Министерство образования и науки Российской Федерации

ГОУ ВПО “Ивановский государственный химико-технологический университет”

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ
КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Методические указания

Составитель Т.А. Баранова

Иваново 2007

Составитель Т.А. Баранова

УДК 519.237.5(072)

Многомерные статистические методы. Корреляционный анализ: Метод. указания / Иван. гос. хим.-технол. ун-т.; Сост. Т.А. Баранова. – Иваново, 2007. – 40с.

Методические указания разработаны в соответствии с Государственным образовательным стандартом и предназначены для студентов ИГХТУ специальности 061800 ″Математические методы в экономике″.

Рассматриваются многомерные генеральная и выборочная совокупности, статистическое оценивание и сравнение генеральных совокупностей, корреляционный анализ. Приведены основные теоретические положения о корреляционном анализе, примеры решения задач и практические задания, предназначенные для самостоятельной работы студентов.

Табл. 17. Библиогр.: 7 назв.

Рецензент

доктор экономических наук, профессор А.Н. Ильченко (Ивановский государственный химико–технологический университет)

ВВЕДЕНИЕ

Цели, задачи изучения, сфера профессионального применения дисциплины ”Многомерные статистические методы”

Квалифицированному специалисту недостаточно иметь знания только в области экономики, ему также необходимо уметь учитывать сложную взаимосвязь различных факторов, оказывающих существенное влияние на важнейшие экономические процессы. Анализ производительности труда, объема производства, расхода сырья и других видов ресурсов предполагает знание специалистом теории вероятностей, математической статистики и многомерных статистических методов. Сложность состоит в том, что на строго детерминированные процессы и явления накладываются случайные явления. Следовательно, нельзя проводить экономические и статистические исследования без учета действия случайных факторов, без знания основ теории вероятностей и математической статистики – дисциплин, занимающихся количественным анализом закономерностей массовых случайных явлений.

Случайное в единичном явлении становится закономерным в массовом явлении. Применение методов математической статистики становится возможным благодаря теоретической и практической обоснованности перехода от случайного в единичных явлениях к объективной закономерности в массе таких явлений. Знание закономерностей, которым подчиняются массовые случайные явления, позволяет решать задачи прогнозирования развития этих явлений, расширяет возможности научного анализа и принятия решений. Математическая статистика находит применение в экономике, бизнесе, технике, производстве, научных исследованиях, социологии, страховом деле, демографии.

Социально–экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т.е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно–статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

Многомерный экономико–статистический анализ опирается на широкий спектр методов. Проведение системного анализа до изучения взаимосвязей в многомерной совокупности требует иметь представление о связях между отдельной зависимой переменной и группой влияющих на нее показателей. Это может быть осуществлено при помощи множественного корреляционного и регрессионного анализа.

Методы многомерной классификации, которые предназначены разделять рассматриваемые совокупности объектов, субъектов или явлений на группы в определенном смысле однородные. Необходимо учитывать, что каждый из рассматриваемых объектов характеризуется большим количеством разных и стохастически связанных признаков. Для решения столь сложных задач классификации применяют кластерный и дискриминантный анализ. Наличие множества исходных признаков, характеризующих процесс функционирования объектов, заставляет отбирать из них наиболее существенные и изучать меньший набор показателей. Чаще исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации. Такое решение может быть обеспечено методами снижения размерности, куда относятся факторный и компонентный анализ. Эти методы позволяют учитывать эффект существенной многомерности данных, дают возможность лаконичного или более простого объяснения многомерных структур. Они вскрывают объективно существующие, непосредственно не наблюдаемые закономерности при помощи полученных факторов или главных компонент.

Последние дают возможность достаточно просто и точно описать наблюдаемые исходные данные, структуру и характер взаимосвязей между ними. Сжатие информации получается за счет того, что число факторов или главных компонент – новых единиц измерения – значительно меньше, чем было исходных признаков.

Задача оценки тесноты связи между системами показателей приводит к каноническим корреляциям.

Все перечисленные методы могут быть усвоены только при активном применении статистических пакетов прикладных программ для ПЭВМ.

МНОГОМЕРНАЯ НОРМАЛЬНО РАСПРЕДЕЛЕННАЯ ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ

Закономерности, которым подчиняется в математической статистике исследуемая переменная, полностью определяются комплексом условий ее наблюдения. Математически эти закономерности задаются соответствующим законом распределения вероятностей.

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

В простейшем случае генеральная совокупность есть одномерная случайная величина X с функцией распределения $F(x)=P(X<x)$, которая определяет вероятность того, что X примет значение, меньшее фиксированного числа x.

В многомерном статистическом анализе изучаются генеральные совокупности с точки зрения нескольких признаков (обычно более двух). 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 являются случайный вектор X (или случайная точка) 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:

$$F(x) = P(X_1 < x_1, X_2 < x_2, \dots X_k < x_k) = P(X < x), \tag{1}$$

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

Закон распределения полностью характеризует генеральную совокупность, т.е. последняя может быть полностью задана функцией распределения или плотностью распределения. Однако, такая исчерпывающая 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

$$Mx = \sum_{i-1}^{N} x_i p_i , \tag{2}$$

где   вероятность появлений i–го возможного значения x_i случайной величины X (i=1, 2,…, N).

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

Матрица

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1j} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2j} & \dots & \sigma_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{i1} & \sigma_{i2} & \dots & \sigma_{ij} & \dots & \sigma_{ik} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kj} & \dots & \sigma_{kk} \end{pmatrix} \tag{3}$$

называется ковариационной матрицей случайного вектора X. Она симметрическая и неотрицательно определена. Симметрическая матрица, совпадает со своей транспонированной, $a_{ij} = a_{ji}$. В случае кососимметрической матрицы $a_{ij} = -a_{ji}$. На главной диагонали ковариационной матрицы находятся дисперсии, а на побочной диагонали – корреляционные моменты (коэффициенты ковариации) случайного вектора X.

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

$$\rho_{ij} = M\left(\frac{x_i - Mx_i}{\sigma_i} \cdot \frac{x_j - Mx_j}{\sigma_j}\right) = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \tag{4}$$

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

Квадрат коэффициента корреляции называют коэффициентом детерминации.

Матрица

$$R = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix} \tag{5}$$

называется корреляционной. Она симметрическая и неотрицательно определена.

При рассмотрении различных моделей статистического анализа часто предполагается нормальное распределение всех или некоторых признаков генеральной совокупности. Говорят, что непрерывная $\boldsymbol{k}$–мерная случайная величина распределена нормально, если плотность распределения имеет вид

$$p(x) = \left[(2\pi)^k |\Sigma|\right]^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}\ , \qquad (6)$$

где $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_k \end{pmatrix}$ — k-мерный вектор математических ожиданий; Σ^{-1}-матрица, обратная ковариационной матрице Σ размерности $(k \times k)$; $|\Sigma|$-определитель этой матрицы. Напомним, что матрица Σ является симметрической и положительно определенной.

Таким образом, многомерный нормальный закон распределения определяется вектором математических ожиданий μ и ковариационной матрицей Σ, т.е. $\left[k + \frac{k(k+1)}{2}\right]$ параметрами генеральной совокупности.

Пример 1. Показать, что при k=1 имеет место одномерный нормальный закон распределения.

Решение. При k=1 $\Sigma = \sigma_{11} = \sigma^2$. Тогда $|\Sigma| = \sigma^2$, а обратная матрица $\Sigma^{-1} = \frac{1}{\sigma^2}$. Подставляя найденные значения в выражение для плотности распределения (6), имеем

$$P(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\ . \qquad (7)$$

Мы получили плотность распределения одномерного нормального закона, зависящего от двух параметров: математического ожидания μ и среднего квадратического отклонения σ.

Пример 2. Приняв k=2 в выражении (6) для плотности нормально распределенной k–мерной случайной величины, получить плотность двумерного нормального закона распределения.

Решение. При k=2 имеем

$$(x-\mu) = \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix};\ \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

где $\sigma_{12}=\sigma_{21}$., т. к. матрица Σ симметрическая.

Тогда, приняв $\sigma_{11}=\sigma_1^2$ и $\sigma_{22}=\sigma_2^2$, σ_{11} , σ_{22} – дисперсии, σ_1 и σ_2 - среднеквадратические отклонения, получим

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}.$$

Откуда, согласно выражению (6) для плотности нормального распределения k-мерной случайной величины

$$Q(x_1,x_2) = (x-\mu)^T\Sigma^{-1}(x-\mu) = (x_1-\mu_1\ \ x_2-\mu_2)\frac{1}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\times$$

$$\times\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix} = \frac{1}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\times$$

$$\times\left((x_1-\mu_1)\sigma_1^2-(x_2-\mu_2)\sigma_{12}-\sigma_{12}(x_1-\mu_1)+\sigma_1^2(x_2-\mu_2)\right)\times$$

$$\times\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix} = \frac{1}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\times$$

$$\times\left[(x_1-\mu_1)^2\sigma_1^2-2(x_1-\mu_1)(x_2-\mu_2)\sigma_{12}+(x_2-\mu_2)^2\sigma_2^2\right].$$

где $Q(x_1,\ x_2)$ так переобозначили то, что стоит в фигурных скобках в выражении (6).

Разделив числитель и знаменатель на $\sigma_1^2\sigma_2^2$ и учитывая, что $\frac{\sigma_{12}}{\sigma_1\sigma_2}=\rho_{12}=\rho$, получим

$$Q(x_1,x_2) = \frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)}{\sigma_1}\frac{(x_2-\mu_2)}{\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]. \qquad (8)$$

Учитывая также, что $\frac{1}{\sqrt{(2\pi)^2|\Sigma|}} = \frac{1}{2\pi\sqrt{(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}} = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$, окончательно получаем

$$p(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}Q(x_1,x_2)\right\}. \qquad (9)$$

Из выражений (8), (9) следует, что плотность двумерного нормального закона распределения определяется пятью параметрами: математическими ожиданиями μ_1 и μ_2 случайных величин x_1 и x_2, их средними квадратическими отклонениями σ_1 и σ_2 и коэффициентом корреляции ρ. Определенная таким образом нормально распределенная генеральная совокупность является невырожденной. В случае вырожденности совокупности компоненты случайного вектора X являются линейно зависимыми. Следовательно, часть их представляет собой линейные комбинации остальных компонент, образующих линейно независимую подсистему.

Выборкой из генеральной совокупности (*X*) называют результаты ограниченного ряда наблюдений $x_1, x_2,\ldots, x_n$, где n–объем выборки.

Выборку рассматривают как некий эмпирический аналог генеральной совокупности, с которым чаще всего на практике имеют дело, поскольку обследование всей генеральной совокупности бывает либо слишком трудоемко, либо принципиально невозможно.

Задачи математической статистики фактически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам выборки. Достоверность выводов, получаемых в результате статистической обработки данных, во многом зависит от успешного решения вопроса представительности выборки, т.е. полноты и адекватности представления свойств анализируемой генеральной совокупности (так называемое свойство репрезентативности выборки). Это достигается случайностью отбора, когда каждый элемент генеральной совокупности имеет одинаковую вероятность быть отобранным. При оценке представительности выборки учитывается и то, как выборка получена, и то, насколько распределение показателей в выборке характерно для анализируемой генеральной совокупности в целом.

