.$$

Полученные результаты вычислений и критические значения статистики F удобно для обозрения и сравнения расположить в таблице.

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Оценка	Модель M1		Модель M2		Модель M3 (A - случайный фактор)	
				F _{набл}	F _{кр}	F _{набл}	F _{кр}	F _{набл}	F _{кр}
A	1842,53	3	614,2	66,8	2,84	6,8	4,76	66,8	2,84
B	900,4	2	450,2	48,9	3,24	5,0	5,14	5,0	5,14
AB	537,47	6	89,6	9,7	2,34	9,7	2,34	9,7	2,34
Случайные	441,2	48	9,2	-	-	-	-	-	-
Общий	3721,6	59	63,08	-	-	-	-	-	-

В таблице рассмотрены все модели. Разумеется, на практике следует заранее выбирать одну из них, в зависимости от того, желательно ли распространить статистические выводы на всю совокупность по какому-либо фактору или выводы достаточно сделать относительно включенных в модель уровней факторов.

Таким образом, возможны следующие статистические выводы.

Если наблюдения проводятся в условиях модели M1, то результаты анализа показывают существенное влияние факторов A и B, а также их совместного действия на результативный признак - показатель производительности труда.

Если наблюдения проводятся в условиях модели M2, то результаты анализа показывают существенное влияние фактора A и взаимодействия AB, влияние же фактора B данными обследования не доказано. Следовательно, можно утверждать, что на производительность труда изучаемой генеральной совокупности работников существенно влияют стаж и совместное действие возраста и стажа; влияние возраста в результате эксперимента не доказано.

Наконец, если обследование проводится в условиях смешанной модели, причем для фактора A уровни были получены случайно, в результате выборки, по которой фиксировались единицы отбора по фактору B, то данные расчетов показывают существенное влияние на производительность труда действия фактора A и взаимодействия факторов AB. Относительно же действия фактора B в результате проведенного анализа сказать ничего нельзя.

Пример 6. Применение компонентного анализа.

По данным о численности (x_1) и фонде зарплаты (x_2) пяти строительных организаций провести компонентный анализ.

$$X = \begin{pmatrix} 3 & 4 \\ 6 & 5 \\ 8 & 9 \\ 2 & 3 \\ 7 & 6 \end{pmatrix}$$

Решение.

Рассчитаем выборочные характеристики переменных x_1 и x_2 :

$$\overline{x_1} = 5,2 \quad s_1 = 2,315$$

$$\overline{x_2} = 5,4 \quad s_2 = 2,059$$

Выборочный коэффициент корреляции равен:

$$\rho = \frac{\overline{x_1 x_2} - \overline{x_1} \cdot \overline{x_2}}{s_1 \cdot s_2} = \frac{32,4 - 5,2 \cdot 5,4}{2,315 \cdot 2,059} = 0,906.$$

Преобразуем матрицу X в матрицу нормированных значений Z , с элементами $z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}$, где $i = 1, 2, 3, 4, 5; j = 1, 2$.

$$Z = \begin{pmatrix} -0,950 & -0,680 \\ 0,346 & -0,194 \\ 1,210 & 1,748 \\ -1,382 & -1,166 \\ 0,778 & 0,291 \end{pmatrix}$$

Матрица парных коэффициентов корреляций имеет вид:

$$R = \begin{pmatrix} 1 & 0,906 \\ 0,906 & 1 \end{pmatrix}$$

Для определения собственных значений матрицы R рассмотрим характеристическое уравнение

$$\begin{vmatrix} \lambda - 1 & -0,906 \\ 0,906 & \lambda - 1 \end{vmatrix} = \begin{vmatrix} 1 - \lambda & 0,906 \\ 0,906 & 1 - \lambda \end{vmatrix} = 0.$$

Отсюда следует,

$$(1 - \lambda)^2 - (0,906)^2 = 0 \text{ или } \boxed{}$$

Так как по условию компонентного анализа $\lambda_1 > \lambda_2$, то $\lambda_1 = 1,906$, $\lambda_2 = 0,094$, где λ_1 , λ_2 соответственно дисперсии и вклад 1-й и 2-й главных компонент в суммарную дисперсию, равную $\lambda_1 + \lambda_2 = k = 2$.

Относительный вклад компонент в суммарную дисперсию равен:

$$\frac{\lambda_1}{k} 100\% = \frac{1,906}{2} 100\% = 95,3\%,$$

$$\frac{\lambda_2}{k} 100\% = \frac{0,094}{2} 100\% = 4,7\%.$$

Таким образом,

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} 1,906 & 0 \\ 0 & 0,094 \end{pmatrix}$$

Определим матрицу собственных векторов из уравнения $(R - \lambda E)V = 0$. Собственный вектор V_1 находим из условия:

$$\begin{pmatrix} 1 - \lambda_1 & \rho \\ \rho & 1 - \lambda_1 \end{pmatrix} \begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad V_1 = \begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix}$$

Подставляя полученные значения получим:

$$\begin{pmatrix} -0,906 & 0,906 \\ 0,906 & -0,906 \end{pmatrix} \begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Откуда $-0,906v_{11} + 0,906v_{21} = 0$ или $v_{11} = v_{21} = 1$, то есть

$$V_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Нормированный собственный вектор, соответствующий λ_1 , равен:

$$U_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Собственный вектор V_2 найдем решая уравнение:

$$\begin{pmatrix} 0,906 & 0,906 \\ 0,906 & 0,906 \end{pmatrix} \begin{pmatrix} V_{12} \\ V_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

откуда $0,906v_{12} + 0,906v_{22} = 0$ или $-v_{12} = v_{22}$, то есть

$$V_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Нормированный собственный вектор, соответствующий λ_2 , равен:

$$U_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Тогда нормированная матрица собственных векторов имеет вид:

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 0,707 & -0,707 \\ 0,707 & 0,707 \end{pmatrix}.$$

Матрицу факторных нагрузок найдем по формуле:

$$A = U\Lambda^{1/2},$$

$$\text{где } \Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix}.$$

Подставив полученные значения, получим:

$$A = \begin{pmatrix} 0,707 & -0,707 \\ 0,707 & 0,707 \end{pmatrix} \times \begin{pmatrix} 1,3806 & 0 \\ 0 & 0,3066 \end{pmatrix} = \begin{pmatrix} 0,9761 & -0,2167 \\ 0,9761 & 0,2167 \end{pmatrix}.$$

Матрицу факторных нагрузок используют для интерпретации главных компонент, так как ее элементы $a_{jv} = \rho_{jv}$ характеризуют степень связи между x_j -м признаком и f_v -й главной компонентой. В рассматриваемой задаче первая главная компонента тесно связана с показателями x_1 и x_2 , f_1 характеризует размер предприятия.

Матрицу значений главных компонент F можно получить по формуле:

$$F = Z(A^T)^{-1}.$$

Предварительно найдем обратную матрицу $(A^T)^{-1}$. Так как

$$A^T = \begin{pmatrix} 0,9761 & 0,9761 \\ -0,2167 & 0,2167 \end{pmatrix}, \text{ то } (A^T)^{-1} = \begin{pmatrix} 0,5123 & -2,3076 \\ 0,5123 & 2,3076 \end{pmatrix}.$$

Тогда

$$F = \begin{pmatrix} -0,950 & -0,680 \\ 0,346 & -0,194 \\ 1,210 & 1,748 \\ -1,382 & -1,166 \\ 0,778 & 0,291 \end{pmatrix} \times \begin{pmatrix} 0,5123 & -2,3076 \\ 0,5123 & 2,3076 \end{pmatrix} = \begin{pmatrix} -0,835 & 0,623 \\ 0,078 & -1,246 \\ 1,515 & 1,242 \\ 1,305 & 0,498 \\ 0,548 & -1,124 \end{pmatrix}.$$

Как уже отмечалось, матрица F , которую мы получили, характеризует пять строительных организаций в пространстве главных компонент. Ее можно использовать в задачах класси-

фикации и регрессионного анализа. Например, классификация организаций по первой главной компоненте f_1 , характеризующей размер предприятия, позволяет их ранжировать в порядке возрастания - 4; 1; 2; 5; 3, что согласуется с матрицей X.

Пример 7. Применение методов определения оценок общностей для составления редуцированной матрицы парных коэффициентов корреляции в факторном анализе.

По данным матрицы парных коэффициентов корреляции R требуется определить оценки общностей h_j^2 , используя следующие методы:

- 1) наибольшего элемента по строке (столбцу);
- 2) среднего коэффициента корреляции;
- 3) триад;
- 4) первого центроидного фактора.

$$R = \begin{pmatrix} 1 & 0,846 & 0,805 & 0,859 & 0,473 & 0,398 & 0,301 & 0,382 \\ 0,846 & 1 & 0,881 & 0,826 & 0,376 & 0,326 & 0,277 & 0,415 \\ 0,805 & 0,881 & 1 & 0,801 & 0,380 & 0,319 & 0,237 & 0,345 \\ 0,859 & 0,826 & 0,801 & 1 & 0,436 & 0,329 & 0,327 & 0,365 \\ 0,473 & 0,376 & 0,380 & 0,436 & 1 & 0,762 & 0,730 & 0,629 \\ 0,398 & 0,326 & 0,319 & 0,329 & 0,762 & 1 & 0,583 & 0,577 \\ 0,301 & 0,277 & 0,237 & 0,327 & 0,730 & 0,583 & 1 & 0,539 \\ 0,382 & 0,415 & 0,345 & 0,365 & 0,629 & 0,577 & 0,539 & 1 \end{pmatrix}$$

Решение.

1. Запишем матрицу R^h , используя метод наибольшего элемента по строке (столбцу). Для этого перепишем матрицу R без ее диагональных элементов. В качестве диагональных элементов проставляются наибольшие элементы по строке. Например, для первой строки наибольший элемент равен 0,859. Следовательно, оценка h_1^2 равна 0,859. Обычно, для того, чтобы отличать оценки общностей от внедиагональных элементов их помещают в скобки.

$$R^h = \begin{pmatrix} (0,859) & 0,846 & 0,805 & 0,859 & 0,473 & 0,398 & 0,301 & 0,382 \\ 0,846 & (0,881) & 0,881 & 0,826 & 0,376 & 0,326 & 0,277 & 0,415 \\ 0,805 & 0,881 & (0,881) & 0,801 & 0,380 & 0,319 & 0,237 & 0,345 \\ 0,859 & 0,826 & 0,801 & (0,859) & 0,436 & 0,329 & 0,327 & 0,365 \\ 0,473 & 0,376 & 0,380 & 0,436 & (0,762) & 0,762 & 0,730 & 0,629 \\ 0,398 & 0,326 & 0,319 & 0,329 & 0,762 & (0,762) & 0,583 & 0,577 \\ 0,301 & 0,277 & 0,237 & 0,327 & 0,730 & 0,583 & (0,730) & 0,539 \\ 0,382 & 0,415 & 0,345 & 0,365 & 0,629 & 0,577 & 0,539 & (0,629) \end{pmatrix}$$

2. Определим оценки общностей методом среднего коэффициента корреляции по строке (столбцу).

