УДК 330.44

АНАЛИЗ ПОТРЕБЛЕНИЯ ПРОДУКТОВ ПИТАНИЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДА МНОГОМЕРНОГО ДИСПЕРСИОННОГО АНАЛИЗА (MANOVA)

Макжанова Я.В., Швед Е.В.

ФГБОУ ВО «Российский экономический университет им. Г.В. Плеханова», Москва, e-mail: kafedra_vm@mail.ru

Одним из первичных методов обработки групп данных является проверка на однородность, что осуществляется с помощью специального метода математической статистики – дисперсионного анализа. В статье дано краткое описание многомерного дисперсионного анализа (MANOVA) и используемых в нем критериев. Рассматривается применение многомерного дисперсионного анализа для исследования влияния территориального фактора на уровень потребления всех основных продуктов питания в совокупности в регионах Российской Федерации в 2014 году и на изменение уровня потребления каждого из основных продуктов за 2004–2014 г. Расчеты показывают: 1) территориальный фактор оказывает существенное влияние на уровень потребления основных продуктов питания, поэтому рассматриваемые группы данных (по округам) нельзя объединять в одну многомерную генеральную совокупность; 2) территориальный фактор не оказывает влияние на изменение уровня потребления мясных продуктов, сахара и растительного масла с течением времени и оказывает существенное влияние на изменение уровня потребления остальных продуктов.

Ключевые слова: многомерный дисперсионный анализ, MANOVA, потребление, основные продукты питания, территориальный фактор

ANALYSIS OF FOODSTUFF CONSUMPTION USING METHOD OF MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

Makzhanova Ya.V., Shved E.V.

Russian Economic University named after G.V. Plekhanov, Moscow, e-mail: kafedra_vm@mail.ru

One of primary techniques in processing of data groups is testing for homogeneity that can be performed using a special statistical method – analysis of variance. The study gives a short description of multivariate analysis of variance (MANOVA) and its common statistics. The multivariate analysis of variance in this research examines the influence of the regional factor on consumption of all the basic foodstuffs in total in the Russian Federation in 2014 and on change in consumption of each basic foodstuff in 2004–2014. Conclusions: (1) the consumption of basic foodstuffs varied significantly between the regions, therefore the groups under study cannot be joined together into one multivariate population; (2) in 2004–2014 the regional factor did not affect the change in consumption of meat products, sugar and vegetable oil significantly, while its influence on the change in consumption of other foodstuffs was significant.

Keywords: multivariate analysis of variance, MANOVA, consumption, basic foodstuffs, regional factor

Большинство анализируемых экономических данных представляют собой многомерные генеральные совокупности со взаимосвязанными компонентами. Причем связь между компонентами чаще всего не функциональная, а стохастическая. Поэтому представляется логичным использование многомерных методов статистического анализа для их исследования.

Одной из проблем, решаемых математической статистикой, является проблема проверки однородности данных, т.е. принадлежности различных групп данных к одной генеральной совокупности, одномерной или многомерной, с целью их дальнейшего объединения в одну генеральную совокупность и последующей статистической обработки всего массива данных целиком. Если сравниваются две группы данных, то задача сводится к проверке статистической гипотезы о равенстве средних, если групп данных более двух, то к дисперсионному анализу.

В данной статье многомерный дисперсионный анализ (MANOVA) используется для проверки однородности многомерных экономических данных – уровня потребления основных продуктов питания. Прежде чем приступить непосредственно к применению дисперсионного анализа, напомним читателю его основные положения.

Однофакторный одномерный дисперсионный анализ (One-way ANOVA)

В дисперсионном анализе изучается влияние одного или нескольких факторов (независимых категориальных переменных) на результаты наблюдений (значения зависимой переменной X) [11].

Рассмотрим влияние на зависимую переменную X одного фактора A, принимающего m значений – *уровней* – $A_1, A_2, ..., A_m$. На i-том уровне фактора имеется выборка $x_{1i}, x_{2i}, ..., x_{n_i,i}$ объема n_i, $i = 1, 2, \ldots, m$. В об-

щем случае объемы выборок могут быть неравными. Общее число наблюдений равно $n = n_1 + n_2 + \ldots + n_m$. Исходные данные обычно представляют в виде таблицы (табл. 1).

Предполагается, что все выборки, их еще называют *группами*, независимы и извлечены из нормально распределенных совокупностей с равными дисперсиями: $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$, …, $N(\mu_m,\sigma^2)$. В однофакторном дисперсионном анализе проверяется гипотеза о равенстве математических ожиданий этих генеральных совокупностей: $H_0 : \mu_1 = \mu_2 = ... = \mu_m$, т.е. о несущественном влиянии фактора А на результаты наблюдений. Альтернативная гипотеза утверждает, что не все μ_i равны между собой.

Для описания результатов наблюдений используется линейная модель:

$$x_{ij} = \mu_j + \varepsilon_{ij},$$

где ε_{ij} – неизвестные одинаково распределенные случайные величины, характеризующие случайную ошибку, вызванную влиянием неконтролируемых факторов.