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

- генеральная совокупность настолько многочисленна, что проведение обследования всех элементов совокупности (сплошное обследование) слишком трудоемко. С такой ситуацией приходится встречаться при контроле качества продукции крупносерийного и массового производства;
- 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 процессе проведения испытания происходит разрушение отбираемых образцов (например, испытание предела прочности изделия);
- встречаются обстоятельства, когда мы располагаем результатами испытания всей совокупности, реально существующей на данный момент времени, но рассматриваем их как выборку из гипотетической генеральной совокупности. Так поступают в тех случаях, когда хотят выявить общую закономерность, по отношению к которой имеющаяся совокупность представляется лишь частным случаем.

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ И СРАВНЕНИЕ МНОГОМЕРНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

Точечные оценки параметров многомерной генеральной совокупности

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

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \qquad (10)$$

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 (или n точек в k–мерном пространстве).

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

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{11} \\ s_{21} & s_{22} & \dots & s_{2k} \\ \dots & \dots & \dots & \dots \\ s_{k1} & s_{k2} & \dots & s_{kk} \end{pmatrix}, \; s_{lj} = \frac{1}{n}\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j); \; l, j = \overline{1,k}. \qquad (11)$$

Вместо S используют также несмещенную оценку Σ:

$$\hat{S} = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} & \dots & \hat{s}_{1k} \\ \hat{s}_{21} & \hat{s}_{22} & \dots & \hat{s}_{2k} \\ \dots & \dots & \dots & \dots \\ \hat{s}_{k1} & \hat{s}_{k2} & \dots & \hat{s}_{kk} \end{pmatrix} = \frac{n}{n-1}S, \; \hat{s}_{lj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{ij} - \bar{x}_j); \; l, j = \overline{1,k}. \qquad (12)$$

Оценку корреляционной матрицы (ρ_{ij}) можно получить по формуле

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}, \qquad (13)$$

где $r_{ij} = \dfrac{s_{ij}}{s_i s_j}$ – оценка парного коэффициента корреляции между i-й и j-й компонентами x.

Для получения оценок параметров условных распределений требуются специально организованные выборки, т.е. выборки при фиксированных значениях части компонент генеральной совокупности. На практике такой подход можно осуществить с помощью группировки данных по закрепленным значениям части признаков, равных либо их естественным дискретным значениям, либо серединным, центральным значениям областей (интервалов, прямоугольников, брусов и т.д.) группирования.

Доверительная область для вектора математического ожидания

При малых объемах выборок или подвыборок точечные оценки могут достаточно далеко отклоняться от оцениваемых параметров и поэтому вводится понятие интервальной оценки параметра генеральной совокупности.

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

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

1) Предположим, что ковариационная матрица известна, тогда с надежностью γ можно утверждать, что вектор μ накрывается доверительной областью, задаваемой неравенством

$$n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \le (\chi^2)^{-1}(1 - \gamma). \tag{14}$$

2) Пусть теперь ковариационная матрица Σ неизвестна. В этом случае с использованием статистики T^2 Хотеллинга и F–распределения получают уравнение поверхности, ограничивающей доверительную область с надежностью γ

$$(\bar{x} - \mu)^T \hat{S}^{-1} (\bar{x} - \mu) = \frac{k(n-1)}{n(n-k)} F_{\alpha,k,n-k}, \tag{15}$$

где $F_{\alpha,k,n-k}$ –точка F–распределения, соответствующая уровню значимости α и числам степеней свободы k и n-k. Уравнение (15) определяет k-мерный эллипсоид (эллипс при k=2) с центром $\bar{x}$, так как его левая часть представляет положительно определенную квадратичную форму относительно μ.

Пример 3. В таблице приводятся данные о фонде заработной платы работников централизованных бухгалтерий x_1 и товарообороте обслуживаемых аптек x_2.

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

Найти оценки математических ожиданий, дисперсии и коэффициенты корреляции, доверительную область для вектора математических ожиданий с надежностью γ =0,95.

Решение.

1. Найдем средние арифметические $\overline{x_1}$ и $\overline{x_2}$:

$$\overline{x_1} = \frac{20+36+28+51+70+45+30+56}{8} = 42$$

$$\overline{x_2} = \frac{3,5+5,4+2,7+9,8+10,1+6,2+2,4+9,5}{8} = 6,2$$

2. Перейдем к центрированным величинам $u_{ij} = x_{ij} - \overline{x_i}$:

$u_{11} = x_{11} - \overline{x_1} = 20 - 42 = -22$ $\quad$ $u_{21} = x_{21} - \overline{x_2} = 3,5 - 6,2 = -2,7$

$u_{12} = x_{12} - \overline{x_1} = 36 - 42 = -6$ $\quad$ $u_{22} = x_{22} - \overline{x_2} = 5,4 - 6,2 = -0,8$

$u_{13} = x_{13} - \overline{x_1} = 28 - 42 = -14$ $\quad$ $u_{23} = x_{23} - \overline{x_2} = 2,7 - 6,2 = -3,5$

$u_{14} = x_{14} - \overline{x_1} = 51 - 42 = 9$ $\quad$ $u_{24} = x_{24} - \overline{x_2} = 9,8 - 6,2 = 3,6$

$u_{15} = x_{15} - \overline{x_1} = 70 - 42 = 28$ $\quad$ $u_{25} = x_{25} - \overline{x_2} = 10,1 - 6,2 = 3,9$

$u_{16} = x_{16} - \overline{x_1} = 45 - 42 = 3$ $\quad$ $u_{26} = x_{26} - \overline{x_2} = 6,2 - 6,2 = 0$

$u_{17} = x_{17} - \overline{x_1} = 30 - 42 = -12$ $\quad$ $u_{27} = x_{27} - \overline{x_2} = 2,4 - 6,2 = -3,8$

$u_{18} = x_{18} - \overline{x_1} = 56 - 42 = 14$ $\quad$ $u_{28} = x_{28} - \overline{x_2} = 9,5 - 6,2 = 3,3$

3. Составляем матрицу U из центрированных величин и транспонированную U^T:

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

4. Перемножаем матрицы

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

5. Находим оценки дисперсий и среднеквадратических отклонений

$$\hat{S} = \frac{1}{n-1} \cdot U^T \cdot U = \frac{1}{8-1} \cdot \begin{pmatrix} 1930 & 346,6 \\ 346,6 & 73,68 \end{pmatrix} = \frac{1}{7} \cdot \begin{pmatrix} 1930 & 346,6 \\ 346,6 & 73,68 \end{pmatrix} = \begin{pmatrix} 275,714 & 49,514 \\ 49,514 & 10,526 \end{pmatrix}$$

$\hat{S}_{11} = 275{,}714$; $\hat{S}_{22} = 10{,}526$ – оценки дисперсий;

$\hat{S}_1 = \sqrt{\hat{S}_1^2} = 16{,}605$; $\hat{S}_2 = \sqrt{\hat{S}_2^2} = 3{,}244$ – оценки средних квадратических отклонений.

6. Находим выборочный коэффициент корреляции

$$r_{ij} = \frac{S_{ij}}{S_i \cdot S_j} = \frac{49{,}514}{16{,}605 \cdot 3{,}244} = 0{,}919\,.$$

7. Находим матрицу, обратную к $\hat{S}$:

$$\hat{S}^{-1} = \frac{1}{\det \hat{S}} \cdot \begin{pmatrix} 10{,}526 & -49{,}514 \\ -49{,}514 & 275{,}714 \end{pmatrix} = \frac{1}{450{,}523} \cdot \begin{pmatrix} 10{,}526 & -49{,}514 \\ -49{,}514 & 275{,}714 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}023365 & -0{,}109909 \\ -0{,}109909 & 0{,}6120157 \end{pmatrix}$$

$$\det \hat{S} = 275{,}714 \cdot 10{,}526 - 49{,}514^2 = 450{,}523$$

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

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

где $F(0{,}05;2;6)=5{,}14$ находим по таблице распределения Фишера–Снедекора (F–распределение). После преобразований (матрицы в левой части перемножаются друг на друга) получаем уравнение эллипса $0{,}0234 \cdot (42 - \mu_1)^2 + 0{,}0242 \cdot (42 - \mu_1) \cdot (6{,}2 - \mu_2) + 0{,}612 \cdot (6{,}2 - \mu_2)^2 = 1{,}499$, которое определяет границы доверительной области для вектора $(\mu_1, \mu_2)^T$.

ГИПОТЕЗЫ О ПАРАМЕТРАХ МНОГОМЕРНОЙ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Сравнение вектора генеральных средних со стандартом

Рассматривается k–мерная генеральная совокупность, имеющая нормальное распределение $N_k(\mu, \Sigma)$, $|\Sigma| \neq 0$. По выборке объема n из этой генеральной совокупности определены вектор средних арифметических и несмещенная оценка $\hat{S}$ ковариационной матрицы Σ. Возможны два случая.

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

$$\chi^2 = n(\bar{x} - \mu_0)^T \Sigma^{-1} (\bar{x} - \mu_0), \tag{16}$$

имеющая распределение хи–квадрат с числом степеней свободы $\nu = k$ при справедливости гипотезы H_0.

2) Если ковариационная матрица неизвестна, то можно воспользоваться статистикой Хотеллинга

$$T^2 = n(\bar{x} - \mu_0)^T \hat{S}^{-1} (\bar{x} - \mu_0). \tag{17}$$

При истинности H_0 имеет место равенство

$$T^2_{\alpha,k,n-k} = \frac{k(n-1)}{n-k} F_{\alpha,k,n-k}. \tag{18}$$

Поэтому для критической области $T^2 > T^2_{kp} = T^2_{\alpha,k,n-k}$ можно вычислить $T^2_{\hat{e}\eth}$ с помощью таблиц F–распределения. 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 χ^2 с одной степенью свободы, а квадрат случайной величины, имеющей t-распределение Стьюдента, распределен по F–распределению с числом степеней свободы 1 и n–1.

Пример 4. По данным годовых отчетов n=10 промышленных предприятий (данные в таблице) проверить при $\alpha = 0{,}05$ гипотезу о соответствии средних уровней экономических показателей работы группы предприятий (x_1–объем валовой продукции, x_2–производительность труда, x_3–себестоимость товарной продукции) контрольным значениям при известных значениях указанных в таблице параметров генеральной совокупности, которая является нормально распределенной.

Таблица

Исходная информация для сравнения параметров

№ i	1	2	3	4	5	6	7	8	9	10	σ_j^2	Парная корреляция	Контроль-ные значения μ_{0j}
x_{i1}	3	4	5	5	5	5	6	7	10	10	4	$\rho_{12} = 0{,}14\sqrt{5}$	7
x_{i2}	1,2	1,2	1,4	1,2	1,2	1,5	1,5	1,3	1,7	1,6	0,2	$\rho_{23} = 0{,}12\sqrt{5}$	1,5
x_{i3}	2,1	2,8	3,2	4,5	4,8	4,9	5,5	6,5	8,5	8,2	4	ρ_{31}=0,9	5

Решение. Исходя из условия задачи, требуется проверить гипотезу H$_0$:$\mu = \mu_0$ против альтернативной гипотезы H$_1$:$\mu \neq \mu_0$, причем $\mu_0 = (7;1{,}5;5)^T$. Здесь матрица Σ генеральных коэффициентов ковариации считается известной.