Для первой строки найдем среднюю арифметическую ее элементов (без диагонального элемента первой строки матрицы R):

$$\frac{1}{7}(0,846 + 0,805 + 0,859 + 0,473 + 0,398 + 0,301 + 0,382) = \frac{1}{7}4,064 = 0,581.$$

Аналогично рассчитываются оценки общностей и для всех остальных строк матрицы парных коэффициентов корреляции R.

Сведем их в таблицу.

Значение оценок общностей, полученные методом среднего коэффициента корреляции

	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
Значение h_i^2	0,581	0,564	0,538	0,563	0,541	0,471	0,428	0,465

3. Для того, чтобы реализовать метод триад при определении оценки общности h_1^2 в первой строке матрицы выберем два наибольших коэффициента корреляции. Это $r_{14} = 0,859 = r_{41}$ и $r_{12} = 0,846 = r_{21}$. Найдем $r_{41} = r_{42} = 0,826$. Затем определим необходимое значение оценки общности:

$$h_1^2 = \frac{0,859 \cdot 0,846}{0,826} = 0,880.$$

Составим таблицу значений $r_{jk}, r_{jl}, r_{kl}, \hat{h}_i^2$.

Значения $r_{jk}, r_{jl}, r_{kl}, \hat{h}_j^2$

	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
r_{jk}	0,859	0,881	0,881	0,859	0,762	0,762	0,730	0,629
r_{jl}	0,846	0,846	0,805	0,826	0,730	0,583	0,583	0,577
r_{kl}	0,826	0,805	0,846	0,846	0,583	0,730	0,762	0,762
$\hat{h}_j^2$	0,880	0,926	0,838	0,839	0,954	0,609	0,559	0,476

4. Используя метод первого центроидного фактора найдем оценки общностей $\hat{h}_j^2$.

а) Методом выбора наибольшего элемента по строке (столбцу) составим матрицу (первый пункт данного примера);

б) Определим сумму всех коэффициентов корреляции новой матрицы (из пункта первого данного примера)

$$\sum_{j=1}^8 r_{jj} + \sum_{m,l=1}^8 r_{ml} = 0,859 + 0,881 + \dots + 0,730 + 0,629 + 2(0,846 + 0,805 + \dots + 0,577 \cdot 0,539) = 35,411;$$

в) Вычислим квадраты сумм элементов строк новой матрицы (из пункта первого данного примера). Например, для первой строки искомый квадрат суммы элементов будет равен

$$\left(\sum_{l=1}^8 r_{1l} \right)^2 = (4,064 + 0,859)^2 = 24,236.$$

Отсюда

$$\epsilon_{h_1}^2 = \frac{24,236}{35,411} = 0,684.$$

Составим таблицу значений квадратов сумм элементов строк новой матрицы и вычисленных значений оценок.

Значения квадратов сумм и оценок общностей

	$\hat{h}_1^2$	$\hat{h}_2^2$	$\hat{h}_3^2$	$\hat{h}_4^2$	$\hat{h}_5^2$	$\hat{h}_6^2$	$\hat{h}_7^2$	$\hat{h}_8^2$
$\left(\sum_{m,l=1}^8 r_{ml} \right)^2$	24,23 6	23,31 0	21,61 3	23,05 9	20,68 4	16,45 1	13,94 3	15,06 2
$\epsilon_{h_j}^2$	0,684	0,658	0,610	0,651	0,584	0,465	0,394	0,425

ПРАКТИЧЕСКАЯ ЧАСТЬ

Приводимые ниже задачи включаются в состав контрольной работы для студентов заочного отделения специальности «Статистика».

Контрольная работа предполагает решение пяти задач в соответствии с вариантом, определяемым последней цифрой номера зачетной книжки, и выполнение лабораторной работы с применением ПК в соответствии с вариантом, определяемым по алфавитному списку группы.

Перед выполнением контрольной работы следует изучить теоретический материал по соответствующим темам курса и проанализировать примеры решения типовых задач. Затем, руководствуясь основными расчетными алгоритмами и формулами, решить задачи. Причем, расчеты, используемые формулы и пояснения к ним должны приводиться полностью. Результатами выполнения лабораторной работы должны явиться распечатки промежуточных и итоговых данных, получаемых при проведении корреляционно-регрессионного анализа, применения метода главных компонент, факторного и кластерного анализа. Необходимое условие качественного выполнения лабораторной работы - экономико-статистическая интерпретация полученных результатов и соответствующие этой интерпретации выводы.

Варианты контрольной работы

№ варианта	Последняя цифра номера зачетной книжки	Задание
1	1	1.1; 2.1.1.; 2.2.1.; 3.1.; 4.1.
2	2	1.2; 2.1.2.; 2.2.2.; 3.2.; 4.2.
3	3	1.3.; 2.1.3.; 2.2.3.; 3.3.; 4.3.
4	4	1.4.; 2.1.4.; 2.2.4.; 3.4.; 4.4.
5	5	1.5.; 2.1.5.; 2.2.5.; 3.5.; 4.5.
6	6	1.6.; 2.1.6.; 2.2.6.; 3.6.; 4.6.
7	7	1.7.; 2.1.7.; 2.2.7.; 3.7.; 4.7.
8	8	1.8.; 2.1.8.; 2.2.8.; 3.8.; 4.8.
9	9	1.9.; 2.1.9.; 2.2.9.; 3.9.; 4.9.
10	0	1.10.; 2.1.10.; 2.2.10.; 3.10.; 4.10.

1. Задачи по кластерному анализу

Задача 1.1. По иерархическому агломеративному алгоритму провести классификацию $n = 6$ предприятий машиностроения, деятельность которых характеризуется показателями: x_1 - рентабельность, % и x_2 - производительность труда, тыс.руб./чел.

Номер предприятия	1	2	3	4	5	6
x_1	23,4	17,5	9,7	18,2	6,6	8,0
x_2	9,1	5,2	5,5	9,4	7,5	5,7

В качестве расстояния между объектами принять обычное евклидово расстояние. Расстояние между кластерами определять по принципу: а) «ближайшего соседа»; б) «дальнего соседа»; в) «центрам тяжести» групп.

Классификацию проводить, не нормализуя предварительно исходные данные.

Задача 1.2. Решить задачу 1.1., приняв в качестве расстояния между объектами взвешенное евклидово расстояние с весами $\omega_1 = 0,3$ и $\omega_2 = 0,7$.

Задача 1.3. По иерархическому агломеративному алгоритму провести классификацию $n = 5$ хозяйств, работа которых характеризуется показателями объема реализованной продукции, тыс. руб./га: x_1 - растениеводства и x_2 - животноводства.

Номер хозяйства	1	2	3	4	5
x_1	2,49	1,51	1,17	1,67	2,73
x_2	0,38	0,51	0,28	0,29	0,34

В качестве расстояния между объектами принять обычное евклидово расстояние. Расстояние между кластерами опреде-

лать по принципу: а) «ближайшего соседа»; б) «дальнего соседа»; в) «центрам тяжести» групп.

Классификацию проводить, не нормализуя предварительно исходные данные.

Задача 1.4. По иерархическому агломеративному алгоритму провести классификацию $n = 6$ регионов по уровню медицинского обслуживания населения, который характеризуется показателями: x_1 - число врачей на 10 тыс. жителей и x_2 - число больничных коек на 10 тыс. жителей.

Номер региона	1	2	3	4	5	6
x_1	34,8	31,2	32,1	35,7	30,2	34,2
x_2	126	112	123	128	115	123

В качестве расстояния между объектами принять взвешенное евклидово расстояние с весами $\omega_1 = 0,4$ и $\omega_2 = 0,6$.

Расстояние между кластерами определить по принципу: а) «ближайшего соседа»; б) «дальнего соседа»; в) «центрам тяжести» групп.

Классификацию проводить, не нормализуя предварительно исходные данные.

Задача 1.5. По иерархическому агломеративному алгоритму провести классификацию $n = 5$ хозяйств района, работа которых оценивается по показателям: x_1 - производительность труда (тыс. руб./чел.); объем реализованной продукции, тыс. руб./га; x_2 - растениеводства и x_3 - животноводства.

Номер хозяйства	1	2	3	4	5
x_1	8,22	4,33	6,43	6,39	4,92
x_2	0,25	0,49	0,51	0,27	0,32
x_3	0,41	0,38	0,51	0,42	0,55

В качестве расстояния между объектами принять обычное евклидово расстояние.

Расстояние между кластерами определить по принципу: а) «ближайшего соседа»; б) «дальнего соседа»; в) «центрам тяжести» групп.

Классификацию проводить, не нормализуя предварительно исходные данные.

Задача 1.6. Решить задачу 1.5., приняв в качестве расстояния между объектами взвешенное евклидово расстояние с весами $\omega_1 = 0,2$; $\omega_2 = 0,4$ и $\omega_3 = 0,4$.

Задача 1.7. По иерархическому агломеративному алгоритму провести классификацию $n = 6$ машиностроительных предприятий, деятельность которых характеризуется показателями: x_1 - производительность труда (тыс.руб./чел.); x_2 - удельный вес потерь от брака, %; x_3 - фондоотдача активной части основных производственных фондов (руб./руб.).

Номер предприятия	1	2	3	4	5	6
x_1	9,4	6,6	7,4	10,0	6,6	9,1
x_2	0,15	0,48	0,62	0,32	0,50	0,90
x_3	1,90	0,88	1,09	2,62	1,32	1,89

В качестве расстояния между объектами принять обычное евклидово расстояние.

Расстояние между кластерами определить по принципу: а) «ближайшего соседа»; б) «дальнего соседа»; в) «центрам тяжести» групп.

Классификацию проводить, нормализуя и не нормализуя предварительно исходные данные.

Задача 1.8. Решить задачу 1.1., приняв в качестве расстояния между объектами взвешенное евклидово расстояние с весами $\omega_1 = 0,3$ и $\omega_2 = 0,7$.