В качестве статистического критерия принимается *F*-критерий:

$$F = \frac{MS_B}{MS_W},$$

в случае справедливости нулевой гипотезы имеющий распределения Фишера $F(m - 1, n - m)$,

где $MS_B = \dfrac{\sum_{j=1}^{m} n_j (\overline{x}_j - \overline{x})^2}{m-1} = \dfrac{SSH}{m-1}$ – межгрупповая (факторная) дисперсия (Mean Square between groups),

$MS_W = \dfrac{\sum_{j=1}^{m}\sum_{i=1}^{n_m} \left(x_{ij} - \overline{x}_j\right)^2}{n-m} = \dfrac{SSE}{n-m}$ – внутригрупповая (остаточная) дисперсия (Mean Square within groups),

$\overline{x}$ – общая средняя,

$SSH = \sum_{j=1}^{m} n_j (\overline{x}_j - \overline{x})^2$ – межгрупповая сумма квадратов отклонений (Hypothesis Sum of Squares),

$SSE = \sum_{j=1}^{m}\sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_j\right)^2$ – внутригрупповая сумма квадратов отклонений (Error Sum of Squares).

MS_B и MS_W являются статистическими оценками генеральной дисперсии σ^2, характеризующими разброс данных между выборками и внутри выборок соответственно. Если математические ожидания μ_i не равны между собой, то разброс данных внутри групп относительно групповых средних должен быть меньше, чем разброс групповых средних относительно общей средней, и тогда значение *F*-критерия будет большим. Нулевая гипотеза отвергается, если $F > F_{кр}$. На этом основана идея одномерного дисперсионного анализа.

Однофакторный многомерный дисперсионный анализ (One-way MANOVA)

Естественным обобщением ANOVA на многомерный случай является многомерный дисперсионный анализ – MANOVA (Multivariable Analysis of Variances).

В многомерном случае предполагается, что *m* независимых выборок $\mathbf{X}_{1j}, \mathbf{X}_{2j}, ..., \mathbf{X}_{n_j,j}$, $j = 1,2,\ldots,m$, извлечены из *k*-мерных генеральных совокупностей, имеющих многомерное нормальное распределение с одинаковыми ковариационными матрицами Σ. Таким образом, изучается влияние одного фактора *A* – одной независимой категориальной переменной – на *k*-мерный вектор зависимых переменных $\mathbf{X} = \left(X_1, X_2, ..., X_k\right)$, коррелирующих между собой. Общее число векторов наблюдений равно $n = n_1 + n_2 + \ldots + n_m$.

Исходные данные можно представить в виде таблицы (табл. 2), где $\mathbf{X}_{ij} = (x_{ij1}, x_{ij2}, ..., x_{ijk})^{\mathrm{T}}$ – *k*-мерные векторы наблюдений, $i = 1,2,\ldots,n_j$, $j = 1,2,\ldots,m$, $\overline{\mathbf{X}}_j = (\overline{x}_{j1}, \overline{x}_{j2}, ..., \overline{x}_{jk})^{\mathrm{T}}$ – *k*-мерные векторы выборочных средних с компонентами $\overline{x}_{jr} = \dfrac{1}{n_j}\sum_{i=1}^{n_j} x_{ijr}$, $r = 1, 2,\ldots,k$.

Вектор общих средних имеет вид

$$\overline{\mathbf{X}} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_k)^{\mathrm{T}},$$

где $\overline{x}_r = \dfrac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{ijr}$, $r = 1, 2,\ldots,k$.

Результаты наблюдений описываются линейной моделью

$$\mathbf{X}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_{ij},$$

или в векторной форме

$$\begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \ldots \\ x_{ijk} \end{pmatrix} = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \\ \ldots \\ \mu_{jk} \end{pmatrix} + \begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \\ \ldots \\ \varepsilon_{ijk} \end{pmatrix},$$

где ε_{ijr} – неизвестные одинаково распределенные случайные величины, характеризующие случайную ошибку, $r = 1, 2,\ldots,k$.

Таблица 1

Исходные данные для ANOVA

	A_1 (Выборка 1)	A_2 (Выборка 2)	…	A_m (Выборка m)
	x_{11} x_{21} … $x_{n_1,1}$	x_{12} x_{22} … $x_{n_2,2}$	… … … … …	x_{1m} x_{2m} … $x_{n_m,m}$
Выборочное среднее	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_m$
Объем выборки	n_1	n_2	…	n_m

Таблица 2

Исходные данные для MANOVA

	A_1 (Выборка 1)	A_2 (Выборка 2)	…	A_m (Выборка m)
	$\mathbf{X}_{11}$ $\mathbf{X}_{21}$ … $\mathbf{X}_{n_1,1}$	$\mathbf{X}_{12}$ $\mathbf{X}_{22}$ … $\mathbf{X}_{n_2,2}$	… … … … …	$\mathbf{X}_{1m}$ $\mathbf{X}_{2m}$ … $\mathbf{X}_{n_m,m}$
Вектор выборочных средних	$\bar{\mathbf{X}}_1$	$\bar{\mathbf{X}}_2$	…	$\bar{\mathbf{X}}_m$
Объем выборки	n_1	n_2	…	n_m

Нулевая гипотеза в многомерном случае имеет вид

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_m,$$

где $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2}, ..., \mu_{jk})^{\mathrm{T}}$ – k-мерные векторы математических ожиданий зависимых переменных, $j = 1, 2, \ldots, m$, то есть требуется проверить выполнение равенств для соответствующих компонент векторов $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_m$.