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

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

$\sigma_2 = \sqrt{0{,}2}; \sigma_{23} = \sigma_{32} = \rho_{23}\sigma_2\sigma_1 = 0{,}12\sqrt{5}\cdot 2\sqrt{0{,}2} = 0{,}24; \sigma_3 = 2.$ 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}, используя метод обратной матрицы. Будем иметь

$$|\Sigma| = 4\cdot 4\cdot 0{,}2 + 0{,}28\cdot 0{,}24\cdot 3{,}6 + 3{,}6\cdot 0{,}28\cdot 0{,}24 - 3{,}6\cdot 0{,}2\cdot 3{,}6 - 0{,}24\cdot 0{,}24\cdot 4 -$$
$$- 0{,}28\cdot 0{,}28\cdot 4 = 0{,}54784$$

$$\sigma_{11}^{-1} = (-1)^2\begin{vmatrix} 0{,}2 & 0{,}24 \\ 0{,}24 & 4 \end{vmatrix} : 0{,}54784 = (0{,}8 - 0{,}24^2) : 0{,}54784 = 1{,}3551400;$$

$$\sigma_{12}^{-1} = (-1)^3\begin{vmatrix} 0{,}28 & 0{,}24 \\ 3{,}6 & 4 \end{vmatrix} : 0{,}54784 = -0{,}4672897 = \sigma_{21}^{-1};$$

$$\sigma_{13}^{-1} = (-1)^4\begin{vmatrix} 0{,}28 & 0{,}2 \\ 3{,}6 & 0{,}24 \end{vmatrix} : 0{,}54784 = -1{,}1915887 = \sigma_{31}^{-1};$$

$$\sigma_{22}^{-1} = (-1)^4\begin{vmatrix} 4 & 3{,}6 \\ 3{,}6 & 4 \end{vmatrix} : 0{,}54784 = 5{,}5490654;$$

$$\sigma_{23}^{-1} = (-1)^5\begin{vmatrix} 4 & 0{,}28 \\ 3{,}6 & 0{,}24 \end{vmatrix} : 0{,}54784 = 0{,}0876168 = \sigma_{32}^{-1};$$

$$\sigma_{33}^{-1} = (-1)^6\begin{vmatrix} 4 & 0{,}28 \\ 0{,}28 & 0{,}2 \end{vmatrix} : 0{,}54784 = 1{,}3171728.$$

Таким образом, 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 неравенством $n(\bar{x} - \mu_0)^T \Sigma^{-1}(\bar{x} - \mu_0) > \chi^2(\alpha, k)$. Подставляя в него n=10, $\mu_0 = (7;1,5;5)^T$, Σ^{-1}, указанные выше, а также $\chi^2(0,05;3) =$ 7,815, получим неравенство относительно средних арифметических $\bar{x}_1, \bar{x}_2, \bar{x}_3$ – компонент вектора x. Далее находим наблюдаемое значение вектора x: $x = (6;1,4;5,1)^T$. Подставляя его в левую часть неравенства, получим

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

$$= 10(-1,4275700; -0,0788551; 1,3145443) \times (-1; -0,1; 0,1)^T = 15,669099 = 15,669.$$

Так как $\chi^2_{\text{набл}} = 15,669 > \chi^2(0,05;3) = 7,815$, т.е. $\bar{x}_{\text{набл}}$ 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 последствий обнаруженного несоответствия выясняется планирующими и управляющими органами.

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

При изучении данной темы следует акцентировать внимание на понимании сути корреляционной зависимости. Изучение корреляционной зависимости между переменными сводится к измерению тесноты связи, отбору факторов, оказывающих наиболее существенное влияние на результативный признак, обнаружению неизвестных причин связей, построению корреляционной модели и оценке ее параметров, проверке значимости параметров связи и их интервальному оцениванию.

Задачи и проблемы корреляционного анализа

Корреляционный анализ, разработанный К. Пирсоном и Дж. Юлом, является одним из методов статистического анализа взаимозависимости нескольких признаков – компонент случайного вектора *X*.

При построении корреляционных моделей исходят из условия нормальности многомерного закона распределения генеральной совокупности. Эти условия обеспечивают линейный характер связи между изучаемыми признаками. На практике не всегда строго соблюдаются предпосылки корреляционного анализа, а также возможны случаи, когда с помощью корреляционных моделей обнаруживают достаточно сильную «зависимость» признаков, в действительности не имеющих причинной связи друг с другом. Такие корреляции называют ложными.

Одним из основных показателей взаимозависимости двух случайных величин является парный коэффициент корреляции, служащий мерой линейной статистической зависимости между этими величинами. Этот показатель соответствует своему прямому назначению, когда статистическая связь между соответствующими признаками в генеральной совокупности линейна. Это касается также частных и совокупных коэффициентов корреляции. Одним из требований, определяющих корреляционный метод, является требование линейности статистической связи, т.е. линейности возможных уравнений (средней квадратической) регрессии. Указанные условия выполняются, если генеральная совокупность распределена по нормальному закону. Основная задача корреляционного анализа состоит в оценке $k(k+3)/2$ параметров, определяющих нормальный закон распределения k–мерного вектора X, в частности, корреляционной матрицы генеральной совокупности X по выборке.

Для значимых парных коэффициентов корреляции имеет смысл указать более предпочтительные точечные или интервальные оценки. Далее следует оценить и проверить значимость множественных коэффициентов корреляции или детерминации всевозможных подсистем x_j (j=1, k), содержащих три и более различных случайных величины x_j. Для выяснения «чистых», истинных взаимозависимостей следует проанализировать выборочные частные коэффициенты корреляции.

Таким образом, решение основной задачи корреляционного анализа позволяет определить расположение «облака» точек в пространстве k-измерений, т.е. оценить природу взаимозависимости между наблюдаемыми переменными.

Дополнительная задача корреляционного анализа состоит в оценке уравнений регрессии, где в качестве результативного признака выступает признак, являющийся следствием других признаков (факторов) – причин. Причинно–следственная связь устанавливается из внестатистических соображений, например, из аргументов, касающихся физической природы явлений.

Иногда имеет смысл оценить уравнение регрессии для измерения результативного признака по факторным, несмотря на то, что причинно–следственной связи на самом деле между ними не существует. Здесь причиной могут быть другие факторы, не рассматриваемые в модели, но действующие как на функцию, так и на аргументы уравнения регрессии. Так следует поступать в том случае, когда непосредственное измерение результативного признака затруднительно, но существует тесная корреляционная связь (коэффициент множественной корреляции достаточно близок к единице) между результативным признаком и факторными, измерять и наблюдать которые легче в последующих исследованиях.

Назовем параметр связи в генеральной совокупности значимо отличающимся от нуля (значимым), если гипотеза о равенстве нулю этого параметра отвергается с заданным уровнем значимости α . Если же эта гипотеза принимается, генеральный параметр связи называется незначимым. В корреляционной модели соответствующая связь считается недоказанной или отсутствующей.

На примере трехмерной генеральной совокупности достаточно четко просматриваются основные задачи и особенности многомерного корреляционного анализа. Для изучения многообразия взаимозависимости между переменными в модели используют три меры тесноты корреляционной связи – парный, частный, множественный коэффициенты корреляции.

Среди всех методов статистического оценивания наибольшее предпочтение в практике оценки корреляционных моделей отдают методу наименьших квадратов (МНК). Для собственно линейных моделей этот метод дает несмещенные оценки и удовлетворяет требованиям точности оценивания (см. предпосылки Гаусса-Маркова).

Для изучения взаимосвязи признаков, не поддающихся количественному измерению, используют различные показатели ранговой корреляции. Например, Спирмэна, Кендалла, конкордации, ассоциации, контингенции и другие.

Нелинейная связь между показателями возникает в том случае, если равномерным изменениям одной величины соответствуют неравномерные изменения другой, причем эта неравномерность носит закономерный характер. Для изучения связи между нелинейно зависимыми переменными используют корреляционное отношение.

Знания, умения, навыки по теме «Корреляционный анализ»

Изучив тему «Корреляционный анализ», студент должен

Знать:

1. Особенности корреляционной зависимости между переменными.
2. Задачи корреляционного анализа.
3. Применение метода наименьших квадратов для оценки параметров корреляционных моделей.
4. Меры тесноты связи в корреляционных моделях.
5. Алгоритм проверки значимости параметров связи корреляционных моделей.
6. Расчет интервальных оценок для параметров корреляционных моделей.
7. Процедуру анализа двумерной и трехмерной моделей.
8. Коэффициенты ранговой корреляции.
9. Алгоритм измерения тесноты связи в нелинейных моделях.

Уметь:

1. Самостоятельно строить и анализировать двумерные и трехмерные корреляционные модели.

2. Решать задачи с использованием теоретического материала.
3. Использовать статистические пакеты для решения задач корреляционного анализа.

<u>Получить навыки:</u> построения и анализа корреляционных моделей в профессиональной деятельности, использования для этих целей статистических пакетов.

Двумерная модель

Наиболее простой случай корреляционной зависимости: две нормально распределенные случайные величины. Двумерная линейная корреляционная модель характеризуется пятью параметрами: два математических ожидания, две дисперсии, парный линейный коэффициент корреляции.

Рассмотрим генеральную совокупность с двумя признаками x и y, совместное распределение которых задано плотностью двумерного нормального закона

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}Q_2(x,y)\right\}, \qquad (19)$$

$$Q_2(x,y) = \frac{1}{2\sqrt{1-\rho^2}}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\,\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right], \qquad (20)$$

$$Mx = \mu_x,\ My = \mu_y,\ Dx = \sigma_x^2,\ Dy = \sigma_y^2,\ M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}\right] = \rho,\ \rho^2 \neq 1.$$

Имея эти параметры, можно получить уравнения линий регрессии, показывающих изменение условных математических ожиданий в зависимости от изменения соответствующих значений случайных аргументов: $My/x - My = \beta_{yx}(x - Mx)$ – прямая регрессии y на x; $Mx/y - Mx = \beta_{xy}(y - My)$ – прямая регрессии x на y; $\beta_{yx} = \rho\frac{\sigma_y}{\sigma_x}$ – коэффициент регрессии y на x; $\beta_{xy} = \rho\frac{\sigma_x}{\sigma_y}$ – коэффициент регрессии x на y.

Квадрат коэффициента корреляции ρ^2, т.е. коэффициент детерминации, в рассматриваемой модели указывает долю дисперсии одной случайной величины , обусловленную вариацией другой. Коэффициент β_{yx} показывает, на сколько единиц своего измерения увеличится ($\beta>0$) или уменьшится ($\beta<0$) в среднем y (My/x), если x увеличить на единицу своего измерения.

Задача двумерного корреляционного анализа состоит прежде всего в оценке пяти параметров, определяющих генеральную совокупность.

Точечные оценки неизвестных других параметров: $\bar{x}$ – оценка для μ_x; $\bar{y}$ – оценка для μ_y; $\overline{x^2}$ – оценка для $M(x^2)$; $\overline{y^2}$ – оценка для $M(y^2)$; $\overline{xy}$ – оценка для

$M(xy)$; $s_x^2 = \overline{x^2} - (\bar{x})^2$ – оценка для σ_x^2; $s_y^2 = \overline{y^2} - (\bar{y})^2$ – оценка для σ_y^2; $r = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y}$ – оценка для ρ. Оценки генеральных коэффициентов регрессии β_{yx} и β_{xy} получаются соответственно по формулам $b_{yx} = r\dfrac{s_y}{s_x}$, $b_{xy} = r\dfrac{s_x}{s_y}$.