Расстояние между кластерами определить по принципу: а) «ближайшего соседа»; б) «дальнего соседа»; в) «центрам тяжести» групп.

Классификацию проводить, нормализуя предварительно исходные данные.

Задача 1.9. Решить задачу 1.5., приняв в качестве расстояния между объектами взвешенное евклидово расстояние с весами $\omega_1 = 0,2$; $\omega_2 = 0,4$ и $\omega_3 = 0,4$.

Расстояние между кластерами определять по принципу «средней связи».

Классификацию проводить, нормализуя предварительно исходные данные.

Задача 1.10. Решить задачу 1.3. решить, приняв в качестве расстояния между объектами взвешенное евклидово расстояние с весами $\omega_1 = 0,6$; $\omega_2 = 0,4$.

Расстояние между кластерами определять по принципу «средней связи».

Классификацию проводить, нормализуя предварительно исходные данные.

2. Задачи по дисперсионному анализу

2.1. Однофакторный комплекс

Задача 2.1.1. Исследовать влияние численности населения места реализации продукции на удельный объем выручки (тыс.руб./чел.). Принять $\alpha = 0,05$.

№ опыта	до 10 тыс.чел.	10-50 тыс.чел.	50-100 тыс.чел.
1	29	30	33
2	28	29	32
3	29	32	33

Задача 2.1.2. Исследовать влияние количества осадков за год на урожайность пшеницы (кг / га). Принять $\alpha = 0,01$.

№ опыта	Количество осадков (мм)			
	250-260	260-270	270-280	280-290
1	31	27	35	33
2	30	29	34	37

Задача 2.1.3. Исследовать на уровне значимости $\alpha = 0,05$ существенность влияния размеров предприятий на производительность труда.

№ опыта	Предприятия			
	мелкие	средние	крупные	крупнейшие
1	88	87	93	92
2	89	86	93	94

Задача 2.1.4. На уровне значимости $\alpha = 0,05$ исследовать влияние цвета краски на срок службы покрытия, мес.

№ опыта	синий	зеленый	красный
1	64	59	57
2	62	61	58
3	64	58	60

Задача 2.1.5. Исследовать влияние времени реализации продукции на объем выручки, тыс. руб. Принять $\alpha = 0,01$

№ опыта	I квартал	II квартал	III квартал	IV квартал
1	149	150	140	153
2	148	149	144	153

Задача 2.1.6. Исследовать влияние стажа работы рабочих на производительность труда в механическом цехе завода при $\alpha = 0,05$.

№ опыта	Стаж работы рабочих		
	до 10 лет	от 10 до 15 лет	от 15 до 20 лет
1	154	158	164
2	158	162	162
3	157	164	163

Задача 2.1.7. На уровне значимости $\alpha = 0,05$ проверить существенность влияния температуры на производительность труда, кг/ч.

№ опыта	150-160°C	160-170°C	170-180°C
1	96	99	103
2	99	100	104
3	97	96	102

Задача 2.1.8. Исследовать влияние времени внесения удобрения на урожайность кукурузы, ц/га. Принять $\alpha = 0,01$.

№ опыта	Число дней с момента посадки			
	0-5	5-10	10-15	15-20
1	51	49	50	48
2	52	47	51	49

Задача 2.1.9. При исследовании зависимости себестоимости продукции на предприятии от производительности труда найдены следующие суммы квадратов отклонений: $D^2_x = 36$; $D^2_y = 86$. Среди 36 наблюдений выделено 6 групп. При $\alpha =$

0,01 проверить влияние производительности труда на себестоимость товарной продукции.

Задача 2.1.10. При анализе влияния влажности воздуха в инкубаторе на продолжительность инкубационного периода у цыплят получены следующие суммы квадратов отклонений: $D^2_x = 38$; $D^2_y = 72$. При $\alpha = 0,05$ проверить влияние влажности воздуха в инкубаторе на продолжительность инкубационного периода, если среди 32 наблюдений выделены 4 группы.

2.2. Двухфакторный комплекс

Задача 2.2.1. В двухфакторном комплексе приводится сменная выработка рабочего в зависимости от типа станка (фактор А) и стажа его работы (фактор В). При $\alpha = 0,01$ проверить влияние фактора А. При $\alpha = 0,05$ проверить существенность влияния фактора В на сменную выработку.

	A ₁	A ₂	A ₃
B ₁	195	198	202
B ₂	196	201	203
B ₃	198	202	204

Задача 2.2.2. В двухфакторном комплексе приводятся данные об объеме выручки от реализации продукции (фактор А) и района реализации (фактор В). При $\alpha = 0,05$ проверить существенность влияния факторов А и В на объем выручки.

	A ₁	A ₂	A ₃	A ₄
B ₁	86	88	93	92
B ₂	87	89	93	94

Задача 2.2.3. При анализе влияния средней температуры воздуха (фактор В) и количества осадков в июне месяце (фактор А) на урожайность пшеницы (ц/га) получен двухфакторный комплекс. При $\alpha = 0,01$ проверить существенность влияния фак-

тор В. При $\alpha = 0,05$ проверить существенность влияния фактора А на урожайность пшеницы.

Средняя месячная температура (В)	Количество осадков, мм (А)		
	45-46	46-47	47-48
20-22	30	31	35
22-24	31	33	36
24-26	31	35	37

Задача 2.2.4. При исследовании зависимости товарооборота центральной районной аптеки от товарооборота (фактор А) и штатной численности прикрепленной аптечной сети (фактор В) получен двухфакторный комплекс. При $\alpha = 0,01$ проверить существенность влияния фактора В на товароборот центральной районной аптеки. При $\alpha = 0,05$ проверить существенность влияния фактора А.

	B_1	B_2	B_3
A_1	195	198	202
A_2	196	201	203
A_3	198	202	204

Задача 2.2.5. В результате анализа двухфакторного комплекса найдено, что суммы квадратов отклонений, обусловленные действием факторов А и В и их взаимодействием (АВ), равны следующим величинам: $D^2_A = 42$; $D^2_B = 47$; $D^2_{AB} = 11$; $D^2_Y = 114$. Фактор А имеет $r = 3$ группы, фактор В - $m = 5$ групп, в каждой клетке комплекса имеется по 3 наблюдения.

В предположении модели М1 при $\alpha = 0,05$ проверить существенность влияния фактора А. В предположении модели М2 при $\alpha = 0,05$ проверить влияние фактора В.

Задача 2.2.6. В двухфакторном комплексе суммы квадратов отклонений, обусловленные действием факторов А и В и их взаимодействием (АВ), равны следующим величинам: $D^2_A = 29$; $D^2_B = 15$; $D^2_{AB} = 12$; $D^2_Y = 59$. Фактор А имеет $r = 3$ группы, фактор

В - $m = 5$ групп, в каждой клетке комплекса имеется по 2 наблюдения.

Учитывая, что уровни фактора А случайные, а фактора В фиксированные, при $\alpha = 0,05$ проверить гипотезу о влиянии факторов А и В на результативный признак.

Задача 2.2.7. В двухфакторном комплексе суммы квадратов отклонений, обусловленные действием факторов А и В и их взаимодействием (АВ), равны следующим величинам: $D^2_A = 29$; $D^2_B = 31$; $D^2_{AB} = 5$; $D^2_Y = 69$. Фактор А имеет $r = 4$ группы, фактор В - $m = 3$ группы, в каждой клетке комплекса имеется по 3 наблюдения.

Учитывая, что уровни факторов А и В случайные, при $\alpha = 0,01$ проверить гипотезу о влиянии этих факторов на результативный признак.

Задача 2.2.8. При исследовании влияния времени реализации продукции (фактор А) и района реализации (фактор В) на объем выручки (тыс. руб.) найдены следующие суммы квадратов: $D^2_A = 19$; $D^2_B = 16$; $D^2_{AB} = 11$; $D^2_Y = 50$. Фактор А имеет $r = 4$ группы, фактор В - $m = 5$ групп, в каждой клетке комплекса имеется по 2 наблюдения.

В предположении модели М1 при $\alpha = 0,05$ проверить влияние времени реализации продукции на объем выручки. В предположении модели М2 проверить влияние фактора В на объем выручки при $\alpha = 0,01$.

Задача 2.2.9. При исследовании зависимости себестоимости товарной продукции на предприятии от объема товарной продукции (фактор А) и производительности труда (фактор В) найдены следующие суммы квадратов: $D^2_A = 42$; $D^2_B = 36$; $D^2_{AB} = 3$; $D^2_Y = 85$. Фактор А имеет $r = 6$ групп, фактор В - $m = 2$ группы, в каждой клетке комплекса имеется по 3 наблюдения.

В предположении, что фактор А имеет фиксированные, а фактор В случайные уровни, при $\alpha = 0,01$ проверить влияние объема товарной продукции на ее себестоимость.

Задача 2.2.10. При исследовании зависимости себестоимо-

сти от объема товарной продукции (фактор А) и фондовооруженности (фактор В) найдены следующие суммы квадратов: $D^2_A = 13$; $D^2_B = 15$; $D^2_{AB} = 6$; $D^2_Y = 39$. Фактор А имеет $r = 5$ групп, фактор В - $m = 3$ группы, в каждой клетке комплекса имеется по 3 наблюдения.

В предположении модели М2 при $\alpha = 0,05$ проверить влияние фактора А.

3. Задачи по компонентному анализу

Задача 3.1. Хозяйственная деятельность $n = 10$ предприятий приборостроения характеризуется $k = 2$ показателями: x_1 - общие затраты на рубль товарной продукции; x_2 - фондоотдача.

№	x_1	x_2
1	0,92	0,51
2	0,72	0,59
3	0,83	1,03
4	0,81	1,21
5	0,82	0,63
6	0,93	0,68
7	0,84	0,57
8	0,89	1,52
9	0,89	1,04
10	0,95	0,99

Требуется:

а) провести ранжирование предприятий по первой главной компоненте;

б) графически изобразить предприятия в пространстве двух первых главных компонент.

в) дать интерпретацию полученным результатам.

Задача 3.2. Деятельность $n=10$ предприятий приборостроения характеризуется $k=2$ показателями: x_1 - общие затраты на единицу товарной продукции; x_2 - среднегодовая заработная плата на единицу промышленно-производственного персонала.

№	x_1	x_2	y
1	0,92	2,33	3,5
2	0,72	2,31	4,4
3	0,83	2,39	3,9
4	0,81	2,34	3,8
5	0,82	2,38	3,6
6	0,93	2,43	3,5
7	0,84	2,20	3,0
8	0,79	2,33	4,1
9	0,88	2,23	4,0
10	0,95	2,13	3,1

Приняв за результативный показатель у производительность труда, построить уравнение регрессии на главной компоненте, наиболее тесно связанной с у. Дать интерпретацию полученным результатам.

