Инженерный вестник Дона, №3 (2019)
ivdon.ru/ru/magazine/archive/n3y2019/5808

Программное обеспечение многофакторного дисперсионного анализа

А.И. Джангаров, Х. А. Ахметова

Чеченский государственный университет, г. Грозный

Аннотация: объектом исследования в данной статье стала довольно важная тема описания такого статистического метода, как дисперсионный анализ. Дело в том, что многие программные продукты рассматривают, в основном, случай, когда число факторов не более двух. За исключением такого мощного продукта, как SPSS Statistica. То же можно отнести и к научным трудам, где очень редко рассматривается и описывается метод дисперсионного анализа с большим количеством факторов (зачастую, когда речь заходит о многофакторном дисперсионном анализе, также рассматривается два фактора). Поэтому был произведен анализ имеющегося математического аппарата, предложена собственная идея реализации данного статистического метода и создан программный продукт, демонстрирующий его работу.

Ключевые слова: дисперсионный анализ, общий случай, программная реализация.

Метод многофакторного дисперсионного анализа является одним из важнейших инструментов в статистике и призван дать оценку влияния, как одного фактора, так и нескольких на исследуемые выходные переменные (признаки).

Актуальность данного статистического метода определяется тем, что область применения дисперсионного анализа достаточно широка. Его применяют в медицинских исследованиях, в химических экспериментах, в инженерных исследованиях, в методике воспитания физической дисциплины и т.д. Современные технологии позволяют достаточно быстро реализовать данный метод и получить результаты с довольно низкой вероятностью ошибки. Это способствует росту производительности во многих сферах нашей жизни и позволяет быстрее принимать верные и наименее рискованные решения [1].

Цель данной работы состояла в разборе, изучении данного статистического метода и создании программы, реализующей многофакторный дисперсионный анализ в среде программирования Delphi.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Подробно были рассмотрены и изучены различные математические модели, а также предложена собственная модель для реализации данного метода.

Многофакторный дисперсионный анализ

Алгоритм проведения дисперсионного анализа выглядит следующим образом:

1) Разбиение сумм квадратов;
2) Нахождение и оценка дисперсий;
3) Оценка действия фактора.

Рассмотрим одновременное действие факторов x_1 и x_2. Соответствующая таблица (таблица 1) хранит в себе результаты наблюдений из серии параллельных измерений $u_1 \times u_2 \times m$ при многочисленных опытах над экспериментальными данными y_{jgl}, где j – исследуемый уровень изменения первого фактора $x_1 (j = 1, 2, \ldots, u_1)$; g – исследуемый уровень второго фактора $x_2 (g = 1, 2, \ldots, u_2)$; l – порядковый номер исследования серии наблюдений jg-м в случае, когда факторов несколько ($l = 1, 2, \ldots, m_{jg}$) [2]. Прежде всего необходимо вычислить суммы результатов наблюдений для всех возможных вариантов сочетания факторов x_1 и x_2 при исследуемых уровнях j и g:

$$\overline{y}_{jg} = \frac{1}{m} \sum_{l=1}^{m} y_{jgl}$$

далее необходимо вычислить средние арифметические сумм $\overline{y}_j$, для фактора x_1:

$$\overline{y}_j = \frac{1}{u_2 m} \sum_{g=1}^{u_2} \sum_{l=1}^{m} y_{jgl} = \frac{1}{u_2} \sum_{g=1}^{u_2} \overline{y}_{jg}$$

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

средние арифметические сумм $\overline{y}_g$ для фактора x_2

$$\overline{y}_g = \frac{1}{u_1 m}\sum_{j=1}^{u_1}\sum_{l=1}^{m} y_{jgl} = \frac{1}{u_1}\sum_{j=1}^{u_1}\overline{y}_{jg}$$

и наконец – общее среднее арифметическое всех сумм по строкам таблицы 1.