Важнейшим преимуществом проведения однократной процедуры MANOVA для k-мерного вектора зависимых переменных $\mathbf{X} = (X_1, X_2, ..., X_k)$ вместо проведения k процедур ANOVA отдельно для каждой из зависимых переменных X_r, $r = 1, 2, \ldots, k$, является учет корреляции зависимых переменных друг с другом, что позволяет учесть все связи, скрытые в массивах многомерных числовых данных [2]. Кроме того, многомерный дисперсионный анализ является первым этапом на пути к решению проблемы классификации исследуемых объектов, а также предварительным шагом перед проведением процедуры снижения размерности данных путем исключения наименее важных признаков. Еще одним достоинством MANOVA является тот факт, что он менее чувствителен к условию нормальности исходных данных, чем ANOVA. Поэтому его применяют вместо ANOVA еще и для анализа повторных данных, для которых не выполняется условие сферичности.

В отличие от одномерного случая вместо межгрупповой и внутригрупповой сумм квадратов SSH и SSE рассматриваются их обобщения – межгрупповая и внутригрупповая матрицы **H** и **E** (симметричные квадратные матрицы порядка k), определяемые следующим образом [12]:

$$\mathbf{H} = \sum_{j=1}^{m} n_j \left(\bar{\mathbf{X}}_j - \bar{\mathbf{X}}\right)\left(\bar{\mathbf{X}}_j - \bar{\mathbf{X}}\right)^{\mathrm{T}} = \begin{pmatrix} SSH_{11} & SPH_{12} & ... & SPH_{1k} \\ SPH_{12} & SSH_{22} & ... & SPH_{2k} \\ ... & ... & ... & ... \\ SPH_{1k} & SPH_{2k} & ... & SSH_{kk} \end{pmatrix},$$

$$\mathbf{E} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_j\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_j\right)^{\mathrm{T}} = \begin{pmatrix} SSE_{11} & SPE_{12} & \ldots & SPE_{1k} \\ SPE_{12} & SSE_{22} & \ldots & SPE_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ SPE_{1k} & SPE_{2k} & \ldots & SSE_{kk} \end{pmatrix},$$

где $SSH_{rr} = \sum_{j=1}^{m} n_j \left(\overline{x}_{jr} - \overline{x}_r\right)^2$ – межгрупповые суммы квадратов (Hypothesis Sums of Squares),

$SPH_{qr} = \sum_{j=1}^{m} n_j \left(\overline{x}_{jr} - \overline{x}_r\right)\left(\overline{x}_{jq} - \overline{x}_q\right)$ – межгрупповые суммы произведений (Hypothesis Sums of Products),

$SSE_{rr} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijr} - \overline{x}_{jr}\right)^2$ – внутригрупповые суммы квадратов (Error Sums of Squares),

$SPE_{qr} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijr} - \overline{x}_{jr}\right)\left(x_{ijq} - \overline{x}_{jq}\right)$ – внутригрупповые суммы произведений (Error Sums of Products), $r = 1,2,...,k$, $q = 1,2,...,k$, $q \neq r$.

Для проверки нулевой гипотезы используются четыре статистических критерия [2, 12], приведенных в табл. 3. Все критерии – это скалярные величины, рассчитанные с помощью матриц **H** и **E** и основанные на различных подходах к определению критических статистик. Критические значения критериев можно найти в специальных таблицах [12]. В табл. 3 также приведены выражения для критериев через собственные значения $\lambda_1, \lambda_2, ..., \lambda_k$ матрицы $\mathbf{E}^{-1}\mathbf{H}$ или матрицы $\mathbf{H}\mathbf{E}^{-1}$ (собственные значения обеих матриц совпадают, но собственные векторы различны!).

Данные статистические критерии можно аппроксимировать *F*-статистикой Фишера со степенями свободы df_1, df_2 [2, 12]. Соответствующие формулы приведены в табл. 4.

Необходимо отметить, что максимальный корень по методу Роя является более мощным критерием по сравнению с тремя другими только в том случае, если векторы средних значений коллинеарны между собой [12]. Это возможно в ситуации, когда наибольшее собственное значение λ_1 существенно превосходит (в несколько раз) все остальные собственные значения. В остальных случаях его можно не принимать во внимание.