Задача 3.3. Значения экономических показателей, отражающих работу $n=10$ предприятий приборостроения, приводятся в таблице: x_1 - фондоотдача; x_2 - среднегодовая заработная плата на единицу промышленно-производственного персонала.

№	x_1	x_2
1	0,88	2,24
2	0,73	2,29
3	0,73	2,40
4	1,47	2,46
5	0,75	2,58
6	0,81	2,18
7	0,83	2,12
8	0,92	2,10
9	0,46	2,29
10	1,11	2,36

Требуется:

- определить собственные значения главных компонент и их вклад в суммарную дисперсию исходных показателей;
- определить ортогональную матрицу собственных векторов корреляционной матрицы R;
- дать экономическую интерпретацию главных компонент.

Задача 3.4. Хозяйственная деятельность $n=10$ машиностроительных предприятий характеризуется $k=3$ показателями: x_1 - трудоемкость единицы продукции; x_2 - удельный вес рабочих в составе промышленно-производственного персонала; x_3 - коэффициент сменности оборудования.

№	x_1	x_2	x_3
1	0,38	0,72	1,40
2	0,24	0,70	1,20

№	x_1	x_2	x_3
3	0,31	0,66	1,15
4	0,42	0,69	1,09
5	0,51	0,71	1,26
6	0,31	0,73	1,36
7	0,37	0,65	1,15
8	0,16	0,82	1,87
9	0,18	0,80	2,17
10	0,43	0,83	1,61

Требуется:

- провести ранжирование предприятий по первой главной компоненте;
- графически изобразить предприятия в пространстве двух первых главных компонент;
- дать интерпретацию полученным результатам.

Задача 3.5. Значения показателей, характеризующих экономическую деятельность $n=10$ предприятий машиностроения, приводятся в таблице: x_1 - удельный вес покупных изделий; x_2 - коэффициент сменности оборудования; x_3 - премии и вознаграждения на одного работника.

№	x_1	x_2	x_3	у
1	0,30	1,40	0,67	9,8
2	0,56	1,20	0,98	13,2
3	0,42	1,15	1,16	17,3
4	0,26	1,09	0,54	7,1
5	0,16	1,26	1,23	2,5
6	0,45	1,36	0,78	12,1
7	0,31	1,15	1,16	15,2
8	0,08	1,87	2,44	31,3
9	0,63	2,16	1,06	11,6
10	0,03	1,61	2,13	30,1

Приняв за результативный показатель у - рентабельность, построить уравнение регрессии на главной компоненте, наиболее тесно связанной с у, дать интерпретацию полученным результатам.

Задача 3.6. Значения экономических показателей $n = 10$ предприятий машиностроения: x_1 - удельный вес рабочих в составе промышленно-производственного персонала; x_2 - удельный вес покупных изделий; x_3 - премии и вознаграждения на одного работника даны в таблице.

№	x_1	x_2	x_3
1	0,68	0,33	0,82
2	0,74	0,25	0,84
3	0,66	0,32	0,67
4	0,72	0,02	1,04
5	0,68	0,06	0,66
6	0,77	0,15	0,86
7	0,78	0,08	0,79
8	0,78	0,20	0,34
9	0,81	0,20	1,60
10	0,79	0,30	1,46

Требуется:

- определить собственные значения главных компонент и их вклад в суммарную дисперсию исходных показателей;
- определить ортогональную матрицу собственных векторов;
- дать экономическую интерпретацию главных компонент.

Задача 3.7. Деятельность $n = 10$ предприятий приборостроения характеризуется $k=3$ показателями: x_1 - трудоемкость единицы продукции; x_2 - удельный вес покупных изделий; x_3 - коэффициент сменности оборудования.

№	x_1	x_2	x_3	y
1	0,51	0,20	1,47	21,9
2	0,36	0,64	1,27	48,4
3	0,23	0,42	1,51	173,5
4	0,26	0,27	1,46	74,1
5	0,27	0,37	1,27	68,6
6	0,29	0,38	1,43	60,8
7	0,01	0,35	1,50	355,6

№	x_1	x_2	x_3	y
8	0,02	0,42	1,35	264,8
9	0,18	0,32	1,41	526,6
10	0,25	0,33	1,47	118,6

Приняв за результативный показатель y - индекс снижения себестоимости продукции, построить уравнение регрессии на главной компоненте, наиболее тесно связанной с ним. Дать экономическую интерпретацию полученным результатам.

Задача 3.8. Найдены значения главных компонент для i -го предприятия $f_{i1} = 0,661$, $f_{i2} = -2,151$ и матрица факторных нагрузок

$$A = \begin{pmatrix} -0,756 & 0,654 \\ 0,756 & 0,654 \end{pmatrix}$$

Требуется найти значения исходных показателей для этого предприятия, если выборка характеризуется векторами средних $X = (0,850 \ 0,877)'$ и средних квадратических отклонений $S = (0,072 \ 0,333)'$.

Задача 3.9. В матрице значений главных компонент F оказалось пропущенным последнее десятое наблюдение ($i = 10$).

$$F = \begin{pmatrix} -0,484 & 1,053 \\ 1,227 & -1,572 \\ 0,773 & 0,527 \\ 0,606 & -0,175 \\ 0,799 & 0,314 \\ 0,136 & 2,097 \\ -0,671 & -1,109 \\ 0,720 & -0,507 \\ -0,921 & -0,231 \\ *** & *** \end{pmatrix}$$

Матрица факторных нагрузок равна

$$A = \begin{pmatrix} -0.791 & 0.611 \\ 0.791 & 0.611 \end{pmatrix}$$

Требуется определить:

- а) значения главных компонент для десятого наблюдения;
б) значения исходных показателей для 10-го наблюдения,

если выборка характеризуется векторами $X = (0.850 \ 2.307)^T$ и $S = (0.072 \ 0.093)^T$.

Задача 3.10. В матрице значений главных компонент F оказалось пропущенным восьмое наблюдение ($i = 8$).

$$F = \begin{pmatrix} -0.25 & 0.34 & -0.54 \\ 0.22 & -1.30 & 1.31 \\ -0.75 & -1.08 & -0.14 \\ -0.97 & 0.24 & 0.39 \\ -0.89 & 1.39 & 0.87 \\ -0.03 & -0.24 & 0.60 \\ -1.00 & -0.62 & -0.92 \\ 1.86 & -0.33 & -1.91 \\ 0.58 & 2.07 & 1.09 \end{pmatrix}$$

Матрица факторных нагрузок равна

$$A = \begin{pmatrix} -0.72 & 0.69 & -0.08 \\ 0.88 & 0.44 & 0.19 \\ 0.96 & 0.12 & -0.24 \end{pmatrix}$$

Требуется определить:

- а) значения главных компонент для восьмого наблюдения;
б) значения исходных показателей для 8-го наблюдения,

если выборка характеризуется векторами $X = (0.33 \ 0.73 \ 1.43)^T$ и $S = (0.11 \ 0.06 \ 0.34)^T$.

4. Задачи по факторному анализу

Задача 4.1. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом наибольшего коэффициента корреляции по строке и методом триад, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1.00 & 0.43 & 0.37 & 0.40 & 0.58 \\ & 1.00 & 0.85 & 0.98 & 0.11 \\ & & 1.00 & 0.88 & 0.46 \\ & & & 1.00 & 0.28 \\ & & & & 1.00 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.2. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом наибольшего коэффициента корреляции по строке и методом триад, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1 & 0.740 & 0.565 & 0.875 & 0.362 & 0.730 \\ & 1 & 0.478 & 0.690 & 0.802 & 0.558 \\ & & 1 & 0.672 & 0.475 & 0.650 \\ & & & 1 & 0.525 & 0.820 \\ & & & & 1 & 0.539 \\ & & & & & 1 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.3. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом среднего коэффициента корреляции по строке и методом триад, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1,00 & 0,43 & 0,37 & 0,40 & 0,58 \\ & 1,00 & 0,85 & 0,98 & 0,11 \\ & & 1,00 & 0,88 & 0,46 \\ & & & 1,00 & 0,28 \\ & & & & 1,00 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.4. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом наибольшего коэффициента корреляции по строке и методом первого центроидного фактора, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1,00 & 0,43 & 0,37 & 0,40 & 0,58 \\ & 1,00 & 0,85 & 0,98 & 0,11 \\ & & 1,00 & 0,88 & 0,46 \\ & & & 1,00 & 0,28 \\ & & & & 1,00 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.5. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом среднего коэффициента корреляции по строке и методом триад, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1 & 0,740 & 0,565 & 0,875 & 0,362 & 0,730 \\ & 1 & 0,478 & 0,690 & 0,802 & 0,558 \\ & & 1 & 0,672 & 0,475 & 0,650 \\ & & & 1 & 0,525 & 0,820 \\ & & & & 1 & 0,539 \\ & & & & & 1 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.6. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом первого центроидного фактора и методом триад, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1 & 0,740 & 0,565 & 0,875 & 0,362 & 0,730 \\ & 1 & 0,478 & 0,690 & 0,802 & 0,558 \\ & & 1 & 0,672 & 0,475 & 0,650 \\ & & & 1 & 0,525 & 0,820 \\ & & & & 1 & 0,539 \\ & & & & & 1 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.7. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом первого центроидного фактора и методом среднего коэффициента корреляции по строке, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1 & 0,740 & 0,565 & 0,875 & 0,362 & 0,730 \\ & 1 & 0,478 & 0,690 & 0,802 & 0,558 \\ & & 1 & 0,672 & 0,475 & 0,650 \\ & & & 1 & 0,525 & 0,820 \\ & & & & 1 & 0,539 \\ & & & & & 1 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

Задача 4.8. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом квадрата коэффициента множественной корреляции, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1 & 0,740 & 0,565 & 0,875 & 0,362 & 0,730 \\ & 1 & 0,478 & 0,690 & 0,802 & 0,558 \\ & & 1 & 0,672 & 0,475 & 0,650 \\ & & & 1 & 0,525 & 0,820 \\ & & & & 1 & 0,539 \\ & & & & & 1 \end{pmatrix}$$

Задача 4.9. От заданной матрицы парных коэффициентов корреляции R , пользуясь методом квадрата коэффициента множественной корреляции, перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1,00 & 0,43 & 0,37 & 0,40 & 0,58 \\ & 1,00 & 0,85 & 0,98 & 0,11 \\ & & 1,00 & 0,88 & 0,46 \\ & & & 1,00 & 0,28 \\ & & & & 1,00 \end{pmatrix}$$

Задача 4.10. От заданной матрицы парных коэффициентов корреляции R , пользуясь методами: квадрата коэффициента множественной корреляции, триад, наибольшего коэффициента корреляции по строке (столбцу), перейти к редуцированной матрице R_h .

$$R = \begin{pmatrix} 1,00 & 0,43 & 0,37 & 0,40 & 0,58 \\ & 1,00 & 0,85 & 0,98 & 0,11 \\ & & 1,00 & 0,88 & 0,46 \\ & & & 1,00 & 0,28 \\ & & & & 1,00 \end{pmatrix}$$

Проанализировать графическое изображение полученных оценок общностей.