Таблица № 1

Результаты исследований

фактор x_2 / фактор x_1	1	2	…	g	…	u_2	$\overline{y}_j = \frac{1}{u_2 m}\sum_{g=1}^{u_2}\sum_{l=1}^{m} y_{jgl}$
1	y_{111} y_{112} ⋮ y_{11l} ⋮ y_{11m}	y_{121} y_{122} ⋮ y_{12l} ⋮ y_{12m}	… … … … … … … …	y_{1g1} y_{1g2} ⋮ y_{1gl} ⋮ y_{1gm}	… … … … … … … …	$y_{1u_2 1}$ $y_{1u_2 2}$ ⋮ $y_{1u_2 l}$ ⋮ $y_{1u_2 m}$	$\overline{y}_1$
2	y_{211} y_{212} ⋮ y_{21l} ⋮ y_{21m}	y_{221} y_{222} ⋮ y_{22l} ⋮ y_{22m}	… … … … … … … …	y_{2g1} y_{2g2} ⋮ y_{2gl} ⋮ y_{2gm}	… … … … … … … …	$y_{2u_2 1}$ $y_{2u_2 2}$ ⋮ $y_{2u_2 l}$ ⋮ $y_{2u_2 m}$	$\overline{y}_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

j	y_{j11} y_{j12} $\vdots$ y_{j1l} $\vdots$ y_{j1m}	y_{j21} y_{j22} $\vdots$ y_{j2l} $\vdots$ y_{j2m}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{jg1} y_{jg2} $\vdots$ y_{jgl} $\vdots$ y_{jgm}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{ju_21} y_{ju_22} $\vdots$ y_{ju_2l} $\vdots$ y_{ju_2m}	$\overline{y}_j$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
u_1	$y_{u_11 1}$ y_{u_112} $\vdots$ y_{u_11l} $\vdots$ y_{u_11m}	y_{u_121} y_{u_122} $\vdots$ y_{u_12l} $\vdots$ y_{u_12m}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{u_1g1} y_{u_1g2} $\vdots$ y_{u_1gl} $\vdots$ y_{u_1gm}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	$y_{u_1u_21}$ $y_{u_1u_22}$ $\vdots$ $y_{u_1u_2l}$ $\vdots$ $y_{u_1u_2m}$	$\overline{y}_{u_1}$
$\overline{y}'_g = \frac{1}{u_1 m}\sum_{j=1}^{u_1}\sum_{l=1}^{m} y_{jgl}$	$\overline{y}'_1$	$\overline{y}'_2$	$\cdots$	$\overline{y}'_g$	$\cdots$	$\overline{y}'_{u_2}$	$\overline{\overline{y}} = \frac{1}{u_1 u_2 m}\sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m} y_{jgl}$

Такое распределение наблюдений приводит к тому, что возникновение грубых ошибок в наблюдениях зависит от дисперсии σ_ε^2. Подобная тенденция при исследовании экспериментальных данных возникает под действием фактора x_1x_2 (фактор взаимодействия) с соответствующей дисперсией $\sigma^2_{x_1x_2}$ [3].

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Разбиение сумм квадратов

Так как основная задача дисперсионного анализа заключается в оценке влияния факторов на выходную переменную посредством разбиения сумм квадратов отклонений исследуемых данных, выполним данное действие для суммы квадратов от общего среднего:

$$S_0 = \sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(y_{jgl} - \overline{\overline{y}})^2 =$$

$$= \sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(y_{jgl} - \overline{y}_{jg} + \overline{y}_{jg} - \overline{y}_j + \overline{y}_j - \underline{y}'_g + \overline{y}'_g - \overline{\overline{y}} + \overline{\overline{y}} - \overline{\overline{y}})^2 =$$

$$= \sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(y_{jgl} - \overline{\overline{y}})^2 + \sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(\overline{y}_j - \overline{\overline{y}})^2 +$$

$$+\sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(y'_g - \overline{\overline{y}})^2 + \sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(\overline{y}_{jg} - \overline{y}_j - \overline{y}'_g + \overline{\overline{y}})^2 =$$

$$= S_\varepsilon + S_{x_1} + S_{x_2} + S_{x_1x_2}.$$

здесь S_0 – искомая сумма квадратов, характеризующая распределение наблюдений под влиянием фактора взаимодействия;