Таблица 3

Статистические критерии для MANOVA

Статистический критерий	Формула	Критическая область	Выражение критерия через собственные значения матрицы $\mathbf{E}^{-1}\mathbf{H}$
Лямбда Уилкса (Wilks' Lambda)	$\Lambda = \frac{\lvert E\rvert}{\lvert E+H\rvert}$, $0 \le \Lambda \le 1$	$\Lambda \le \Lambda_\alpha$	$\Lambda = \prod_{r=1}^{k} \frac{1}{1+\lambda_r}$
След Хотеллинга (Hotelling's Trace)	$T_0^2 = trace\left(E^{-1}H\right)$	$T_0^2 \ge T_{0\,\alpha}^2$	$T_0^2 = \sum_{r=1}^{k} \lambda_r$
След Пиллая (Pillai's Trace)	$V = trace\left((E+H)^{-1}H\right)$	$V \ge V_\alpha$	$V = \sum_{r=1}^{k} \frac{\lambda_r}{1+\lambda_r}$
Максимальный корень по методу Роя (Roy's Largest Root)	$\theta = \lambda_1$ или $\theta = \frac{\lambda_1}{1+\lambda_1}$, где λ_1 – наибольшее собственное значение матрицы $\mathbf{E}^{-1}\mathbf{H}$	$\theta \ge \theta_\alpha$	$\theta = \lambda_1$ или $\theta = \frac{\lambda_1}{1+\lambda_1}$

Таблица 4

Аппроксимация критериев *F*-статистикой

Статистический критерий	Соответствующая аппроксимирующая *F*-статистика со степенями свободы df_1, df_2
Лямбда Уилкса (Wilks' Lambda)	$F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \cdot \frac{df_2}{df_1}$, где $df_1 = k(m-1); \quad df_2 = wt - \frac{k(m-1)-2}{2}; \quad w = n-1-\frac{k+m}{2};$ $t = \sqrt{\frac{k^2(m-1)^2-4}{k^2+(m-1)^2-5}}$
След Хотеллинга (Hotelling's Trace)	$F = \frac{T_0^2}{s} \cdot \frac{df_2}{df_1}$, где $df_1 = s(2t+s+1); \quad df_2 = 2(su+1); \quad s = \min(k, m-1);$ $u = \frac{n-m-k-1}{2}; \quad t = \frac{\lvert k-m+1\rvert -1}{2}$
След Пиллая (Pillai's Trace)	$F = \frac{V}{s-V} \cdot \frac{df_2}{df_1}$, где $df_1 = s(2t+s+1); \quad df_2 = s(2u+s+1); \quad s = \min(k, m-1)$ $u = \frac{n-m-k-1}{2}; \quad t = \frac{\lvert k-m+1\rvert -1}{2}$
Максимальный корень по методу Роя (Roy's Largest Root)	Ввиду отсутствия удовлетворительной аппроксимации критерия *F*-статистикой используется ее приблизительная «верхняя граница»: $F = \lambda_1 \frac{n-m-d-1}{d}$, где $d = \max(k, m-1); \quad df_1 = d; \quad df_2 = n-m-d-1$

Анализ потребления продуктов питания

Показатели душевого потребления основных продуктов питания входят в число критериев оценки уровня жизни населения в регионе. Неоднородность потребления продуктов в регионах Российской Федерации объясняется различиями в доходах, уровнем развития сельского хозяйства в регионе, степенью обеспеченности региона определенным видом продукта, климатическими особенностями, наконец, исторически сложившимися традициями в питании в конкретном регионе. Например, спрос на мясо и мясопродукты выше в регионах с высокими среднедушевыми доходами; в регионах со средним уровнем доходов и высоким уровнем развития сельхозпроизводства высокий уровень потребления всех продуктов питания обеспечивается более низкими ценами из-за конкуренции сельхозпроизводителей [3] и т.д. Возникают закономерные вопросы: существенны ли различия в потреблении продуктов питания в регионах России и существенно ли меняется потребление с течением времени?

Статистические данные по потреблению основных продуктов питания в регионах Российской Федерации за разные временные промежутки анализировались во многих работах [1, 3, 4, 6–9], в том числе и с применением одномерного дисперсионного анализа [5, 10]. Как правило, в подобных исследованиях регионы (или федеральные округа) сравниваются по уровню потребления друг с другом или с рациональными нормами питания отдельно по каждому из основных продуктов питания, проводится разбиение регионов на кластеры, делаются попытки построить уравнения регрессии и дать прогноз. Мы же попробуем сравнить

между собой федеральные округа по уровню потребления совокупности всех основных продуктов одновременно. Представляется, что такой подход является обоснованным в силу существующих взаимосвязей (соотношений) между уровнем потребления разных видов продуктов: недостаточное потребление одного вида продукта компенсируется избыточным потреблением другого (например, мясо – картофель), или увеличенное потребление одного влечет увеличенное потребление другого сопутствующего продукта (например, ягоды – сахар).