Оценки уравнений регрессии имеют вид $\overline{y/x} - \bar{y} = b_{xy}(x - \bar{x})$, $\overline{x/y} - \bar{x} = b_{xy}(y - \bar{y})$.

Следует отметить, что вышеприведенные точечные оценки являются состоятельными, а $\bar{x}$ и $\bar{y}$ несмещенными и эффективными. Кроме того, распределение выборочных средних ($\bar{x}, \bar{y}$) не зависит от распределения (s_x^2, s_y^2, r). Выборочный коэффициент корреляции по абсолютной величине не превосходит единицы.

Приемы вычисления выборочных характеристик

Если объем выборки невелик, то наблюдаемые точки располагают в таблице в порядке их регистрации и обрабатывают по следующей схеме:

x	y	x^2	y^2	xy
. . x_j . .	. . y_j . .	. . x_j^2 . .	. . y_j^2 . .	. . $x_j\, y_j$ . .
Σx_j	Σy_j	Σx_j^2	Σy_j^2	$\Sigma x_j y_j$
$\bar{x}_j$	$\bar{y}_j$	$\overline{x_j^2}$	$\overline{y_j^2}$	$\overline{x_j y_j}$

В схеме последовательно заполняют столбцы таблицы результатами операций, указанных сверху. В предпоследней строке вычисляются соответствующие суммы элементов столбцов, в последней строке вычисляются средние значения элементов по столбцам.

Если выборка многочисленна, то данные группируются путем построения двумерного интервального ряда, корреляционная таблица для которого имеет вид:

y\x	…	$(a_k\text{-}b_k]$	…	m_y
. . $(c_l\text{-}d_l]$ . .	. . … . .	. . m_{kl} . .	. . … . .	. . m_{*1} . .
m_x	…	m_{k*}	…	n

Далее рассчитываются групповые характеристики: групповые средние, групповые исправленные дисперсии, коэффициент корреляции, коэффициенты регрессии. Следует отметить, что при группировке вычисленные характеристики могут сильно отличаться от выборочных. Оценки по группированным данным s_x^2, s_y^2 можно улучшить поправками Шеппарда $s_x^2 \cong s_{x_{гр}}^2 - \frac{1}{12}h_x^2$; $s_y^2 \cong s_{y_{гр}}^2 - \frac{1}{12}h_y^2$.

Эти поправки часто сглаживают ошибки, возникающие от группировки, если длина интервала (h) не превосходит восьмой части размаха соответствующего признака.

Проверка значимости параметров связи

В двумерной модели достаточно проверить значимость только коэффициента корреляции. Осуществляется это с помощью *t*–статистики, имеющей распределение Стьюдента с $\nu = n - 2$ степенями свободы.

Интервальные оценки параметров связи

Для значимых параметров связи имеет смысл найти интервальные оценки для генерального среднего и для значимого коэффициента корреляции. Осуществляют вышеназванные оценки с помощью статистики Фишера

$$Z_r = \frac{1}{2}\ln\frac{1+r}{1-r} \qquad (21)$$

Корреляционный анализ оплаты и затрат труда

Пример 5. На основе следующих данных годового отчета 100 хозяйств об оплате труда (y) в хлопководстве, измеряемой в тыс. руб. за 1 чел.–ч., и затратах труда (x), измеряемых количеством человеко–часов на 1 га хлопчатника, провести корреляционный анализ.

Таблица

Взаимосвязь между оплатой труда за единицу времени работы и затратами труда на единицу обрабатываемой площади

Y\X	500-650	650-800	800-950	950-1100	1100-1250	1250-1400	1400-1550	m_y
0,6-0,9					1	1	2	4
0,9-1,2			1	4	2	3	2	12
1,2-1,5		4	7	6	1	1		19
1,5-1,8	2	4	4	3	1			14
1,8-2,1	6	8	3	4				21
2,1-2,4	6	5	6	1				18

2,4-2,7	2	5						7
2,7-3,0	2	3						5
m_x	18	29	21	18	5	5	4	100

<u>Решение.</u> Будем считать, что приведенные в примере данные являются выборкой из двумерной нормально распределенной совокупности, на основании которой построен двумерный интервальный вариационный ряд (корреляционная таблица).

1) Используя формулы для условных вариантов, получим из вспомогательной таблицы корреляционного анализа значения величин, необходимых для вычисления точечных оценок основных параметров:

$x_i' = \dfrac{x_i - x_0}{h_x}$; $y_i' = \dfrac{y_i - y_0}{h_y}$, где h_x=150 – длина интервалов по x; h_y=0,3 – длина интервала по y; x_0, y_0 – рабочие средние выбираются обычно равными центрам интервалов, лежащих в середине соответствующих одномерных рядов. Тогда $x_0 = \dfrac{1100 + 950}{2} = 1025$, $y_0 = \dfrac{1{,}8 + 2{,}1}{2} = 1{,}95$.

2) Используя данные, рассчитаем из вспомогательной таблицы корреляционного анализа значения величин, необходимых для вычисления точечных оценок основных параметров: n=100.

а) $x_i' = \dfrac{x_i - x_0}{h_x}$

$$x_1' = \frac{575 - 1025}{150} = \frac{-450}{150} = -3$$

$$x_2' = \frac{725 - 1025}{150} = \frac{-300}{150} = -2$$

$$x_3' = \frac{875 - 1025}{150} = \frac{-150}{150} = -1$$

$$x_4' = \frac{1025 - 1025}{150} = \frac{0}{150} = 0$$

$$x_5' = \frac{1175 - 1025}{150} = \frac{150}{150} = 1$$

$$x_6' = \frac{1325 - 1025}{150} = \frac{300}{150} = 2$$

$$x_7' = \frac{1475 - 1025}{150} = \frac{450}{150} = 3$$

б) $y_i' = \dfrac{y_i - y_0}{h_y}$

$$y_1' = \frac{0{,}75 - 1{,}95}{0{,}3} = -4$$

$$y_2' = \frac{1{,}05 - 1{,}95}{0{,}3} = -3$$

$$y_3' = \frac{1{,}35 - 1{,}95}{0{,}3} = -2$$

$$y_4' = \frac{1{,}65 - 1{,}95}{0{,}3} = -1$$

$$y_5' = \frac{1{,}95 - 1{,}95}{0{,}3} = 0$$

$$y_6' = \frac{2{,}25 - 1{,}95}{0{,}3} = 1$$

$$y_7' = \frac{2{,}55 - 1{,}95}{0{,}3} = 2$$

$$y_8' = \frac{2{,}85 - 1{,}95}{0{,}3} = 3$$

â) $\Sigma x' m_x = 18 \cdot (-3) + 29 \cdot (-2) + 21 \cdot (-1) + 18 \cdot 0 + 5 \cdot 1 + 5 \cdot 2 + 4 \cdot 3 =$

$= -54 - 58 - 21 + 0 + 5 + 10 + 12 = -106$

ã) $\Sigma x'^2 m_x = (-3)^2 \cdot 18 + (-2)^2 \cdot 29 + (-1)^2 \cdot 21 + 0 \cdot 18 + 1^2 \cdot 5 + 2^2 \cdot 5 + 3^2 \cdot 4 =$

$= 162 + 116 + 21 + 0 + 5 + 20 + 36 = 360$

ä) $\Sigma x'y'm_{xy} = (-3)\cdot(-1)\cdot 2 + 0 + (-3)\cdot 1\cdot 6 + (-3)\cdot 2\cdot 2 + (-3)\cdot 3\cdot 2 +$

$+ (-2)\cdot(-2)\cdot 4 + (-2)\cdot(-1)\cdot 4 + 0 + (-2)\cdot 1\cdot 5 + (-2)\cdot 2\cdot 5 + (-2)\cdot 3\cdot 3 + (-1)\cdot(-3)\cdot 1 +$

$+ (-1)\cdot(-2)\cdot 7 + (-1)\cdot(-1)\cdot 4 + 0 + (-1)\cdot 1\cdot 6 + 0 + 1\cdot(-4)\cdot 1 + 1\cdot(-3)\cdot 2 +$

$+ 1\cdot(-2)\cdot 1 + 1\cdot(-1)\cdot 1 + 2\cdot(-4)\cdot 1 + 2\cdot(-3)\cdot 3 + 2\cdot(-2)\cdot 1 + 3\cdot(-4)\cdot 2 +$

$+ 3\cdot(-3)\cdot 2 = 6 - 18 - 12 - 18 + 16 + 8 - 10 - 20 - 18 + 3 + 14 + 4 - 6 - 4 -$

$- 6 - 2 - 1 - 8 - 18 - 4 - 24 - 18 = -187 + 51 = -136$

å) $\Sigma y'm_y = (-4)\cdot 4 + (-3)\cdot 12 + (-2)\cdot 19 + (-1)\cdot 14 + 0 + 1\cdot 18 + 2\cdot 7 + 3\cdot 5 =$

$= -16 - 36 - 38 - 14 + 0 + 18 + 14 + 15 = -57$

æ) $\Sigma y'^2 m_y = (-4)^2 \cdot 4 + (-3)^2 \cdot 12 + (-2)^2 \cdot 19 + (-1)^2 \cdot 14 + 0 + 1^2 \cdot 18 + 2^2 \cdot 7 + 3^2 \cdot 5 =$

$= 64 + 108 + 76 + 14 + 0 + 18 + 28 + 45 = 353$

3) После расчетов по пункту 2, будем иметь следующие точечные оценки параметров:

а) $\bar{x} = \frac{1}{n}\Sigma x'm_x h_x + x_0 = -\frac{106}{100}150 + 1025 = 866 \text{ чел.-ч/га} = 8{,}66 \text{ чел.-ч/га}$;

б) $\bar{y} = \frac{1}{n}\Sigma y'm_y h_y + y_0 = -\frac{57}{100}0{,}3 + 1{,}95 = 1{,}779 \text{ руб/чел-ч.}$;

в) $s_x^2 = \left[\frac{\Sigma x'^2 m_x}{n} - \left(\frac{\Sigma x'm_x}{n}\right)^2\right]\cdot h_x^2 = \left[\frac{360}{100} - \left(\frac{-106}{100}\right)^2\right]\cdot 150^2 = 55719$;

г) $s_y^2 = \left[\frac{\Sigma y'^2 m_y}{n} - \left(\frac{\Sigma y'm_y}{n}\right)^2\right]\cdot h_y^2 = \left[\frac{353}{100} - \left(\frac{-57}{100}\right)^2\right]\cdot 0{,}3^2 = 0{,}288459$;

д) $s_x = \sqrt{S_x^2} = \sqrt{55719} = 236{,}04972 \approx 2{,}36 \text{ чел.-ч/га}$

е) $s_y = \sqrt{S_y^2} = \sqrt{0{,}288459} = 0{,}5370834 \approx 0{,}54 \text{ тыс. руб/чел-ч.}$

$$\text{æ) } r = \frac{n\Sigma x'y'm_{xy} - \Sigma x'm_x \Sigma y'm_y}{\sqrt{\left[n\Sigma x'^2 m_x - (\Sigma x'm_x)^2\right]\cdot\left[n\Sigma y'^2 m_y - (\Sigma y'm_y)^2\right]}} =$$

$$= \frac{100(-136) - (-106)(-57)}{\left\{\left[100\cdot 360 - (-106)^2\right]\cdot\left[100\cdot 353 - (-57)^2\right]\right\}^{\frac{1}{2}}} =$$

$$= \frac{-19642}{(24764\cdot 32051)^{\frac{1}{2}}} = -\frac{19642}{28172{,}875} = -0{,}6972.$$

4) Далее проверим при уровне значимости α*=0,05* значимость выборочного коэффициента корреляции. По таблицам Фишера–Иейтса получаем $r_{кр}=r(0{,}05;98)<r(0{,}05;90)=0{,}205$. Критическая область имеет вид $|r| > r_{êð}$. Так как $r_{íàáë} = -0{,}6972$ попадает в критическую область, то нулевая гипотеза о

незначимости выборочного коэффициента корреляции отвергается с вероятностью ошибки 0,05. Следовательно, генеральный коэффициент корреляции ρ значимо отличается от нуля. Таким образом, можно считать доказанной связь между случайными величинами x и y.