ТЕСТ I

1. Известно, что в однофакторном дисперсионном комплексе факторная дисперсия и общая дисперсия равны соответственно 45,74 и 67,5. Можно ли считать статистически обоснованным влияние исследуемого фактора, если рассматриваемая выборка содержит 16 наблюдений в четырех группах. Определить разность между наблюдаемым и табличным значением F -критерия при $\alpha = 0,01$:

- а) -3,4101; б) 2,4596; в) 28; г) 0.

2. Для расчета суммы квадратов отклонений, обусловленных случайными факторами в однофакторном комплексе, используется формула:

а) $\sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$; б) $\sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$;
в) $\sum_{j=1}^r (\bar{y}_j - \bar{y})^2 n_j$; г) $\sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$.

3. Чему равна оценка дисперсии, обусловленной фактором В в двухфакторном дисперсионном комплексе, если известны следующие суммы квадратов отклонений: $D_A^2 = 10,0$; $D_{AB}^2 = 1,0$; $D_x^2 = 0$; $D_y^2 = 67,5$. Дисперсионный анализ проводится по наблюдениям, в которых по фактору А выделены четыре группы, а по фактору В - три группы:

- а) -56,5; б) 0,3; в) 0,001; г) 28,25.

4. Для проверки статистической значимости выводов однофакторного дисперсионного анализа используется статистика критерия:

$$\begin{aligned} \text{а) } F &= S_x^2 \cdot S_z^2; & \text{б) } F &= \frac{\frac{1}{k} S_x^2}{\frac{1}{n-k} S_z^2}; \\ \text{в) } F &= \frac{S_x^2}{S_z^2}; & \text{г) } F &= S_z^2 - S_x^2. \end{aligned}$$

5. Известно, что в двухфакторном дисперсионном комплексе дисперсия по фактору А равна 64,5, а остаточная дисперсия равна 1,0. Проверить статистическую значимость влияния фактора А при $\alpha = 0,05$, если исследования проводятся по выборке объема $n = 16$ и по фактору А выделено 4 группы, а по фактору В - две. Определить разность между наблюдаемым и табличным значением F критерия:

$$\text{а) } 167,93; \quad \text{б) } -0,128; \quad \text{в) } 11,5; \quad \text{г) } 43,213.$$

6. Для расчета сумм квадратов отклонений, обусловленных совместным влиянием факторов А и В в двухфакторном комплексе, используется формула:

$$\begin{aligned} \text{а) } \sum_{k=1}^m \sum_{j=1}^r \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{j\cdot} - \bar{y}_{k\cdot} + \bar{y})^2 n_{jk}; \\ \text{б) } \sum_{j=1}^r \sum_{i=1}^{n_j} (\bar{y}_{j\cdot} - \bar{y})^2 n_{jk}; \end{aligned}$$

$$\begin{aligned} \text{в) } \sum_{k=1}^m \sum_{i=1}^r (\bar{y}_{jk} - \bar{y}_{j\cdot} - \bar{y}_{k\cdot} + \bar{y})^2 n_{jk}; \\ \text{г) } \sum_{k=1}^m \sum_{j=1}^r \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{jk})^2. \end{aligned}$$

7. Для проверки основной гипотезы о влиянии фактора А на результативный признак у в двухфакторном дисперсионном комплексе в условиях модели II используется статистика:

$$\text{а) } F = \frac{S_A^2}{S_{Ae}}; \quad \text{б) } F = \frac{S_A^2}{S_y}; \quad \text{в) } F = \frac{S_A^2}{S_e}; \quad \text{г) } F = \frac{S_A^2}{S_z}.$$

8. Чему равна оценка факторной дисперсии, если в условиях однофакторного дисперсионного комплекса суммы квадратов отклонений равны следующим величинам: $D_y^2 = 116$; $D_z^2 = 46$. Исследование ведется по выборке объема $n = 15$ наблюдений, в которых выделены 3 группы:

$$\text{а) } 3,83; \quad \text{б) } 35; \quad \text{в) } 14,1; \quad \text{г) } 6,93.$$

9. Для проверки основной гипотезы о влиянии фактора В на результативный признак у в двухфакторном дисперсионном комплексе в условиях модели III (В - случайный фактор) используется статистика:

$$\text{а) } F = \frac{S_e^2}{S_{Ae}}; \quad \text{б) } F = \frac{S_e^2}{S_f}; \quad \text{в) } F = \frac{S_A^2}{S_e}; \quad \text{г) } F = \frac{S_e^2}{S_z}.$$

10. В двухфакторном комплексе суммы квадратов отклонений, обусловленных действием факторов А и В, равны следующим величинам: $D_A^2 = 29$; $D_B^2 = 15$; $D_{AB}^2 = 12$; $D_y^2 = 59$. Фактор А имеет $r = 3$ уровня, фактор В - $m = 5$ уровней и в каждой клетке комплекса $n = 2$ наблюдения. Учитывая, что уровни фактора А случайные, а фактора В фиксированные, при $\alpha = 0,01$ проверить гипотезу о влиянии фактора А. Определить разность между наблюдаемым и табличным значениями F-критерия.

- а) $F_n - F_{кр} = 63,0$; б) $F_n - F_{кр} = -0,9$;
в) $F_n - F_{кр} = 0,12$; г) $F_n - F_{кр} = 66,14$.

ТЕСТ II

1. В кластер S_1 входят четыре объекта, расстояние от которых до объекта № 5 составляет соответственно: 2, 5, 6, 7. Чему равно расстояние от объекта № 5 до кластера S_1 , если исходить из принципа «ближайшего соседа»:

- а) 2; б) 5; в) 6; г) 7.

2. Определить по данным п. 1 настоящего теста, расстояние от объекта № 5 до кластера S_1 , если исходить из принципа «дальнего соседа»:

- а) 2; б) 5; в) 6; г) 7.

3. Чему равно по данным п. 1 настоящего теста, расстояние от объекта № 5 до кластера S_1 , исходя из принципа «средней связи»:

- а) 2; б) 5; в) 6; г) 7.

4. Расстояние между пятью объектами ($n = 5$) характеризуется матрицей расстояний

$$R = \begin{pmatrix} 0 & 2,2 & 3,0 & 5,1 & 5,8 \\ 2,2 & 0 & 1,4 & 5,0 & 6,4 \\ 3,0 & 1,4 & 0 & 6,4 & 7,8 \\ 5,1 & 5,0 & 6,4 & 0 & 2,0 \\ 5,8 & 6,4 & 7,8 & 2,0 & 0 \end{pmatrix}$$

Чему равно расстояние между кластерами $S_{(1,2)}$ и $S_{(3,4,5)}$, в которые входят соответственно объекты (1, 2) и (3, 4, 5), если исходить из принципа «ближайшего соседа»:

- а) 2,2; б) 3,0; в) 1,4; г) 2,0.

5. По данным п. 4 определите, чему равно расстояние между кластерами $S_{(12)}$ и $S_{(345)}$, если исходить из принципа «дальнего соседа»:

- а) 5,8; б) 5,1; в) 4,89; г) 6,4.

6. Чему равно, по данным п. 4, расстояние между кластерами $S_{(12)}$ и $S_{(345)}$, если исходить из принципа «средней связи»:

- а) 4,45; б) 3,37; в) 4,89; г) 2,86.

7. Данные о четырех фирмах, деятельность которых характеризуется показателями x_1 и x_2 , представлены в таблице:

№	1	2	3	4
x_1	1	7	1	9
x_2	5	9	3	7

Чему равно расстояние между 1-м и 2-м объектами, если в качестве метрики принять обычное евклидово расстояние:

- а) 3,78; б) 9,34; в) 7,21; г) 5,19.

8. Чему равно, по данным п. 7, расстояние между 3-м и 4-м объектами, если в качестве метрики принять взвешенное евклидово расстояние с весами $\omega_1 = 0,7$ и $\omega_2 = 0,3$:

- а) 5,35; б) 7,04; в) 6,66; г) 9,38.

9. Чему равно, по данным п. 7, расстояние между 2-м объектом и кластером $S_{(13)}$, включающим 1-й и 3-й объекты, если расстояние измерять с помощью обычной евклидовой метрики и принципа «центра тяжести»:

- а) 6,37; б) 7,18; в) 8,93; г) 5,72.

10. Чему равно, по данным п. 7, расстояние между кластерами $S_{(13)}$ и $S_{(24)}$, если исходить из принципа «центра тяжести» «взвешенной» евклидовой метрики с весами $\omega_1 = 0,2$ и $\omega_2 = 0,8$:

- а) 7,45; б) 9,73; в) 5,39; г) 4,75.

ТЕСТ III

1. Финансовая устойчивость предприятия характеризуется $k = 8$ показателями. В результате расчетов получены собственные значения трех первых главных компонент: $\lambda_1 = 4.0$; $\lambda_2 = 1.6$ и $\lambda_3 = 0.8$. Чему равен относительный вклад двух первых главных компонент, %?

- а) 30; б) 70; в) 60; г) 80.

2. Дана матрица факторных нагрузок:

$$A = \begin{pmatrix} -0.72 & 0.69 & -0.08 \\ 0.88 & 0.44 & 0.19 \\ 0.96 & 0.12 & -0.24 \end{pmatrix}$$

Чему равен относительный вклад второй главной компоненты f_2 в суммарную дисперсию, %?

- а) 74; б) 37; в) 4; г) 23.

3. В матрице факторных нагрузок A оказались пропущенными элементы a_{13} , a_{22} , a_{31} :

$$A = \begin{pmatrix} -0.90 & -0.30 & a_{13} \\ -0.54 & a_{22} & 0.03 \\ a_{31} & 0.20 & 0.33 \end{pmatrix}$$

Чему равно собственное значение λ_3 , соответствующее третьей главной компоненте f_3 ?