В статье рассматривается анализ потребления основных продуктов питания в различных округах Российской Федерации по данным Росстата за 2014 год. Изучается влияние территориального фактора – федерального округа – на восьмимерный вектор показателей потребления

$$\mathbf{X} = (X_1, X_2,\ldots, X_8),$$

где X_1 – душевое потребление мяса и мясопродуктов (кг за год),
X_2 – душевое потребление молока и молочных продуктов (кг за год),
X_3 – душевое потребление яиц (шт за год),
X_4 – душевое потребление сахара (кг за год),
X_5 – душевое потребление картофеля (кг за год),
X_6 – душевое потребление овощей и бахчевых культур (кг за год),
X_7 – душевое потребление растительного масла (кг за год),
X_8 – душевое потребление хлебных продуктов (кг за год).

Исходные данные частично представлены в табл. 5.

Таким образом, в переводе на язык математической статистики, изучается влияние территориального фактора, имеющего восемь уровней (по числу федеральных округов), на вектор показателей потребления $\mathbf{X} = (X_1, X_2,\ldots, X_8)$. Число наблюдений (объемы выборок n_j, $j = 1,2,...,8$) на каждом уровне фактора различно, так как в каждом федеральном округе разное число регионов.

Перечисленные продукты питания хотя и слабо, но коррелируют между собой. Корреляционная матрица по данным за 2014 год выглядит следующим образом (табл. 6).

Несмотря на то, что метод MANOVA не очень чувствителен к требованию многомерной нормальности, проверим, нет ли очевидных свидетельств того, что распределение рассматриваемой генеральной совокупности существенно отличается от нормального. Для этого рекомендуется [12], во-первых, проверить на нормальность компоненты вектора $\mathbf{X} = (X_1, X_2,\ldots, X_8)$, во-вторых, визуально проанализировать диаграммы рассеяния для всех возможных пар компонент вектора X. Если облако рассеяния для какой-то пары переменных отличается от эллиптического, то есть основание сомневаться в нормальной распределенности генеральной совокупности.

Таблица 5

Исходные данные (по данным Росстата)

		Продукты питания							
Округ	Регион	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
ЦФО	Белгородская область	97	261	318	47	119	110	13,8	139
ЦФО	Брянская область	64	208	230	34	155	100	11,2	114
…	…	…	…	…	…	…	…	…	…
ДФО	Чукотский авт. округ	51	109	147	34	59	26	17,8	61

Таблица 6

Корреляционная матрица

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1	0,342	0,224	0,268	– 0,004	– 0,028	0,140	0,147
X_2	0,342	1	0,335	0,104	0,020	0,165	0,085	0,222
X_3	0,224	0,335	1	0,293	0,092	0,230	0,241	0,126
X_4	0,268	0,104	0,293	1	– 0,008	0,244	0,238	0,371
X_5	– 0,004	0,020	0,092	– 0,008	1	0,171	– 0,074	0,333
X_6	– 0,028	0,165	0,230	0,244	0,171	1	0,043	0,218
X_7	0,140	0,085	0,241	0,238	– 0,074	0,043	1	– 0,025
X_8	0,147	0,222	0,126	0,371	0,333	0,218	– 0,025	1

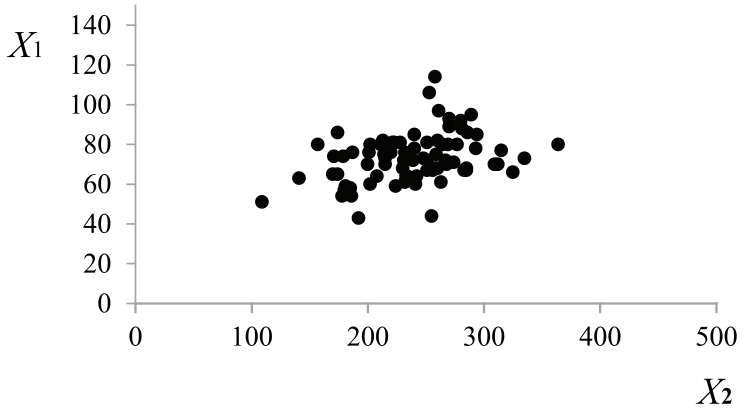

Диаграмма рассеяния переменных X_1 и X_2

Таблица 7

Наблюдаемые частоты

	Переменные							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Номер интервала	Частоты							
1	3	1	1	6	3	2	7	1
2	8	4	4	8	6	3	13	0
3	18	14	7	17	17	36	24	6
4	26	19	14	20	25	28	19	20
5	17	22	25	15	15	6	9	30
6	5	14	17	7	7	4	4	18
7	1	4	11	4	4	0	2	3
8	2	2	1	3	3	1	2	2

Каждая из рассматриваемых переменных X_1, X_2,…, X_8, взятая по отдельности, имеет распределение, близкое к нормальному, что подтверждается наблюдаемыми частотами распределения каждой переменной, приведенными в табл. 7.