5) Корреляционный анализ можно углубить:

а) Найдем с надежностью $\gamma = 1 - \alpha = 1 - 0{,}05 = 0{,}95$ интервальную оценку генерального коэффициента корреляции ρ. Получим значение статистики Z по формуле: $Z = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{0{,}3028}{1{,}6972} = -0{,}86183$.

б) Найдем точность интервальной оценки MZ по формуле $\Delta_Z = \Phi^{-1}(\gamma) : \sqrt{n-3} = 1{,}96 : \sqrt{97} = 0{,}19901$ (смотрим значение r=0,6972 по “Таблице Z-преобразования Фишера”).

Интервальная оценка для MZ имеет вид $Z - \Delta_Z \le MZ \le Z + \Delta_Z$, $-0{,}86138 - 0{,}199 \le MZ \le -0{,}86183 + 0{,}199$, т.е. $-1{,}06084 \le MZ \le -0{,}66282$.

С помощью обратной функции Z^{-1} получаем интервальную оценку коэффициента корреляции ρ: $Z^{-1}(Z - \Delta_Z) \le \rho \le Z^{-1}(Z + \Delta_Z)$ или $-0{,}7860 \le \rho \le -0{,}5802$.

в) Найдем точечные оценки коэффициентов регрессии с учетом измерения x в человеко-часах (в противном случае измерения окажутся неточными).

$$b_{yx} = r\frac{s_y}{s_x} = -0{,}6972\sqrt{\frac{0{,}288459}{5{,}5719}} = -0{,}1586344;$$

$$b_{xy} = r\frac{s_x}{s_y} = -0{,}6972\sqrt{\frac{5{,}5719}{0{,}288459}} = -3{,}064199.$$

Точечные оценки уравнений регрессии имеют вид

$\overline{y/x} - \bar{y} = b_{yx}(x - \bar{x})$ или $\overline{y/x} - 1{,}78 = -0{,}15863(x - 8{,}66)$ – регрессия y на x;

$\overline{x/y} - \bar{x} = b_{xy}(y - \bar{y})$ или $\overline{x/y} - 8{,}66 = -3{,}0642(y - 1{,}78)$ – регрессия x на y.

г) Далее можно получить интервальные оценки коэффициентов регрессии:

$$b_{yx} - St^{-1}(\alpha)\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}} \le \beta_{yx} \le b_{yx} + St^{-1}(\alpha)\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}};$$

$$b_{xy} - St^{-1}(\alpha)\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}} \le \beta_{xy} \le b_{xy} + St^{-1}(\alpha)\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}},$$

где $St^{-1}(\alpha) = St^{-1}(0{,}05; 98) = 1{,}967$ находится по таблицам распределения Стьюдента с помощью линейной интерполяции. Таким образом, интервальные оценки коэффициентов регрессии будут $-0{,}19104 \le \beta_{yx} \le -0{,}12623$; $-3{,}6902 \le \beta_{xy} \le -2{,}4382$.

6) Далее поясним содержательный смысл полученных результатов. Доказана значимость, существенность связи между оплатой 1 чел.-ч затрат в хлопководстве и затратах труда, приходящихся на 1 га площади хлопчатника. Связь является отрицательной ($r = -0{,}6972$), т.е. увеличению одного из показателей соответствует уменьшение среднего значения другого показателя.
При этом увеличению оплаты труда на 1 тыс. руб. за человеко-час соответствует уменьшение затрат на 1 га на величину 3,0642 чел.-ч в среднем. С вероятностью 0,95 можно утверждать, что указанная величина может оказаться равной любому значению из интервала [2,4382; 3,6902]. Увеличению затрат на одну сотню человеко-часов на 1 га будет соответствовать в среднем уменьшение оплаты за 1 чел.-ч на величину 0,15863 тыс.руб., причем с вероятностью 0,95 можно утверждать, что эта величина может принимать любое значение из интервала [0,12623; 0,19104]. Так как коэффициент детерминации ρ^2 при доверительной вероятности γ=0,95 лежит в границах от $(-0{,}5802)^2=0{,}3366$ до $(-0{,}7860)^2=0{,}6178$, то можно сказать, что изменение одного из контролируемых в модели показателей соответствует изменению другого на величину, принимающую значение от 33,66% до 61,78%. Далее в таблице представлен корреляционный анализ двумерной модели.

Вычислительная таблица корреляционного анализа двумерной модели.

y’\x’	-3	-2	-1	0	1	2	3	m_y	$y'm_y$	y'^2m_y
-4					1	1	2	4	-16	64
-3			1	4	2	3	2	12	-36	108
-2		4	7	6	1	1		19	-38	76
-1	2	4	4	3	1			14	-14	14
0	6	8	3	4				21	0	0
1	6	5	6	1				18	18	18
2	2	5						7	14	28
3	2	3						5	15	45
m_x	18	29	21	18	5	5	4	100	-57	353
$x'm_x$	-54	-58	-21	0	5	10	12	-106		
x'^2m_y	162	116	21	0	5	20	36	360		
$\Sigma\ y'm_{xy}$	14	12	-15	-26	-13	-15	-14	-57		
$x'\Sigma\ y'm_{xy}$	-42	-24	15	0	-13	-30	-42	-136		

Трехмерная модель

Для изучения основных задач и особенностей корреляционного анализа удобно рассматривать генеральную совокупность трех признаков *X*, *Y* и *Z*.

Трехмерная непрерывная случайная величина (*X*, *Y*, *Z*) называется нормально распределенной, если плотность совместного распределения однородных случайных величин *X*, *Y*, и *Z* задается в виде

$$\rho(x,y,z) = (2\pi)^{-\frac{3}{2}}\left[\sigma_x^2\sigma_y^2\sigma_z^2\left|\Re_3\right|\right]^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}u^T\Re_3^{-1}u\right\}, \qquad (22)$$

где $\Re_3 = \begin{pmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & 1 & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & 1 \end{pmatrix}$ – симметрическая положительно определенная матрица парных коэффициентов корреляции, соответствующих частным двумерным распределениям случайных величин (*X*, *Y*), (*X*, *Z*) и (*Y*, *Z*); $\left|\Re_3\right|$ - определитель матрицы $\Re_3$, обобщенная дисперсия случайной величины (*X, Y, Z*); $\left|\Re_3\right| = 1 + 2\rho_{xy}\rho_{xz}\rho_{yz} - \rho_{xz}^2 - \rho_{yz}^2 - \rho_{xy}^2 > 0$; $\left|\Re_3\right| \neq 0$ (по условию существования обратной матрицы).

$\left|\Re_3^{-1}\right| = \begin{pmatrix} \rho^{(11)} & \rho^{(12)} & \rho^{(13)} \\ \rho^{(21)} & \rho^{(22)} & \rho^{(23)} \\ \rho^{(31)} & \rho^{(32)} & \rho^{(33)} \end{pmatrix}$ – матрица, обратная $\Re_3$.

Таким образом, трехмерная нормально распределенная случайная величина определяется девятью параметрами:

1) тремя математическими ожиданиями:

а) $Mx = \mu_x$, б) $My = \mu_y$, в) $Mz = \mu_z$;

2) тремя дисперсиями (или средними квадратическими отклонениями):

а) $Dx = \sigma_x^2$, б) $Dy = \sigma_y^2$, в) $Dz = \sigma_z^2$ $(\sigma_x, \sigma_y, \sigma_z)$;

3) тремя парными коэффициентами корреляции:

а) $\rho_{xy} = M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}\right]$,

б) $\rho_{xz} = M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{z-\mu_z}{\sigma_z}\right]$,

в) $\rho_{yz} = M\left[\frac{y-\mu_y}{\sigma_y}\cdot\frac{z-\mu_z}{\sigma_z}\right]$.

Пример 6. Корреляционный анализ показателей эффективности работы предприятий.

С целью анализа взаимосвязи показателей эффективности производства продукции: производительности труда, фондоотдачи и материалоемкости продукции была отобрана группа 25 однотипных машиностроительных предприятий.

На основании годовых отчетов предприятий были получены следующие данные (таблица 1): x – выработка валовой продукции в неизменных ценах на одного работающего средней списочной численности ППП (млн. руб.); y – выпуск валовой продукции на один рубль среднегодовой стоимости основных промышленно-производственных фондов (руб.); z – материалоемкость в стоимостном выражении: стоимость материалов в валовой продукции в неизменных ценах (%).

Таблица 1

Исходные данные

Номер предприятия	x	y	z	Номер предприятия	x	y	z
1	6,0	2,0	25	14	5,7	2,2	25
2	4,9	0,8	30	15	5,1	1,3	30
3	7,0	2,7	20	16	5,2	1,5	14
4	6,7	3,0	21	17	7,3	2,7	20
5	5,8	1,0	28	18	6,1	2,4	27
6	6,1	2,1	26	19	6,2	2,2	28
7	5,0	0,9	30	20	5,9	2,0	26
8	6,9	2,6	22	21	6,0	2,0	26
9	6,8	3,0	20	22	4,8	0,9	31
10	5,9	1,1	29	23	7,3	3,2	19
11	5,0	0,8	27	24	7,2	3,3	20
12	5,6	2,2	25	25	7,0	3,0	20
13	6,0	2,4	24				

Предположим, что рассматриваемые признаки x, y и z в генеральной совокупности подчиняются нормальному закону распределения и указанные данные представляют выборку из этой совокупности.

Для получения точечных оценок генеральных средних, дисперсий, средних квадратических отклонений и парных коэффициентов корреляции результаты промежуточных вычислений удобно поместить в расчетную таблицу (таблица 2). С целью контроля вычислений данные разбиты на пятерки. Для каждой пятерки в итоговой таблице (таблица 3) в последней строке приведены суммы элементов каждого столбца. В последнем контрольном столбце

приводятся суммы элементов соответствующих строк. Сумма пяти элементов контрольного столбца должна совпадать с суммой элементов итоговой строки. Например, 6,0 + 4,9 + 7,0 + 6,7 + 5,8 = 30,4;
6,0 + 2,0 + 25,0 + 36,0 + 4,0 + 625,0 + 12,0 + 150,0 + 50,0 =910,00
Контроль:
30,4 + 9,5 + 124,0 + 187,54 + 21,93 + 3150,0 + 60,72 + 740,1 + 219,0 = 910,00 + +1135,27 + 698,89 + 749,39 + 1049,64 =4543,19.