S_ε – искомая сумма квадратов, характеризующая распределение всех изолированных данных y_{jgl};

S_{x_1} – сумма квадратов отклонений (по строкам таблицы 1). $S_{x_1}/(u_2 m)$ характеризует распределение данных, с учетом влияния «случайного» фактора x_1;

S_{x_2} – сумма квадратов отклонений (по столбцам таблицы 1). $S_{x_2}/(u_1 m)$ характеризует распределение усредненного значения наблюдений по столбцам с учетом влияния «случайного» фактора x_2;

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

$S_{x_1x_2}$ – сумма квадратов отклонений (по строкам и по столбцам таблицы 1). $S_{x_1x_2}/m$ характеризует распределение усредненного значения наблюдений с учетом влияния фактора x_1, x_2 и фактора, характеризующего их взаимодействие [4].

Нахождение и оценка дисперсий

Все перечисленные ранее суммы S_0, S_ε, S_{x_1}, S_{x_2}, $S_{x_1x_2}$, способны предоставить оценку искомой дисперсии, если произвести деление каждой из них на соответствующее число степеней свободы (количество значений в итоговом вычислении данных наблюдений) v_0, v_ε, v_{x_1}, v_{x_2}, $v_{x_1x_2}$ [5]:

1) дисперсия по всем $M = u_1u_2m$ наблюдениям распределения данных вычисляется следующим образом:

$$s_0^2\{y\} = \frac{1}{M-1}\sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\sum_{l=1}^{m}(y_{jgl} - \overline{\overline{y}})^2$$

соответствующие степени свободы $v_0 = u_1u_2m - 1 = M - 1$;

2) дисперсия распределения изолированных наблюдений:

$$s_\varepsilon^2\{y\} = \frac{1}{u_1u_2}\sum_{j=1}^{u_1}\sum_{g=1}^{u_2}\frac{1}{m-1}\sum_{l=1}^{m}(y_{jgl} - \overline{y}_{jg})^2 = \frac{S_\varepsilon}{u_1u_2(m-1)} \approx \sigma_\varepsilon^2$$

соответствующие степени свободы $v_\varepsilon = u_1u_2(m-1)$;

3) дисперсия распределения данных под действием «случайного» фактора x_1:

$$s_{x_1}^2\{y\} = \frac{u_2m}{u_1-1}\sum_{j=1}^{u_1}(\overline{y}_j - \overline{\overline{y}})^2 = \frac{S_{x_1}}{u_1-1} \approx \sigma_\varepsilon^2 + u_2m\sigma_{x_1}^2 + m\sigma_{x_1x_2}^2$$

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

соответствующие степени свободы $v_{x_1} = u_1 - 1$;

4) дисперсия распределения данных под действием «случайного» фактора x_2:

$$s^2_{x_2}\{y\} = \frac{u_1 m}{u_2 - 1}\sum_{g=1}^{u_2}(\overline{y}'_g - \overline{\overline{y}})^2 = \frac{S_{x_2}}{u_2 - 1} \approx \sigma^2_\varepsilon + u_1 m\sigma^2_{x_2} + m\sigma^2_{x_1 x_2}$$

соответствующие степени свободы $v_{x_2} = u_2 - 1$;

5) дисперсия распределения наблюдений с учетом влияния фактора x_1, x_2 и фактора, характеризующего их взаимодействие:

$$s^2_{x_1 x_2}\{y\} = \frac{m}{(u_1 - 1)(u_2 - 1)}\sum_{j=1}^{u_1}\sum_{g=1}^{u_2}(\overline{y}_{jg} - \overline{y}_j - \overline{y}'_g + \overline{\overline{y}})^2$$

соответствующие степени свободы $v_{x_1 x_2} = (u_1 - 1)(u_2 - 1)$.