Диаграммы рассеяния для рассматриваемых переменных – показателей потребления – похожи на показанную на рисунке диаграмму для переменных X_1 и X_2. Как видно, облако рассеяния имеет форму, близкую к эллиптической. Таким образом, можно принять, что выборка извлечена из многомерной генеральной совокупности, имеющей распределение, близкое к нормальному.

Одним из условий применимости многомерного дисперсионного анализа является равенство ковариационных матриц рассматриваемых групп данных, что проверяется многомерным критерием *M*-Бокса. В нашем случае данный критерий невозможно применить ко всей совокупности данных, так как число наблюдений в некоторых выборках меньше, чем размерность вектора *X*, а именно: в Юго-Восточном, Северо-Кавказском и Уральском федеральных округах число регионов равно 6, 7 и 4 соответственно, что меньше, чем число основных продуктов питания.

Применим метод MANOVA к статистическим данным. Проведенные расчеты подтверждают существенность влияния территориального фактора на потребление всех продуктов питания в совокупности. Значения статистических критериев, соответствующая аппроксимирующая *F*-статистика и ее *p*-значение даны в табл. 8.

Так как для всех четырех критериев *p*-значение существенно меньше 0,05, то гипотеза о несущественности влияния территориального фактора на уровень потребления продуктов питания отвергается. Таким образом, территориальный фактор оказывает значимое влияние на потребле-

ние всей совокупности продуктов. Это означает, что рассматриваемые данные по разным федеральным округам нельзя объединять в одну многомерную генеральную совокупность, следовательно, нельзя их использовать для построения уравнения множественной регрессии, а также сомнительно, что можно применять к ним метод главных компонент.

Если в рамках многомерного дисперсионного анализа гипотеза о несущественности влияния фактора отвергается, то рекомендуется провести k процедур однофакторного дисперсионного анализа, чтобы проанализировать, какие из зависимых переменных вносят наиболее существенный вклад в неоднородность данных. В табл. 9 даны результаты проверки гипотезы о влиянии территориального фактора на уровень потребления каждого из продуктов питания, взятых по отдельности, проведенной с помощью однофакторного дисперсионного анализа по данным за 2014 год.

Полученные результаты свидетельствуют о том, что территориальный фактор не оказывает влияние на потребление трех продуктов питания (мясопродукты, растительное масло и хлебопродукты) при уровне значимости 0,05. Потребление остальных основных продуктов питания существенно различается по федеральным округам.

Если рассмотреть вектор потребления **X**, включающий только три продукта: мясопродукты, растительное масло и хлебопродукты, и применить многомерный дисперсионный анализ для проверки значимости влияния территориального фактора, то результаты будут следующими (табл. 10).

Так как p-значение первых трех критериев превышает 0,05, то гипотезу об отсутствии влияния территориального фактора на потребление данных трех продуктов питания можно принять.

Применение MANOVA к повторным данным

Воспользуемся методом многомерного дисперсионного анализа для выяснения, существенно ли влияет территориальный фактор на изменение потребления продуктов с течением времени. Для этого используем статистические данные Росстата за 2004–2014 годы с шагом 2 года по каждому из основных продуктов питания по отдельности. Фрагмент массива исходных данных по переменной X_1 (мясо и мясопродукты) представлен в табл. 11.

Таблица 8

Результаты расчетов

Статистический критерий	Значение критерия	Значение аппроксимирующей F-статистики	p-значение F-статистики
Лямбда Уилкса	0,116	3,114	$8{,}05\times10^{-11}$
След Хотеллинга	2,783	3,145	$1{,}64\times10^{-11}$
След Пиллая	1,716	2,883	$4{,}03\times10^{-10}$
Максимальный корень по методу Роя	0,939	8,333	< 0,001

Таблица 9

Результаты ANOVA для компонент вектора X

	Зависимые переменные (продукты питания)							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Значение F-критерия	1,679	4,538	2,213	2,601	2,511	4,474	0,5685	1,7134
p-значение критерия	0,1278	0,0003	0,0429	0,0189	0,0229	0,0004	0,7791	0,1191

Таблица 10

Результаты расчетов для трехмерного вектора X

Статистический критерий	Значение критерия	Значение аппроксимирующей F-статистики	p-значение F-статистики
Лямбда Уилкса	0,669	1,439	0,103
След Хотеллинга	0,444	1,451	0,098
След Пиллая	0,365	1,424	0,109
Максимальный корень по методу Роя	0,294	3,028	0,008

Таблица 11

Исходные повторные данные для переменной X_1

		Душевое потребление мяса и мясопродуктов (кг в год)					
Округ	Регион	2004	2006	2008	2010	2012	2014
ЦФО	Белгородская область	66	80	88	92	97	97
ЦФО	Брянская область	58	60	60	61	62	64
…	…	…	…	…	…	…	…
ДФО	Чукотский авт. округ	37	40	50	53	51	51