Далее получаем таблицу итоговых строк (таблица 3). Из последней строки итоговой таблицы получаем

$\Sigma x = 151{,}5$; $\Sigma y = 51{,}3$; $\Sigma z = 623{,}0$; $\Sigma x^2 = 933{,}23$; $\Sigma y^2 = 120{,}97$; $\Sigma z^2 = 15869{,}0$

$\Sigma xy = 324{,}78$; $\Sigma xz = 3712{,}1$; $\Sigma yz = 1213{,}0$.

Следовательно,

$$\bar{x} = \frac{1}{25}\cdot 151{,}5 = 6{,}06; \quad \overline{x^2} = \frac{1}{25}\cdot 933{,}23 = 37{,}3292;$$

$$\bar{y} = \frac{1}{25}\cdot 51{,}3 = 2{,}052; \quad \overline{y^2} = \frac{1}{25}\cdot 120{,}97 = 4{,}8388;$$

$$\bar{z} = \frac{1}{25}\cdot 623 = 24{,}92; \quad \overline{z^2} = \frac{1}{25}\cdot 15869 = 634{,}76;$$

$$s_x^2 = \overline{x^2} - (\bar{x})^2 = 37{,}3292 - (6{,}06)^2 = 0{,}6056; \quad s_x = 0{,}778203;$$

$$s_y^2 = \overline{y^2} - (\bar{y})^2 = 4{,}8388 - (2{,}052)^2 = 0{,}628096; \quad s_y = 0{,}792525;$$

$$s_z^2 = \overline{z^2} - (\bar{z})^2 = 634{,}76 - (24{,}92)^2 = 13{,}7536; \quad s_z = 3{,}7085846.$$

Таблица 2

Расчетная таблица

X	Y	Z	X^2	Y^2	Z^2	XY	XZ	YZ	Контроль
6,0	2,0	25	36,00	4,00	625,00	12,00	150,00	50,0	910,00
4,9	0,8	30	24,01	0,64	900,00	3,92	147,00	24,0	1135,27
7,0	2,7	20	49,00	7,29	400,00	18,90	140,00	54,0	698,89
6,7	3,0	21	44,89	9,00	441,00	20,10	140,7	63,0	749,39
5,8	1,0	28	33,64	1,00	784,00	5,80	162,4	28,0	1049,64
30,4	**9,5**	**124**	**187,54**	**21,93**	**3150**	**60,72**	**740,1**	**219,0**	**4543,19**
6,1	2,1	26	37,21	4,41	676,00	12,81	158,6	54,6	977,83
5,0	0,9	30	25,00	0,81	900,00	4,50	150,0	27,0	1143,21
6,9	2,6	22	47,61	6,76	484,00	17,94	151,8	57,2	796,81
6,8	3,0	20	46,24	9,00	400,00	20,40	136,0	60,0	701,44
5,9	1,1	29	34,81	1,21	841,00	6,49	171,1	31,9	1122,51
30,7	**9,7**	**127**	**190,87**	**22,19**	**3301**	**62,14**	**767,5**	**230,7**	**4741,80**
5,0	0,8	27	25,00	0,64	729,00	4,00	135,0	21,6	948,04
5,6	2,2	25	31,36	4,84	625,00	12,32	140,0	55,0	301,32
6,0	2,4	24	36,00	5,76	576,00	14,40	144,0	57,6	866,16
5,7	2,2	25	32,49	4,84	625,00	12,54	142,5	55,0	905,27

5,1	1,3	30	26,01	1,69	900,00	6,63	153,0	39,0	1162,73
27,4	**8,9**	**131**	**150,86**	**17,77**	**3455**	**49,89**	**714,5**	**228,2**	**4783,52**
5,2	1,5	24	27,04	2,25	576,00	7,80	124,8	36,0	804,59
7,3	2,7	20	53,29	7,29	400,00	19,71	146,0	54,0	710,29
6,1	2,4	27	37,21	5,76	729,00	14,64	164,7	64,8	1051,61
6,2	2,2	28	38,44	4,84	784,00	13,64	173,6	61,6	1112,52
5,9	2,0	26	34,81	4,00	676,00	11,80	153,4	52,0	965,91
30,7	**10,8**	**125**	**190,79**	**24,14**	**3165**	**67,59**	**762,5**	**268,4**	**4644,92**
6,0	2,0	26	36,00	4,00	676,00	12,00	156,0	52,0	970,00
4,8	0,9	31	23,04	0,81	961,00	4,32	148,8	27,9	1202,57
7,3	3,2	19	53,29	10,24	361,00	23,36	138,7	60,8	676,89
7,2	3,3	20	51,84	10,89	400,00	23,76	144,0	66,0	726,99
7,0	3,0	20	43,00	9,00	400,00	21,00	140,0	60,0	709,00
32,3	**12,4**	**116**	**213,17**	**34,94**	**2798**	**84,44**	**727,5**	**266,7**	**4285,45**

Таблица 3

Итоговая таблица

1	2	3	4	5	6	7	8	9	10
30,4	9,5	124,0	187,54	21,93	3150	60,72	740,1	219,0	4543,19
30,7	9,7	127,0	190,87	22,19	3301	62,14	767,5	230,7	4741,8
27,4	8,9	131,0	150,86	17,77	3455	49,89	714,5	228,2	4783,52
30,7	10,8	125,0	190,79	24,14	3165	67,59	762,5	268,4	4644,92
32,3	12,4	116,0	213,17	34,94	2798	84,44	727,5	266,7	4285,45
151,5	**51,3**	**623,0**	**933,23**	**120,97**	**15869**	**324,8**	**3712,1**	**1213,0**	**22798,9**

Для вычисления точечных оценок парных коэффициентов корреляции используем формулу

$$r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} \qquad (23)$$

$$\overline{xy} = \frac{324,78}{25} = 12,9912; \quad \overline{xz} = \frac{3712,1}{25} = 148,484; \quad \overline{yz} = \frac{1213,0}{25} = 48,52$$

$$r_{xy} = \frac{12,9912 - 6,06\cdot 2,052}{0,778203\cdot 0,792525} = \frac{0,55608}{0,61674} = 0,9016;$$

$$r_{xz} = \frac{148,484 - 6,06\cdot 24,92}{0,778203\cdot 3,7085846} = \frac{-2,5312}{2,8860} = -0,87706;$$

$$r_{yz} = \frac{48,52 - 2,052\cdot 24,92}{0,792525\cdot 3,7085846} = \frac{-2,61584}{2,9391} = -0,8900.$$

Вычисляем точечные оценки условных средних квадратических отклонений (при одной фиксированной переменной):

$$s_{x/y} = \sqrt{s_x^2\cdot(1 - r_{xy}^2)} = \sqrt{0,6056\cdot(1 - 0,9016362^2)} = 0,3365691; \qquad (24)$$

$$s_{y/x} = \sqrt{s_y^2 \cdot (1 - r_{xy}^2)} = \sqrt{0{,}628096 \cdot (1 - 0{,}9016362^2)} = 0{,}3427633; \qquad (25)$$

$s_{x/z} = 0{,}3738452$; $s_{z/x} = 1{,}7815879$; $s_{y/z} = 0{,}3613602$; $s_{z/y} = 1{,}6909689$.

Получаем точечные оценки частных коэффициентов корреляции:

$$r_{xy/z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}} = \frac{0{,}9016362 - (-0{,}8770519) \cdot (-0{,}8899999)}{\sqrt{0{,}23078 \cdot 0{,}2079002}} = 0{,}5526811; \qquad (26)$$

$r_{xz/y} = \dfrac{-0{,}8770519 + 0{,}9016362 \cdot 0{,}8899999}{\sqrt{0{,}1870552 \cdot 0{,}2079002}} = -0{,}3782736$ (далее будет показано, что он незначим);

$$r_{yz/x} = \frac{-0{,}8899999 + 0{,}9016362 \cdot 0{,}8770519}{\sqrt{0{,}1870552 \cdot 0{,}23078}} = -0{,}4775413.$$

Вычисляем точечные оценки остаточных дисперсий (при двух фиксированных переменных):

$$s_{x/yz}^2 = s_{x/z}^2 \cdot (1 - r_{xy/z}^2) = 0{,}3738452^2 \cdot (1 - 0{,}5526811^2) = 0{,}0970695; \qquad (27)$$

$$s_{y/xz}^2 = s_{y/z}^2 \cdot (1 - r_{xy/z}^2) = 0{,}0906942; \quad s_{z/xy}^2 = s_{z/x}^2 \cdot (1 - r_{yz/x}^2) = 2{,}450226. \qquad (28)$$

Получаем оценки множественных коэффициентов детерминации и корреляции:

$$r_x^2 = 1 - \frac{s_{x/yz}^2}{s_x^2} = 1 - \frac{0{,}0970695}{0{,}6056} = 0{,}8397136; \quad r_x = 0{,}9163588; \qquad (29)$$

$$r_y^2 = 1 - \frac{s_{y/xz}^2}{s_y^2} = 1 - \frac{0{,}906942}{0{,}628096} = 0{,}8556046; \quad r_y = 0{,}9249889; \qquad (30)$$

$$r_z^2 = 1 - \frac{s_{z/xy}^2}{s_z^2} = 1 - \frac{2{,}450226}{13{,}7536} = 0{,}8218484; \quad r_z = 0{,}9065585. \qquad (31)$$

Для контроля вычислений полезно воспользоваться другими формулами точечных оценок множественных коэффициентов корреляции, например (см. Гмурман В.Е. глава 18, §15)

$$r_x = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2 r_{xy} r_{xz} r_{yz}}{1 - r_{yz}^2}} = \sqrt{\frac{0{,}1745766}{0{,}2079002}} = \sqrt{0{,}8397134} = 0{,}9163587. \qquad (32)$$

Ошибка в последнем разряде на единицу является допустимой.

Проверим с уровнем значимости α =0,05 значимость множественных коэффициентов корреляции. Вычислим наблюдаемые значения F–критерия:

$$F_{набл}(r_x^2) = \frac{r_x^2/2}{(1 - r_x^2)/(n-3)} = \frac{0{,}8397136 \cdot 22}{2 \cdot 0{,}1602864} = 57{,}627157; \qquad (33)$$

$$F_{набл}(r_y^2) = \frac{0{,}8556046 \cdot 22}{2 \cdot 0{,}1443954} = 65{,}179712;$$

$$F_{набл}(r_z^2) = \frac{0{,}8218484 \cdot 22}{2 \cdot 0{,}1781516} = 50{,}745165.$$

Находим по таблице F–распределения критическое значение F–статистик для уровня значимости α =0,05, числа степеней свободы числителя $\nu_1 = 2$ и знаменателя $\nu_2 = n - 3 = 22$: $F_{кр}(0{,}05;2;22) = 3{,}44$.