- а) 1.95; б) 0.63; в) 0.21; г) 0.84.

4. В каких пределах меняются элементы матрицы факторных нагрузок A :

- а) $(-1; 0)$; б) $(0; 1)$; в) $(-1; 1)$; г) $(0; 2)$.

5. Дана матрица факторных нагрузок:

$$A = \begin{pmatrix} -0.72 & 0.69 & -0.08 \\ 0.88 & 0.44 & 0.19 \\ 0.96 & 0.12 & -0.24 \end{pmatrix}$$

Чему равен парный коэффициент корреляции между переменной x_3 и второй главной компонентой f_2 ?

- а) 0.12; б) 0.96; в) -0.24; г) 0.19.

6. Дана матрица факторных нагрузок:

$$A = \begin{pmatrix} -0.79 & 0.61 \\ 0.79 & 0.61 \end{pmatrix}$$

Чему равен коэффициент корреляции между переменными x_1 и x_2 ?

- а) 0.75; б) 1.25; в) 0.25; г) -0.25.

7. По данным наблюдений ($n = 5$) получена матрица значений главных компонент F , в которой пропущено одно значение:

$$F = \begin{pmatrix} -0.84 & 0.60 \\ 0.08 & -1.26 \\ 1.52 & 1.26 \\ -1.30 & 0.50 \\ f_{51} & -1.10 \end{pmatrix}$$

Чему равно пропущенное значение f_{51} ?

- а) -0,58; б) 0,65; в) 0,54; г) 1,08.

8. По матрице F из п. 7 определите, чему равен парный коэффициент корреляции между 1-й и 2-й главными компонентами (f_1 и f_2)?

- а) 0; б) 1; в) -1; г) 0,6.

9. При исследовании взаимосвязи двух показателей x_1 и x_2 получен коэффициент корреляции $\rho_{12} = 0,9$. Чему равно собственное значение λ_1 , соответствующее первой главной компоненте?

- а) 0,1; б) 1,9; в) 1,8; г) 0,2.

10. Деятельность n предприятий региона характеризуется четырьмя показателями. При проведении компонентного анализа получены собственные значения, одно из которых оказалось пропущенным: 1,2; 0,4; 0,8. Чему равно собственное значение λ_3 , соответствующее третьей главной компоненте?

- а) 2,5; б) 1,2; в) 0,4; г) 0,8.

ТЕСТ IV

1. Известно, что в однофакторном дисперсионном комплексе факторная дисперсия и общая дисперсия равны соответственно 45,74 и 67,5. Можно ли считать статистически обоснованным влияние исследуемого фактора, если рассматриваемая выборка содержит 16 наблюдений в четырех группах. Определить разность между наблюдаемым и табличным значением F -критерия при $\alpha = 0,01$:

- а) -3,4101; б) 2,4596; в) 28; г) 0.

2. Для расчета суммы квадратов отклонений, обусловленных случайными факторами в однофакторном комплексе, используется формула:

$$\begin{aligned} \text{а) } \sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 n_j; & \quad \text{б) } \sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2; \\ \text{в) } \sum_{j=1}^r (\bar{y}_j - \bar{y})^2 n_j; & \quad \text{г) } \sum_{j=1}^r \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2. \end{aligned}$$

3. Чему равна оценка дисперсии, обусловленной фактором В в двухфакторном дисперсионном комплексе, если известны следующие суммы квадратов отклонений: $D_A^2 = 10,0$; $D_{AB}^2 = 1,0$; $D_\varepsilon^2 = 0$; $D_B^2 = 67,5$. Дисперсионный анализ проводится по наблюдениям, в которых по фактору А выделены четыре группы, а по фактору В - три группы:

- а) -56,5; б) 0,3; в) 0,001; г) 28,25.

4. Для проверки статистической значимости выводов однофакторного дисперсионного анализа используется статистика критерия:

$$\text{а) } F = S_x^2 \cdot S_z^2; \quad \text{б) } F = \frac{\frac{1}{k} S_x^2}{\frac{1}{n-k} S_z^2};$$

$$\text{б) } F = \frac{S_x^2}{S_z^2}; \quad \text{г) } F = S_z^2 - S_x^2.$$

5. Известно, что в двухфакторном дисперсионном комплексе дисперсия по фактору А равна 64,5, а остаточная дисперсия равна 1,0. Проверить статистическую значимость влияния фактора А при $\alpha = 0,05$, если исследования проводятся по выборке объема $n = 16$ и по фактору А выделено 4 группы, а по фактору В - две. Определить разность между наблюдаемым и табличным значением F критерия:

$$\text{а) } 167,93; \quad \text{б) } -0,128; \quad \text{в) } 11,5; \quad \text{г) } 43,213.$$

6. Для расчета сумм квадратов отклонений, обусловленных совместным влиянием факторов А и В в двухфакторном комплексе, используется формула:

$$\text{а) } \sum_{k=1}^m \sum_{j=1}^r \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_j - \bar{y}_k + \bar{y})^2 n_{jk};$$

$$\text{б) } \sum_{j=1}^r \sum_{i=1}^{n_j} (\bar{y}_j + \bar{y}_k - \bar{y})^2 n_{jk};$$

$$\text{в) } \sum_{k=1}^m \sum_{j=1}^r (\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2 n_{jk};$$

$$\text{г) } \sum_{k=1}^m \sum_{j=1}^r \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{jk})^2.$$

7. Для проверки основной гипотезы о влиянии фактора А на результативный признак у в двухфакторном дисперсионном комплексе в условиях модели II используется статистика:

$$\text{а) } F = \frac{S_A^2}{S_{Ae}}; \quad \text{б) } F = \frac{S_A^2}{S_y}; \quad \text{в) } F = \frac{S_A^2}{S_e}; \quad \text{г) } F = \frac{S_A^2}{S_z}.$$

8. Чему равна оценка факторной дисперсии, если в условиях однофакторного дисперсионного комплекса суммы квадратов отклонений равны следующим величинам: $D_y^2 = 116$; $D_z^2 = 46$. Исследование ведется по выборке объема $n = 15$ наблюдений, в которых выделены 3 группы:

$$\text{а) } 3,83; \quad \text{б) } 35; \quad \text{в) } 14,1; \quad \text{г) } 6,93.$$

9. Для проверки основной гипотезы о влиянии фактора В на результативный признак у в двухфакторном дисперсионном комплексе в условиях модели III (В - случайный фактор) используется статистика:

$$\text{а) } F = \frac{S_e^2}{S_{Ae}}; \quad \text{б) } F = \frac{S_e^2}{S_y};$$

$$в) F = \frac{S_A^2}{S_e^2};$$

$$г) F = \frac{S_e^2}{S_z^2}.$$

10. В двухфакторном комплексе суммы квадратов отклонений, обусловленных действием факторов А и В, равны следующим величинам: $D_A^2 = 29$; $D_B^2 = 15$; $D_{AB}^2 = 12$; $D_y^2 = 59$. Фактор А имеет $r = 3$ уровня, фактор В - $m = 5$ уровней и в каждой клетке комплекса $n = 2$ наблюдения. Учítывая, что уровни фактора А случайные, а фактора В фиксированные, при $\alpha = 0,01$ проверить гипотезу о влиянии фактора А. Определить разность между наблюдаемым и табличным значениями F-критерия.

$$а) F_{н} - F_{кр} = 63,0; \quad б) F_{н} - F_{кр} = -0,9;$$

$$в) F_{н} - F_{кр} = 0,12; \quad г) F_{н} - F_{кр} = 66,14.$$

ТЕСТ V

1. В кластер S_1 входят четыре объекта, расстояние от которых до объекта № 5 составляет соответственно: 2, 5, 6, 7. Чему равно расстояние от объекта № 5 до кластера S_1 , если исходить из принципа «ближайшего соседа»:

$$а) 2; \quad б) 5; \quad в) 6; \quad г) 7.$$

2. Определить по данным п. 1 настоящего теста, расстояние от объекта № 5 до кластера S_1 , если исходить из принципа «дальнего соседа»:

$$а) 2; \quad б) 5; \quad в) 6; \quad г) 7.$$

3. Чему равно по данным п. 1 настоящего теста, расстояние от объекта № 5 до кластера S_1 , исходя из принципа «средней связи»:

$$а) 2; \quad б) 5; \quad в) 6; \quad г) 7.$$

4. Расстояние между пятью объектами ($n = 5$) характеризуется матрицей расстояний

$$R = \begin{pmatrix} 0 & 2,2 & 3,0 & 5,1 & 5,8 \\ 2,2 & 0 & 1,4 & 5,0 & 6,4 \\ 3,0 & 1,4 & 0 & 6,4 & 7,8 \\ 5,1 & 5,0 & 6,4 & 0 & 2,0 \\ 5,8 & 6,4 & 7,8 & 2,0 & 0 \end{pmatrix}.$$

Чему равно расстояние между кластерами $S_{(12)}$ и $S_{(345)}$, в которые входят соответственно объекты (1, 2) и (3, 4, 5), если исходить из принципа «ближайшего соседа»:

$$а) 2,2; \quad б) 3,0; \quad в) 1,4; \quad г) 2,0.$$

5. По данным п. 4 определите, чему равно расстояние между кластерами $S_{(12)}$ и $S_{(345)}$, если исходить из принципа «дальнего соседа»:

- а) 5,8; б) 5,1; в) 4,89; г) 6,4.

6. Чему равно, по данным п. 4, расстояние между кластерами $S_{(12)}$ и $S_{(345)}$, если исходить из принципа «средней связи»:

- а) 4,45; б) 3,37; в) 4,89; г) 2,86.

7. Данные о четырех фирмах, деятельность которых характеризуется показателями x_1 и x_2 , представлены в таблице:

№	1	2	3	4
x_1	1	7	1	9
x_2	5	9	3	7

Чему равно расстояние между 1-м и 2-м объектами, если в качестве метрики принять обычное евклидово расстояние:

- а) 3,78; б) 9,34; в) 7,21; г) 5,19.

8. Чему равно, по данным п. 7, расстояние между 3-м и 4-м объектами, если в качестве метрики принять взвешенное евклидово расстояние с весами $\omega_1 = 0,7$ и $\omega_2 = 0,3$:

- а) 5,35; б) 7,04; в) 6,66; г) 9,38.

9. Чему равно, по данным п. 7, расстояние между 2-м объектом и кластером $S_{(13)}$, включающим 1-й и 3-й объекты, если расстояние измерять с помощью обычной евклидовой метрики и принципа «центра тяжести»:

- а) 6,37; б) 7,18; в) 8,93; г) 5,72.