Таблица 12

Преобразованные данные

		Прирост душевого потребления мяса и мясопродуктов (кг за 2 года)				
Округ	Регион	2004–2006	2006–2008	2008–2010	2010–2012	2012–2014
ЦФО	Белгородская область	14	8	4	5	0
ЦФО	Брянская область	2	0	1	1	2
…	…	…	…	…	…	…
ДФО	Чукотский авт. округ	3	10	3	– 2	0

Таблица 13

Результаты расчетов для повторных данных

		Статистические критерии			
Продукт		L	T_0^2	V	q
Мясные продукты	Значение критерия	0,597	0,564	0,475	0,287
	F-статистика	1,060	1,054	1,064	2,907
	***p*-значение**	0,383	0,390	0,376	0,010
Молочные продукты	Значение критерия	0,423	0,992	0,753	0,464
	F-статистика	1,842	1,853	1,799	4,710
	***p*-значение**	0,004	0,003	0,005	< 0,001
Яйца	Значение критерия	0,493	0,821	0,617	0,479
	F-статистика	1,488	1,534	1,429	4,862
	***p*-значение**	0,043	0,031	0,059	< 0,001
Сахар	Значение критерия	0,725	0,342	0,304	0,176
	F-статистика	0,647	0,640	0,656	1,783
	***p*-значение**	0,940	0,945	0,935	0,104
Картофель	Значение критерия	0,355	1,337	0,834	0,926
	F-статистика	2,268	2,498	2,032	9,397
	***p*-значение**	< 0,001	< 0,001	0,001	< 0,001
Овощи и бахчевые	Значение критерия	0,364	1,203	0,865	0,605
	F-статистика	2,208	2,249	2,121	6,135
	***p*-значение**	< 0,001	< 0,001	< 0,001	< 0,001
Растительное масло	Значение критерия	0,508	0,778	0,598	0,455
	F-статистика	1,421	1,454	1,377	4,618
	***p*-значение**	0,065	0,052	0,081	0,000
Хлебные продукты	Значение критерия	0,441	0,944	0,719	0,475
	F-статистика	1,747	1,763	1,703	4,819
	***p*-значение**	0,008	0,006	0,009	< 0,001

Преобразуем исходные данные: вместо душевого потребления рассмотрим двухгодичный прирост душевого потребления (табл. 12).

К преобразованным данным для всех восьми продуктов питания применим многомерный дисперсионный анализ. Результаты расчетов – значения критериев, аппроксимирующая *F*-статистика и ее *p*-значения – представлены в табл. 13.

Таким образом, на протяжении рассматриваемого периода времени территориальный фактор не оказывал существенного влияния на изменение уровня потребления мясных продуктов, сахара и растительного масла и оказывал весомое влияние на изменение потребления яиц, молочных продуктов, картофеля, овощей и хлебных продуктов. То есть можно считать, что изменение потребления мясных продуктов, сахара и растительного масла в течение 10-летнего периода в 2004–2014 годах было однородным на всей территории страны, а уровень потребления остальных продуктов изменялся по-разному в различных федеральных округах. Отметим, что характер изменения уровня потребления с течением времени в данной статье не рассматривается.

Заключение

Группы экономических многомерных данных перед их объединением в одну многомерную совокупность следует проверять на однородность с помощью многомерного дисперсионного анализа. Такая проверка была проведена для статистических данных об уровне потребления основных продуктов питания в регионах Российской Федерации, объединенных в группы по территориальному признаку (федеральные округа).

С помощью многомерного дисперсионного анализа удалось выяснить, что уровень потребления основных продуктов питания существенно различается в различных федеральных округах, поэтому рассматриваемые группы данных нельзя объединять в одну многомерную генеральную совокупность. Если же рассматривать совокупность продуктов, включающую только мясные продукты, хлебные продукты и растительное масло, то тогда потребление можно считать однородным.

Также оказалось, что изменение уровня потребления мясных продуктов, сахара и растительного масла с течением времени однородно по всей территории страны, для других продуктов питания это изменение неоднородно.

Полученные выводы можно использовать в качестве предварительной информации, предшествующей дальнейшему применению многомерных статистических методов для анализа рассматриваемых данных.

Список литературы

1. Айзинова И.М. Потребление продуктов питания в регионах России // Проблемы прогнозирования. – 2014. – № 6 (147). – С. 44–59.

2. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ/ пер. с нем. и предисл. В.М. Ивановой и Ю. Н. Тюрина. – Москва: Финансы и статистика, 1985. – 230 с.

3. Баканач О.В., Проскурина Н.В. Статистический анализ территориальной дифференциации уровня потребления основных продуктов питания в регионах РФ // Вестник Самарского государственного экономического университета. – 2012. – № 10(96). – С. 29–33.

4. Грешонков А.М., Меркулова Е.Ю. Анализ потребления основных продуктов питания по регионам РФ // Социально-экономические явления и процессы. – 2014. – Т. 9. – № 11. – С. 54–62.