Так как все наблюдаемые значения F–статистик превосходят ее критическое значение, то гипотеза о равенстве нулю каждого множественного коэффициента корреляции генеральной совокупности отвергается с вероятностью ошибки, равной 0,05. Следовательно, эти коэффициенты значимо отличаются от нуля.

Далее проверяют значимость частных коэффициентов корреляции с уровнем значимости α =0,05. По таблице r–статистики для числа степеней свободы $\nu = n - 3 = 25 - 3 = 22$ и α =0,05 находим $r_{кр}(0{,}05;25) < r_{кр}(0{,}05;22) < r_{кр}(0{,}05;20)$ или $0{,}381 < r_{кр}(0{,}05;22) < 0{,}423$. Так как наблюдаемые значения $|r|$, $|r_{xy/z}|$, $|r_{yz/x}|$ больше 0,423, то они будут превосходить и $r_{кр}(0{,}05;22)$. Следовательно, гипотеза о равенстве нулю генеральных частных коэффициентов корреляции $\rho_{xy/z}$ и $\rho_{yz/x}$ отвергается с вероятностью ошибки 0,05. Наблюдаемое значение r–статистики для частного коэффициента корреляции между x и z меньше 0,381, т.е. подавно меньше, чем $r_{кр}(0{,}05;22)$. Следовательно, гипотеза H_0: $\rho_{xy/z} = 0$ не отвергается. Примем эту гипотезу, т.е. будем считать, что генеральный частный коэффициент корреляции между x и z равен нулю, т.е. статистически незначим.

Таким образом, значимыми оказались только два частных коэффициента корреляции $\rho_{xy/z}$ и $\rho_{yz/x}$.

Нужно отметить, что в случае необходимости можно было бы вычислить $r_{кр}(0{,}05;22)$ с помощью линейной интерполяции $r_{кр}(0{,}05;22) = 0{,}381 + \dfrac{0{,}423 - 0{,}381}{25 - 20} \cdot (25 - 20) = 0{,}4062$.

С надежностью $\gamma = 1 - \alpha = 0{,}95$ найдем интервальные оценки значимых коэффициентов корреляции (кроме r_{xz}, т.к. он незначим). Воспользуемся таблицей функции Z–преобразования Фишера. Для $\rho_{xy/z}$ получаем по таблице функции $Z_r = \dfrac{1}{2} \cdot \ln\dfrac{1+r}{1-r}$ значение $Z_{0,55} = 0{,}6184$. Затем вычисляем точность интервальной оценки для MZ $\Delta Z = t_{0,95} \cdot \dfrac{1}{\sqrt{n-4}} = 1{,}96 \cdot \dfrac{1}{25-1} = 0{,}4277$ ($t_{0,95}$ находится по таблице интеграла Лапласа). Интервальная оценка для MZ есть: $0{,}6184 - 0{,}4277 \le MZ \le 0{,}6184 + 0{,}4277$ или $0{,}1907 \le MZ \le 1{,}0461$.

Чтобы получить интервальную оценку для $\rho_{xy/z}$ по таблицам Z_r находим числа, ближайшие к 0,1907 и к 1,0461. Это будут числа $Z_r(\min) = 0{,}1923$ и $Z_r(\max) = 1{,}0454$. Чтобы перейти к аргументу r, надо выполнить обратное преобразование, т.е. из $Z_r = \frac{1}{2}\cdot\ln\frac{1+r}{1-r}$ надо получить r. Тогда имеем $0{,}19 \le \rho_{xy/z} \le 0{,}78$.

Для $\rho_{yz/x}$ будем иметь $Z_{-0,48} = -Z_{0,48} = -0{,}5230$ (с учетом того, что функция Z_r нечетная), тогда $-0{,}5230 - 0{,}4277 \le MZ \le -0{,}5230 + 0{,}4277$ или $-{,}09507 \le MZ \le 0{,}0953$. Ближайшими числами к числам 0,0953 и 0,9507 будут числа 0,1003 и 0,9505. Следовательно, интервальная оценка для $\rho_{yz/x}$ есть $-0{,}74 \le \rho_{yz/x} \le -0{,}10$.

Для изучаемых трех признаков нет односторонней причинно-следственной связи. Поэтому в качестве модели взаимозависимости можно выбрать наиболее надежную в условиях данной выборки. Значимым множественным коэффициентом детерминации, имеющим наибольшую оценку, является $\rho_y^2 = 0{,}9249889$. Ему соответствуют оба значимых частных коэффициента детерминации $\rho_{xy/z}^2$ и $\rho_{yz/x}^2$, имеющих и оценки больше, чем $\rho_{xz/y}^2$.

Оценка соответствующего уравнения регрессии имеет вид

$$\overline{y/(x,z)} - \bar{y} = b_{yx/z}(x - \bar{x}) + b_{yz/x}(z - \bar{z}). \tag{34}$$

Вычислим коэффициенты множественной (частной) регрессии:

$$b_{yx/z} = r_{xy/z}\cdot\frac{s_{y/z}}{s_{x/z}} = 0{,}5526811\cdot\frac{0{,}3613601}{0{,}3738452} = 0{,}5342236; \tag{35}$$

$$b_{yz/x} = r_{yz/x}\cdot\frac{s_{y/x}}{s_{z/x}} = -0{,}4775413\cdot\frac{0{,}3427633}{1{,}7815879} = -0{,}0918751. \tag{36}$$

Таким образом, получаем $\overline{y/(x,z)} - 2{,}052 = 0{,}5342(x - 6{,}06) - 0{,}091875(z - 24{,}92)$.

Далее найдем с надежностью $\gamma = 1 - \alpha = 0{,}95$ интервальные оценки для множественных коэффициентов регрессии. По таблице распределения Стьюдента находим $t(\alpha, n-3) = t(0{,}05;22) = 2{,}074$ и решим относительно β неравенства (исходя из определения доверительных интервалов):

$$\left|\frac{(b_{yx/z} - \beta_{yx/z})\cdot s_{x/z}\cdot\sqrt{n-3}}{s_{y/z}\cdot\sqrt{1 - r_{xy/z}^2}}\right| \le t(\alpha; n-3); \quad \left|\frac{(0{,}5342236 - \beta_{yx/z})\cdot 0{,}3738452\cdot\sqrt{22}}{0{,}3613602\cdot\sqrt{1 - 0{,}5526811^2}}\right| \le 2{,}074$$

$$\left|\frac{(b_{yz/x} - \beta_{yz/x})\cdot s_{z/x}\cdot\sqrt{n-3}}{s_{y/x}\cdot\sqrt{1 - r_{yz/x}^2}}\right| \le t(\alpha; n-3); \quad \left|\frac{(-0{,}0918751 - \beta_{yz/x})\cdot 1{,}7815879\cdot\sqrt{22}}{0{,}3427633\cdot\sqrt{1 - 0{,}4775413^2}}\right| \le 2{,}074$$

Вычислим точности интервальных оценок

$$\Delta b_{yx/z} = t\cdot\frac{s_{y/z}\cdot\sqrt{1 - r_{xy/z}^2}}{s_{x/z}\cdot\sqrt{n-3}}; \quad \Delta b_{yx/z} = 2{,}074\cdot\frac{0{,}3613602\sqrt{1 - 0{,}5526811^2}}{0{,}3738452\cdot\sqrt{22}} = 0{,}3562008$$

$$\Delta b_{yz/x} = t \cdot \frac{s_{y/x} \cdot \sqrt{1 - r_{yz/x}^2}}{s_{z/x} \cdot \sqrt{n-3}}\ ; \quad \Delta b_{yz/x} = 2{,}074 \cdot \frac{0{,}3427633 \cdot \sqrt{1 - 0{,}4775413^2}}{1{,}7815879 \cdot \sqrt{22}} = 0{,}0747445 .$$

Откуда $0{,}5342236 - 0{,}3562008 \le \beta_{yx/z} \le 0{,}5342236 + 0{,}3562008$

$-0{,}0918751 - 0{,}0747445 \le \beta_{yz/x} \le -0{,}0918751 + 0{,}0747445$ или

$0{,}1780 \le \beta_{yx/z} \le 0{,}8904$ и $-0{,}0171306 \le \beta_{yz/x} \le -0{,}1666196$.

Получим предпочтительные точечные оценки изучаемых коэффициентов детерминации:

$$\hat{\rho}^2_{xy/z} = \frac{(n-2) \cdot r_{yx/z}^2 - 1}{n-3} = \frac{23 \cdot 0{,}5526811^2 - 1}{22} = 0{,}2738861 \qquad (37)$$

$$\hat{\rho}^2_{yz/x} = \frac{23 \cdot 0{,}4775413^2 - 1}{22} = 0{,}1929567 \qquad (38)$$

$$\hat{\rho}^2_{y} = \frac{(n-1) \cdot r_y^2 - 2}{n-3} = \frac{24 \cdot 0{,}8556046 - 2}{22} = 0{,}8424777 . \qquad (39)$$

На основании полученных расчетов можно сделать следующие выводы. Доказана тесная взаимосвязь каждого из исследуемых показателей эффективности работы предприятия с другими (все множественные коэффициенты детерминации значимы и превышают 0,8). Особенно тесная связь существует между фондоотдачей и двумя остальными показателями – производительностью труда и материалоемкостью. Изменение фондоотдачи в среднем на 84,25% объясняется изменением производительности труда и материалоемкости (изменение фондоотдачи в среднем на 15,75% объясняется влиянием неконтролируемых факторов, признаков). При этом при увеличении производительности труда на 1 руб. фондоотдача увеличивается в среднем на 0,55 руб. на рубль основных производственных фондов. При уменьшении материалоемкости на 1% фондоотдача увеличивается в среднем на 0,48%. Указанные нормативы относительно стабильны при условии, что изучаемые показатели отклоняются на небольшие величины от своих средних уровней (стабильность указывается доверительными интервалами и надежностью 0,95). Взаимозависимость между материалоемкостью и производительностью труда (без учета фондоотдачи) не доказана (частный коэффициент корреляции $\rho_{xz/y}$ незначим) при данных условиях. Для более надежной проверки такой зависимости необходим большой объем выборки.

Варианты заданий и исходные данные для самостоятельной работы по корреляционному анализу

Вариант 1. По отчетным данным n=6 машиностроительных предприятий, представленных в таблице, провести корреляционный анализ взаимосвязи следующих показателей эффективности: x–фондоотдача активной части на 1 руб. ОПФ (руб.); y–рентабельность (%); z– производительность труда (млн руб/чел.).

Номер предприятия	x	y	z
1	2,0	13,2	9,4
2	1,2	12,9	6,7
3	2,6	17,2	10,0
4	1,8	17,5	5,2
5	1,4	8,0	5,7
6	2,3	14,2	9,4

Вариант 2. По отчетным данным n=6 машиностроительных предприятий, представленных в таблице, провести корреляционный анализ взаимосвязи следующих показателей эффективности: x–фондоотдача активной части на 1 руб. ОПФ (руб.); y–рентабельность (%); z– производительность труда (млн руб/чел.).