10. Чему равно, по данным п. 7, расстояние между кластерами $S_{(13)}$ и $S_{(24)}$, если исходить из принципа «центра тяжести» «взвешенной» евклидовой метрики с весами $\omega_1 = 0,2$ и $\omega_2 = 0,8$:

- а) 7,45; б) 9,73; в) 5,39; г) 4,75.

ТЕСТ VI

1. Финансовая устойчивость предприятия характеризуется $k = 8$ показателями. В результате расчетов получены собственные значения трех первых главных компонент: $\lambda_1 = 4,0$; $\lambda_2 = 1,6$ и $\lambda_3 = 0,8$. Чему равен относительный вклад двух первых главных компонент, %?

- а) 30; б) 70; в) 60; г) 80.

2. Дана матрица факторных нагрузок:

$$A = \begin{pmatrix} -0,72 & 0,69 & -0,08 \\ 0,88 & 0,44 & 0,19 \\ 0,96 & 0,12 & -0,24 \end{pmatrix}$$

Чему равен относительный вклад второй главной компоненты f_2 в суммарную дисперсию, %?

- а) 74; б) 37; в) 4; г) 23.

3. В матрице факторных нагрузок A оказались пропущенными элементы a_{13} , a_{22} , a_{31} :

$$A = \begin{pmatrix} -0,90 & -0,30 & a_{13} \\ -0,54 & a_{22} & 0,03 \\ a_{31} & 0,20 & 0,33 \end{pmatrix}$$

Чему равно собственное значение λ_3 , соответствующее третьей главной компоненте f_3 ?

- а) 1,95; б) 0,63; в) 0,21; г) 0,84.

4. В каких пределах меняются элементы матрицы факторных нагрузок A :

- а) $(-1; 0)$; б) $(0; 1)$; в) $(-1; 1)$; г) $(0; 2)$.

5. Дана матрица факторных нагрузок:

$$A = \begin{pmatrix} -0,72 & 0,69 & -0,08 \\ 0,88 & 0,44 & 0,19 \\ 0,96 & 0,12 & -0,24 \end{pmatrix}$$

Чему равен парный коэффициент корреляции между переменной x_3 и второй главной компонентой f_2 ?

- а) 0,12; б) 0,96; в) -0,24; г) 0,19.

6. Дана матрица факторных нагрузок:

$$A = \begin{pmatrix} -0,79 & 0,61 \\ 0,79 & 0,61 \end{pmatrix}$$

Чему равен коэффициент корреляции между переменными x_1 и x_2 ?

- а) 0,75; б) 1,25; в) 0,25; г) -0,25.

7. По данным наблюдений ($n = 5$) получена матрица значений главных компонент F , в которой пропущено одно значение:

$$F = \begin{pmatrix} -0,84 & 0,60 \\ 0,08 & -1,26 \\ 1,52 & 1,26 \\ -1,30 & 0,50 \\ f_{51} & -1,10 \end{pmatrix}$$

Чему равно пропущенное значение f_{51} ?

- а) -0,58; б) 0,65; в) 0,54; г) 1,08.

ЛАБОРАТОРНАЯ РАБОТА

8. По матрице F из п. 7 определите, чему равен парный коэффициент корреляции между 1-й и 2-й главными компонентами (f_1 и f_2)?

- а) 0; б) 1; в) -1; г) 0,6.

9. При исследовании взаимосвязи двух показателей x_1 и x_2 получен коэффициент корреляции $\rho_{12} = 0,9$. Чему равно собственное значение λ_1 , соответствующее первой главной компоненте?

- а) 0,1; б) 1,9; в) 1,8; г) 0,2.

10. Деятельность n предприятий региона характеризуется четырьмя показателями. При проведении компонентного анализа получены собственные значения, одно из которых оказалось пропущенным: 1,2; 0,4; 0,8. Чему равно собственное значение λ_3 , соответствующее третьей главной компоненте?

- а) 2,5; б) 1,2; в) 0,4; г) 0,8.

Исследовать на основе корреляционного, регрессионного, компонентного и факторного анализов зависимость результативного признака от набора показателей по данным таблицы 4. Провести кластеризацию объектов. Дать экономико-статистическую интерпретацию полученных результатов.

Обозначения и наименования показателей:

$y(x_1)$ - среднедушевой денежный доход населения (в месяц, рублей);

x_2 - доля оплаты труда в денежном доходе населения (в процентах от общего объема денежных доходов);

x_3 - доля социальных трансфертов в денежном доходе населения (в процентах от общего объема денежных доходов);

x_4 - размер прожиточного минимума (в среднем на душу населения в месяц, рублей);

x_5 - уровень общей безработицы (в процентах от экономически активного населения);

x_6 - индекс потребительских цен (декабрь к декабрю предыдущего года, в процентах);

x_7 - доля приватизированного жилья (в процентах от общего числа квартир, подлежащих приватизации);

x_8 - уровень развития инфраструктуры (выпуск специалистов высшими учебными заведениями на 10000 населения);

x_9 - коэффициент миграционного прироста (на 10000 человек);

x_{10} - доля пенсионеров в общей численности населения;

x_{11} - удельный вес сельского населения в общей численности населения (в процентах).

Варианты лабораторной работы

№ варианта	Номера факторных признаков
1	2, 3, 4, 5, 6
2	3, 4, 5, 6, 7
3	4, 5, 6, 7, 8
4	5, 6, 7, 8, 9
5	6, 7, 8, 9, 10
6	7, 8, 9, 10, 11
7	8, 9, 10, 11, 2
8	9, 10, 11, 2, 3
9	10, 11, 2, 3, 4
10	11, 2, 3, 4, 5
11	2, 3, 6, 7, 8
12	2, 3, 7, 8, 9
13	3, 4, 8, 9, 10
14	3, 4, 9, 10, 11
15	4, 5, 7, 8, 9
16	4, 5, 8, 9, 10
17	5, 6, 9, 10, 11
18	2, 5, 6, 10, 11
19	3, 5, 7, 9, 11
20	2, 4, 6, 8, 10

Таблица 4

Факторы, характеризующие межрегиональную дифференциацию уровня жизни населения

Регионы России	$y(x_i)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	2		3	4	5	6	7	8	9	10	11	12
Республика Карелия	1052	39,6	16,3	601	16,6	180,1	34	25,3	-3	27,6	26,2	
Республика Коми	1156	44,3	13,6	529	17,8	170,2	37	17,9	-93	23,4	25,7	
Архангельская область	711	45,4	16,0	578	14,9	171,8	29	19,8	-53	26,9	25,9	
Вологодская область	800	55,6	17,6	475	12,7	163,5	37	25,0	-19	27,5	31,7	
Мурманская область	1533	35,1	9,9	693	21,0	179,3	48	11,9	-157	22,9	8,2	
г. Санкт-Петербург	1060	41,8	14,5	614	11,3	178,0	36	81,2	28	26,7	0,0	
Ленинградская область	649	47,9	18,7	508	15,0	166,2	23	0,2	83	27,6	34,1	
Новгородская область	896	32,3	16,5	438	15,4	172,2	40	25,6	56	29,9	28,8	
Псковская область	544	45,7	24,6	405	16,1	168,7	43	16,4	18	30,5	34,1	
Брянская область	554	34,6	26,9	355	15,7	175,4	45	19,4	19	30,8	31,4	
Владимирская область	589	44,2	20,5	391	12,0	172,3	47	17,0	38	29,0	19,6	

232	432	9	592	15	0121	651	254	491	963	939	Архангельская область
262	352	83	338	45	3921	601	443	291	484	842	Республика Татарстан
309	312	15-	162	35	5891	308	263	042	444	134	Республика Калмыкия
814	403	21	612	24	6691	221	333	921	203	965	Республика Карелия
093	562	84	012	94	3191	111	843	081	234	052	Липецкая область
163	603	21	523	84	5121	201	093	512	634	439	Курская область
623	013	24	463	44	6321	56	393	061	963	239	Воронежская область
543	562	411	482	44	9291	311	653	202	424	612	Белгородская область
812	882	04	282	14	2661	126	993	261	984	559	Нижегородская область
262	522	0	291	53	6291	131	584	281	834	865	Кировская область
163	152	92	823	62	6291	631	283	222	364	824	Чувашская республика
904	482	6-	024	82	2421	541	963	212	314	305	Республика Мордовия
21	11	01	6	8	2	9	3	4	3	2	1

Продолжение таблицы 2

1	2	3	4	5	6	7	8	9	10	11	12
Ивановская область	530	39.2	22.0	369	18.8	176.3	43	34.9	26	29.6	17.6
Калужская область	640	49.0	16.1	404	10.2	173.7	46	16.7	45	25.8	25.6
Костромская область	586	43.8	17.8	385	11.2	172.4	44	26.6	29	28.9	34.5
г. Москва	4014	16.4	5.6	679	4.8	209.6	45	107.9	69	27.1	0.0
Московская область	703	43.2	20.4	448	9.9	182.5	33	22.0	57	27.7	20.3
Орловская область	693	34.9	22.4	385	13.2	167.9	41	37.7	40	30.1	36.8
Рязанская область	568	41.7	22.0	390	7.1	179.4	48	21.9	21	32.1	31.6
Смоленская область	712	33.8	17.5	385	16.4	176.7	53	24.3	13	28.8	29.8
Тверская область	537	43.6	23.2	403	11.3	171.2	49	22.6	34	29.9	26.8
Тульская область	721	38.2	21.8	383	11.6	169.4	50	17.7	18	34.3	18.6
Ярославская область	741	45.6	18.1	428	11.1	172.6	43	32.7	30	29.6	19.6
Республика Марий-Эл	421	48.2	23.2	431	13.1	179.6	40	34.1	16	25.0	38.1

Продолжение таблицы 1

231	692	9	28,3	53	470,1	52,1	287	29,1	477	686	Кемеровская область
474	26,2	10	7,82	19	171,0	09,1	237	26,1	537	505	Алтайский край
952	93,6	47	6,17	52	182,1	58,1	537	19,1	477	595	Республика Алтай
781	592	47	27,2	67	189,1	72,1	217	29,1	677	827	Читинская область
521	892	22	373	47	576,1	50,1	025	95,1	705	767	Свердловская область
732	752	11	852	43	783,1	03,1	227	73,1	707	1001	Пермская область
653	852	62	802	83	587,1	73,1	873	79,1	287	379	Оренбургская область
777	592	52	512	53	837,1	13,1	173	172	297	535	Курганская область
503	832	81	613	53	579,1	13,1	057	29,1	267	179	Удмуртская республика
253	552	52	772	03	787,1	73,1	173	58,1	187	959	Республика Башкортостан
322	062	81	583	39	767,1	75,1	773	95,1	362	817	Ростовская область
097	992	85	162	39	317,1	39,1	807	69,1	973	239	Ставропольский край
21	11	01	6	8	7	9	5	7	3	2	1