5. Жук Т.С., Маслак А.А. Сравнительный анализ округов Российской Федерации по потреблению продуктов питания // Теория и практика измерения и мониторинга компетенций и других латентных переменных в образовании. XXI и XXII Всероссийские научно-практические конференции: сборник научных трудов. под. ред.: А.А. Маслака; Филиал Кубанского гос. ун-та в г. Славянске-на-Кубани. 2014. – С. 227–232.

6. Зарецкая А.С. Статистическая оценка обеспеченности населения региона продуктами питания в системе продовольственной безопасности страны // Вестник Новгородского государственного университета им. Ярослава Мудрого. – 2014. – № 82. – С. 91–95.

7. Игнатьев В.М., Середа М.В. Статистический анализ потребления продуктов питания населением регионов [Электронный ресурс] // Экономические исследования: электронный журнал. – 2015. – № 2. – Режим доступа: http://elibrary.ru/download/63770882.pdf.

8. Исаенко А.В. Потребление продуктов питания в домашних хозяйствах России // Вестник Белгородского университета кооперации, экономики и права. – 2005. – № 1. – С. 243–252.

9. Кудрявцева Л.Н. Региональные особенности платежеспособного спроса населения в контексте реализации принципов здорового питания // Никоновские чтения. – 2014. – № 19. – С. 248–250.

10. Макжанова Я.В. Потребление основных продуктов питания в федеральных округах Российской Федерации за период 1996–2002 гг. // Экономический анализ: теория и практика. – 2004. – № 8. – С. 65–68.

11. Математика для экономистов. Практикум: учеб пособие для академического бакалавриата / под общ. ред. О.В. Татарникова. – Москва: Издательство Юрайт, 2014. – 285 с.

12. Rencher A.C., Christensen W.F. Methods of Multivariate Analysis. – 3rd ed. – John Wiley & Sons, Inc., Hoboken, New Jersey, 2012. – 758 p.

References

1. Ajzinova I.M. Potreblenie produktov pitanija v regionah Rossii // Problemy prognozirovanija. 2014. no. 6 (147). pp. 44–59.

2. Arens H., Lejter Ju. Mnogomernyj dispersionnyj analiz/ per. s nem. i predisl. V.M. Ivanovoj i Ju. N. Tjurina. Moskva: Finansy i statistika, 1985. 230 p.

3. Bakanach O.V., Proskurina N.V. Statisticheskij analiz territorialnoj differenciacii urovnja potreblenija osnovnyh produktov pitanija v regionah RF // Vestnik Samarskogo gosudarstvennogo jekonomicheskogo universiteta. 2012. no. 10(96). pp. 29–33.

4. Greshonkov A.M., Merkulova E.Ju. Analiz potreblenija osnovnyh produktov pitanija po regionam RF // Social-no-jekonomicheskie javlenija i processy. 2014. T. 9. no. 11. pp. 54–62.

5. Zhuk T.S., Maslak A.A. Sravnitelnyj analiz okrugov Rossijskoj Federacii po potrebleniju produktov pitanija // Teorija i praktika izmerenija i monitoringa kompetencij i drugih latentnyh peremennyh v obrazovanii. XXI i XXII Vserossijskie nauchno-prakticheskie konferencii: sbornik nauchnyh trudov. pod. red.: A.A. Maslaka; Filial Kubanskogo gos. un-ta v g. Slavjanske-na-Kubani. 2014. pp. 227–232.

6. Zareckaja A.S. Statisticheskaja ocenka obespechennosti naselenija regiona produktami pitanija v sisteme prodovolstvennoj bezopasnosti strany // Vestnik Novgorodskogo gosudarstvennogo universiteta im. Jaroslava Mudrogo. 2014. no. 82. pp. 91–95.

7. Ignatev V.M., Sereda M.V. Statisticheskij analiz potreblenija produktov pitanija naseleniem regionov [Jelektronnyj resurs] // Jekonomicheskie issledovanija: jelektronnyj zhurnal. 2015. no. 2. Rezhim dostupa: http://elibrary.ru/download/63770882.pdf.

8. Isaenko A.V. Potreblenie produktov pitanija v domashnih hozjajstvah Rossii // Vestnik Belgorodskogo universiteta kooperacii, jekonomiki i prava. 2005. no. 1. pp. 243–252.

9. Kudrjavceva L.N. Regionalnye osobennosti platezhesposobnogo sprosa naselenija v kontekste realizacii principov zdorovogo pitanija // Nikonovskie chtenija. 2014. no. 19. pp. 248–250.

10. Makzhanova Ja.V. Potreblenie osnovnyh produktov pitanija v federalnyh okrugah Rossijskoj Federacii za period 1996–2002 gg. // Jekonomicheskij analiz: teorija i praktika. 2004. no. 8. pp. 65–68.

11. Matematika dlja jekonomistov. Praktikum: ucheb posobie dlja akademicheskogo bakalavriata / pod obshh. red. O.V. Tatarnikova. Moskva: Izdatelstvo Jurajt, 2014. 285 p.

12. Rencher A.C., Christensen W.F. Methods of Multivariate Analysis. 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012. 758 p.