Номер предприятия	x	y	z
1	0,7	6,6	5,5
2	1,3	19,1	6,6
3	1,1	9,9	7,4
4	0,6	5,4	4,3
5	0,9	9,1	6,6
6	1,0	9,7	5,5

Вариант 3. На основании годовых отчетных данных пяти строительно-монтажных предприятий были получены значения следующих показателей. Провести корреляционный анализ взаимосвязи следующих показателей эффективности: x– объем выполненных работ, млрд.руб.; y– численность рабочих, чел.; z– фонд заработной платы, млрд.руб.:

Номер предприятия	Объем выполненных работ, млрд.руб.	Численность рабочих, чел.	Фонд заработной платы, млрд.руб.
1	13	320	3,2
2	14	570	5,5
3	16	780	8,0
4	12	200	2,5
5	15	700	7,2

Вариант 4. По отчетным данным n=7 машиностроительных предприятий провести корреляционный анализ взаимосвязи следующих показателей эффективности: x–производительность труда (млн руб/чел.); y–индекс снижения себестоимости продукци (%); z–рентабельность (%).

Номер предприятия	x	y	z
1	9,4	62	11
2	9,9	53	9
3	9,1	56	23
4	5,6	30	10
5	6,7	18	9
6	4,3	14	5
7	7,4	90	10

Вариант 5. По отчетным данным n=7 машиностроительных предприятий провести корреляционный анализ взаимосвязи следующих показателей эффективности: x–производительность труда (млн руб/чел.); y–индекс снижения себестоимости продукци (%); z–рентабельность (%).

Номер предприятия	x	y	z
1	6,6	77	19
2	5,5	32	7
3	9,4	200	14
4	5,7	91	8
5	5,2	82	18
6	10,0	76	17
7	6,7	37	13

Вариант 6. По данным годовой отчетности n=30 угольных шахт найдены следующие выборочные характеристики

$\bar{x}_1 = 34{,}3$; $s_1 = 12{,}9$; $r_{12} = -0{,}427$;

$\bar{x}_2 = 0{,}02$; $s_2 = 0{,}0053$; $r_{13} = -0{,}750$;

$\bar{x}_3 = 568{,}8$; $s_3 = 167{,}2$; $r_{23} = 0{,}102$.

Требуется провести корреляционный анализ, если x_1 – среднемесячная производительность труда по добыче угля (т); x_2 – фондоотдача; x_3 – трудоемкость работ по добыче (человеко–дни на 1000 т).

Вариант 7. По результатам n=100 наблюдений найдены выборочные характеристики трехмерной генеральной совокупности:

$\bar{x} = 4$; $s_x = 2$; $r_{xy} = -0{,}6$;

$\bar{y} = 5$; $s_y = 2$; $r_{xz} = 0{,}8$;

$\bar{z} = 7$; $s_z = 3$; $r_{yz} = -0{,}6$.

Провести корреляционный анализ.

Вариант 8. Для вычисления взаимозависимости между себестоимостью 1т песка (z), сменной добычей песка (y) и фондоотдачей (x) было обследовано восемь карьеров. В результате получены следующие данные:

X	30	20	40	35	45	25	50	30
Y, т	20	30	50	70	80	20	90	25
Z, тыс.руб.	20	25	20	15	10	30	10	20

Требуется провести корреляционный анализ.

Вариант 9. На основании данных годовых отчетов n=20 цементных заводов вычислены:

$$\bar{x}_1 = 35{,}4; \quad s_1 = 12{,}1; \quad r_{12} = 0{,}429;$$
$$\bar{x}_2 = 1417; \quad s_2 = 633; \quad r_{13} = -0{,}520;$$
$$\bar{x}_3 = 12{,}5; \quad s_3 = 3{,}2; \quad r_{23} = -0{,}706.$$

Требуется провести корреляционный анализ, если x_1 – фондоотдача на 1 руб. ОПФ (кг); x_2 – выработка натурального цемента на одного работающего (т); x_3 – среднезаводская себестоимость 1т цемента (тыс. руб.).

Вариант 10. По данным годовых отчетности n=30 угольных шахт найдены следующие выборочные характеристики

$$\bar{x}_1 = 34{,}3; \quad s_1 = 12{,}9; \quad r_{12} = -0{,}229;$$
$$\bar{x}_2 = 59{,}5; \quad s_2 = 21{,}8; \quad r_{13} = -0{,}750;$$
$$\bar{x}_3 = 568{,}8; \quad s_3 = 167{,}2; \quad r_{23} = 0{,}361.$$

Провести корреляционный анализ, если x_1 – среднемесячная производительность труда рабочего по добыче угля (т); x_2 – фондоемкость (тыс. руб./т); x_3 – трудоемкость работ по добыче (человеко–дни на 1000 т).

Вариант 11. По данным n=7 цементных заводов, представленных в таблице, провести корреляционный анализ взаимосвязи следующих показателей эффективности: x–фондоотдачи; y–выработки натурального цемента на одного работающего (млн.т); и z–среднезаводской себестоимости 1т цемента.

Номер предприятия	x	y	z
1	26,2	1,2	19,0
2	38,4	1,6	10,3
3	31,6	1,1	11,9
4	42,7	2,2	10,5
5	32,9	1,7	11,6
6	58,2	1,6	7,8
7	44,8	0,9	9,7

Вариант 12. По данным n=7 цементных заводов, представленных в таблице, провести корреляционный анализ взаимосвязи следующих показателей эффективности: x–фондоотдачи; y–выработки натурального цемента на одного работающего (млн.т); и z–среднезаводской себестоимости 1т цемента.

Номер предприятия	x	y	z
1	35,2	1,7	11,9
2	62,5	1,8	10,9
3	29,2	1,2	12,8
4	38,7	1,5	9,2
5	56,7	1,8	6,7
6	23,1	0,7	13,2
7	48,4	1,9	9,3

ГЛОССАРИЙ

- **Бесповторная выборка** – это выборка, при которой отобранный объект в генеральную совокупность не возвращается.
- **Вероятность** – количественная мера объективной возможности появления события в условиях данного эксперимента.
- **Вероятность доверительная** – вероятность, признаваемая достаточной для суждения о достоверности характеристик, полученных на основе выборочных наблюдений.
- **Вероятность события классическая** – отношение числа исходов эксперимента, благоприятствующих появлению события, к общему числу исходов.
- **Двусторонняя критическая область –** это область, определяемая как $(-\infty, k_{1-\alpha/2}) \cup (k_{\alpha/2}, +\infty)$. Она определяется в случае, когда альтернативная гипотеза имеет вид: $H_1: \Theta \neq \Theta_0$.
- **Дисперсия случайной величины** – математическое ожидание квадрата отклонений случайной величины от ее математического ожидания.
- **Интервальная оценка** – числовой интервал, относительно которого с вероятностью, близкой к единице можно утверждать, что оцениваемый параметр находится внутри него.
- **Конкурирующая (альтернативная) гипотеза** – гипотеза, противоположная нулевой, которая будет верна в том случае, если нулевая гипотеза противоречит опытным данным.
- **Корреляционная зависимость** – зависимость математического ожидания одной случайной величины от вариации других.
- **Критическая область** – подмножество значений выборочной характеристики, составляющей основу статистического критерия, при которых нулевая гипотеза отвергается.
- **Критические точки –** это точки, разделяющие критическую область и область принятия гипотезы.
- **Левосторонняя критическая область** – это область, определяемая как $(-\infty, k_{1-\alpha})$. Она используется в случае, когда альтернативная гипотеза имеет вид: $H_1: \Theta < \Theta_0$.
- **Математическое ожидание** – средняя величина возможных значений случайной величины, взвешенных по их вероятности.
- **Механический отбор** – это отбор, при котором генеральную совокупность «механически» делят на столько групп, сколько объектов должно войти в выборку, а из каждой группы выбирают один объект.
- **Мощность критерия** – вероятность не совершить ошибку второго рода, отвергнуть ложную нулевую гипотезу.
- **Несмещенная оценка** – точечная оценка параметра, математическое ожидание которой равно самому параметру.

- **Нулевая гипотеза** – статистическая гипотеза, которую необходимо проверить.
- **Правосторонняя критическая область** – это область, определяемая как (k_{α} , + ∞). Она используется в случае, когда альтернативная гипотеза имеет вид: H_1:Θ>Θ_0.
- **Повторная выборка** – это выборка, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.
- **Простой случайный отбор** – это отбор, при котором объекты извлекают по одному из всей генеральной совокупности.
- **Серийный отбор** – это отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию.
- **Состоятельная оценка** – точечная оценка, которая сходится по вероятности к оцениваемому параметру.
- **Статистические выводы –** это заключения о генеральной совокупности (т.е. о законе распределения исследуемой случайной величины и его параметрах либо о наличии и силе связи между исследуемыми переменными) на основе выборки, случайно отобранной из генеральной совокупности.
- **Статистическая гипотеза** – любое предположение либо относительно неизвестного закона распределения, либо относительно неизвестных параметров известного закона распределения.
- **Статистический критерий** – однозначно определенное правило, устанавливающее условия, при которых проверяемую гипотезу следует отвергнуть, либо не отвергать. Основу критерия составляет выборочная характеристика, точное или приближенное распределение которой известно при справедливости нулевой гипотезы. Правила проверки гипотезы определяют, при каких условиях гипотеза будет принята.
- **Типический отбор** – это отбор, при котором объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части.
- **Точечная оценка** – функция результатов наблюдения, значение которой принимается за приближенное значение параметра генеральной совокупности.
- **Уровень значимости** – вероятность не совершить ошибку первого рода, вероятность отвергнуть истинную нулевую гипотезу. С уменьшением вероятности ошибки первого рода увеличивается вероятность ошибки второго рода.
- **Эффективная оценка** – точечная оценка, обладающая наименьшей дисперсией среди всех возможных несмещенных оценок параметра данной генеральной совокупности при фиксированном объеме выборки.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян, С.А. Классификация многомерных наблюдений /С.А. Айвазян, З.И. Бежаева, О.В. Староверов – М.: Статистика, 1974. – 240 с.
2. Айвазян, С.А. Прикладная статистика. Основы моделирования и первичная обработка данных /С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин – М.: Финансы и статистика, 1983. – 472 с.
3. Айвазян, С.А. Прикладная статистика. Исследование зависимостей /С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин – М.: Финансы и статистика, 1985. – 488 с.
4. Прикладная статистика. Классификация и снижение размерности /С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др. – М.: Финансы и статистика, 1989.
5. Андерсон, Т. Введение в многомерный статистический анализ /Т. Андерсон Пер. с англ. – М.: ГИФМЛ, 1963. – 500 с.
6. Болч, Б. Многомерные статистические методы экономики /Б. Болч, К. Хуань Пер. с англ. – М.: Статистика, 1979. – 317 с.
7. Дубров, А.М. Многомерные статистические методы /А.М. Дубров, В.С. Мхитарян, Л.И. Трошин – М.: Финансы и статистика, 2003 . – 352 с.

Составитель
Баранова Татьяна Анатольевна

Многомерные статистические методы.
Корреляционный анализ

Методические указания

Редактор Г.В. Куликова

Подписано в печать 9.03.2007. Формат 60×84 1/16. Печать плоская.
Усл. печ. л. 2,33. Уч. изд. л. 2,58. Тираж 100 экз. Заказ ___
Отпечатано на полиграфическом оборудовании
кафедры экономики и финансов ГОУ ВПО “ИГХТУ”
ГОУ ВПО “Ивановский государственный химико–технологический университет”
153000, г. Иваново, пр. Ф. Энгельса, 7