Продолжение таблицы 3

1	2	3	4	5	6	7	8	9	10	11	12
Волгоградская область	639	41,0	16,7	422	14,7	171,8	54	27,8	35	27,0	25,7
Пензенская область	453	42,1	21,1	438	18,1	175,7	48	23,9	25	27,8	35,6
Самарская область	1164	36,6	12,2	439	8,6	183,2	40	35,3	53	25,9	19,5
Саратовская область	646	37,6	17,2	411	16,1	184,1	45	35,0	51	26,6	26,8
Ульяновская область	617	42,6	16,2	304	11,1	163,2	33	30,1	5	25,9	27,0
Республика Адыгея	532	32,1	23,4	359	16,0	168,3	50	24,3	29	27,9	46,1
Республика Дагестан	364	28,7	22,3	341	30,0	178,0	47	30,6	2	20,7	58,5
Кабардино-Балкарская республика	551	27,8	28,3	417	22,4	168,0	66	26,5	-16	22,8	42,5
Карачаево-Черкесская республика	442	32,6	22,6	369	25,5	173,3	66	32,0	-14	25,8	55,8
Республика Северная Осетия	664	23,0	14,2	443	26,6	174,2	64	43,1	3	28,2	31,1
Краснодарский край	666	33,6	16,5	357	16,2	174,8	58	22,1	43	27,3	46,3

Продолжение таблицы 3

1	Продолжение таблицы 6										
	2	3	4	5	6	7	8	9	10	11	12
Чукотский автономный округ	1686	65,7	12,4	1778	4,7	152,1	35	0,3	-558	16,4	30,7
Приморский край	850	47,3	14,5	577	14,9	170,4	45	27,5	-54	21,8	21,8
Хабаровский край	1024	48,3	15,0	583	12,4	163,9	40	43,5	-41	22,7	19,4
Амурская область	847	39,1	10,5	574	16,9	171,4	47	24,9	-61	20,4	35,8
Камчатская область	1561	52,1	12,8	968	17,6	162,3	42	25,9	-163	19,5	19,4
Магаданская область	1622	44,7	8,3	1098	18,1	153,7	40	24,1	-247	18,9	9,6
Сахалинская область	1113	58,6	14,9	820	17,1	174,3	38	16,4	-163	23,7	13,7
Калининградская область	630	40,4	19,9	429	16,7	202,5	44	21,7	137	22,9	22,3

Продолжение таблицы 5

1	2	3	4	5	6	7	8	9	10	11	12
Новосибирская область	778	36,4	13,5	576	13,7	170,6	48	42,6	55	25,4	26,1
Омская область	785	40,5	13,5	418	15,5	172,2	68	33,3	22	24,2	32,7
Томская область	805	45,6	16,7	480	14,6	174,6	40	55,7	11	22,9	33,5
Тюменская область	2166	56,0	7,4	628	14,0	164,5	42	23,7	21	16,9	23,9
Республика Бурятия	601	42,4	14,4	438	22,1	163,7	47	32,8	-46	22,1	40,3
Республика Тыва	491	44,0	24,0	606	20,9	154,2	54	19,5	-19	23,2	52,5
Республика Хакасия	655	45,8	16,2	557	9,6	182,3	46	21,4	21	23,9	29,2
Красноярский край	1034	56,0	13,4	461	16,4	160,4	40	30,4	-17	22,7	25,8
Иркутская область	1065	41,8	12,2	501	13,7	171,5	42	28,9	-10	23,8	20,4
Читинская область	513	55,4	21,0	609	20,4	165,8	45	17,7	-66	21,2	32,7
Республика Саха	1663	54,5	13,3	1023	13,6	155,4	32	18,2	-19	20,1	35,7
Еврейская автономная область	630	41,5	14,9	526	23,9	166,3	48	16,7	-93	21,7	32,4

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. - М.: Статистика, 1974. - 240с.
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичной обработки данных. - М.: Финансы и статистика, 1983. - 472с.
3. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. - М.: Финансы и статистика, 1985. - 488с.
4. Айвазян С.А., Бухштабер В.М., Енюков И.С. и др. Прикладная статистика. Классификация и снижение размерности. - М.: Финансы и статистика, 1989. - 342с.
5. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики: Учебник для вузов. - М.: ЮНИТИ, 1998. - 1022с.
6. Андерсон Т. Введение в многомерный статистический анализ / Пер. с англ. - М.: ГИФМЛ, 1963. - 500с.
7. Андрукович П.Ф. Применение метода главных компонент в практических исследованиях. - М.: Статистика, 1973. - 234с.
8. Афифи А., Эйзен С. Статистический анализ / Пер. с англ. - М.: Мир, 1985. - 488с.
9. Большев Л.П., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965. - 464с.
10. Болч Б., Хуань К. Многомерные статистические методы экономики / Пер. с англ. - М.: Статистика, 1979. - 317с.
11. Боровков А.А. Теория вероятностей. - М.: Наука, 1976. - 278с.
12. Ван-дер-Варден Б.Л. Математическая статистика / Пер. с нем. - М.: Иностранная литература, 1960. - 434с.
13. Ветров А.А., Ломоватский Д.И. Дисперсионный анализ. - М.: Статистика, 1975. - 214с.
14. Гнеденко Б.В. Курс теории вероятностей. - М.: Наука, 1976. - 534с.
15. Демиденко Е.З. Линейная и нелинейная регрессия. - М.:

Финансы и статистика, 1981. - 302с.

16. Дубров А.М. Последовательный анализ в статистической обработке информации. - М.: Статистика, 1976. - 160с.
17. Дубров А.М. Обработка статистических данных методом главных компонент. - М.: Статистика, 1978. - 136с.
18. Дубров А.П. Факторный и компонентный анализ. - М.: Изд-во МЭСИ, 1989. - 246с.
19. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учебник. - М.: Финансы и статистика, 1998. - 352с.
20. Елисеева И.И.,* Семенова Е.В. Основные процедуры многомерного статистического анализа: Учебное пособие. - СПб.: Изд-во СПб УЭФ, 1993. - 78с.
21. Жамбю М. Иерархический кластер-анализ и соответствия / Пер. с франц. - М.: Финансы и статистика, 1983. - 342с.
22. Жуковская В.М., Мучник И.Б. Факторный анализ в социально-экономических исследованиях. - М.: Статистика, 1976. - 152с.
23. Иберла К. Факторный анализ / Пер. с нем. - М.: Статистика, 1980. - 400с.
24. Информатика в статистике: словарь-справочник. - М.: Финансы и статистика, 1994. - 224с.
25. Кендалл М., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. - 112с.
26. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. - 302с.
27. Крастинь О.П. Разработка и интерпретация моделей корреляционных связей в экономике. - Рига: Зинатне, 1983. - 258с.
28. Лоули Д.Н., Максвелл А.З. Факторный анализ как статистический метод / Пер. с англ. - М.: Мир, 1967. - 144с.
29. Мандель И.Д. Кластерный анализ. - М.: Финансы и статистика, 1988. - 176с.
30. Мхитарян В.С. Статистические методы в управлении качеством продукции. - М.: Финансы и статистика, 1982. - 212с.
31. Мхитарян В.С., Трошин Л.И. Дисперсионный анализ. -

М.: Изд-во МЭСИ 1990. – 79с.

32. Окунь Я. Факторный анализ / Пер. с польск. - М.: Статистика, 1974. - 200с.

33. Сажин Ю.В., Макашкин Н.П., Ивенин Е.Н. Статистические методы исследования производственных процессов. - Саранск: Изд-во МГУ, 1989. – 79с.

34. Сажин Ю.В., Ивенин Е.Н. Многомерный анализ структуры и взаимосвязей технико-экономических показателей. - Саранск: Изд-во МГУ, 1991. – 128с.

35. Сажин Ю.В., Басова В.А., Катунь А.В. Многомерные статистические методы анализа экономических процессов. – Саранск: Изд-во Мордов. ун-та, 2000. – 88с.

36. Сошникова Л.А., Тамашевич В.Н., Шефер М., Усбе Г. Многомерный статистический анализ в экономике. М.: ЮНИТИ, 1999. - 598с.

37. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа: Руководство для экономистов. - М.: Финансы и статистика, 1985. - 98с.

38. Харман Г. Современный факторный анализ / Пер. с англ. - М.: Статистика, 1972. - 486с.

39. Хьюбер Дж. Робастность в статистике. - М.: Мир, 1984. - 304с.

Уважаемые читатели!
Издательство "Компания Спутник +"
и редакция журналов
«Аспирант и соискатель» и «Актуальные проблемы современной науки» предлагают Вам опубликовать:

- монографии, научные труды малыми тиражами (от 50 экз.);
- научные статьи для защиты диссертаций в журналах «Аспирант и соискатель» и «Актуальные проблемы современной науки»;
- книги, стихи малыми тиражами (от 50 экз.);
- авторефераты диссертаций (100 экз. за 1-3 дня)
- Все издания регистрируются в Книжной палате РФ и рассылаются по библиотекам России и СНГ.
- Осуществляем компьютерный набор и верстку, а также полиграфические работы (визитки, бланки, листовки, переплет).
- Приглашаем к сотрудничеству авторов учебной, деловой, справочной, краеведческой и другой популярной литературы.
- Оказываем помощь в реализации книжной продукции.

Тел. (095) 174-32-31 (с 9 до 18)
<http://www.sputnikplus.ru> E-mail: sputnikplus2000@mail.ru

Учебное пособие

*Сажин Юрий Владимирович
Басова Валентина Александровна*

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Издательство «Компания Спутник+»
109428, Москва, Рязанский проспект, д. 8а
Тел.: (095) 174-32-31

ЛР № 066478 от 30.03.99
Налоговые льготы в соответствии с ОК 005-93
Тем- Книги и брошюры

Гигиенический сертификат № 77.99.2.953.11.11082.5.00 от 06.05.2000 г.

Подписано в печать 13.02.2002. Формат 60х90/16.
Бумага офсетная.
Усл. печ. л. 10.19. Тираж 200 экз. Заказ 26.

Отпечатано в ООО «Компания Спутник +»
ПД № 1-00007 от 28.07.2000