Также существует специальная проверка правильности расчёта числа степеней свободы [6]:

$$v_0 = v_\varepsilon + v_{x_1} + v_{x_2} + v_{x_1 x_2}.$$

Оценка влияния факторов и их взаимодействия

Наконец, когда произведены все необходимые результаты расчётов, полученные данные необходимо проанализировать и подвести итог. В методе многофакторного дисперсионного анализа используется так называемый критерий Фишера. Если же исследуется влияние только одного фактора – используется критерий Стьюдента [7].

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Последовательность действий следующая:

1) Оценивается действие факторов x_1 и x_2 при помощи соответствующих дисперсий:

$$\sigma_{x_1}^2 \approx \frac{1}{u_2 m}(s_{x_1}^2 - \sigma_\varepsilon^2 - m\sigma_{x_1x_2}^2) \approx \frac{1}{u_2 m}(s_{x_1}^2 - s_{x_1x_2}^2),$$
$$\sigma_{x_2}^2 \approx \frac{1}{u_1 m}(s_{x_2}^2 - \sigma_\varepsilon^2 - m\sigma_{x_1x_2}^2) \approx \frac{1}{u_1 m}(s_{x_2}^2 - s_{x_1x_2}^2)$$

Полученные результаты необходимо правильно интерпретировать и подвести итог: если отличие $s_{x_1}^2$ от $s_{x_1x_2}^2$ и $s_{x_2}^2$ от $s_{x_1x_2}^2$ значимо и

$$F_1 = s_{x_1}^2 \big/ s_{x_1x_2}^2 > F_q(v_{x_1};\, v_{x_1x_2}),\ F_2 = s_{x_2}^2 \big/ s_{x_1x_2}^2 > F_q(v_{x_2};\, v_{x_1x_2}),$$

то говорят, что действие факторов x_1 и x_2 значимо.

2) Оценивается действие фактора x_1x_2 с соответствующей дисперсией:

$$\sigma_{x_1x_2}^2 \approx \frac{1}{m}(s_{x_1x_2}^2 - \sigma_\varepsilon^2) \approx \frac{1}{m}(s_{x_1x_2}^2 - s_\varepsilon^2) \quad .$$

Действие фактора взаимодействия признается значимым, если разница в значениях $s_{x_1x_2}^2$ и s_ε^2 также значима, т. е. если

$$F_{12} = s_{x_1x_2}^2 \big/ s_\varepsilon^2 > F_q(v_{x_1x_2};\, v_\varepsilon).$$

В обратном случае принято считать, что фактор взаимодействия оказывает несущественное влияние [8].

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Общий случай

Наконец, речь пойдет о том, как производить расчеты в случае, когда число факторов больше двух. На самом деле, какого-то существенного отличия двухфакторного дисперсионного анализа от многофакторного нет. Логика рассмотрения данного метода при случае, когда число факторов больше двух лишь усложняется и не претерпевает серьезных изменений [9].

Пусть сначала имеется один фактор и m уровней с количеством опытов u_1, $\ldots u_n$ в сериях соответственно. $M = \sum_{j=1}^{u} u_j$. При этом формулы разбиения сумм принимают следующий вид:

$$S_0 = \sum_{j=1}^{u}\sum_{t=1}^{m} y_{jt}^2 - \frac{1}{um}\left(\sum_{j=1}^{u}\sum_{t=1}^{m} y_{jt}\right)^2 \quad (1)$$

$$S_\varepsilon = \sum_{j=1}^{u}\sum_{t=1}^{m} y_{jt}^2 - \sum_{j=1}^{u}\frac{1}{M}\left(\sum_{t=1}^{m} y_{jt}\right)^2 \quad (2)$$

$$S_x = \sum_{j=1}^{u}\frac{1}{m_j}\left(\sum_{t=1}^{m} y_{jt}\right)^2 - \frac{1}{um}\left(\sum_{j=1}^{u}\sum_{t=1}^{m} y_{jt}\right) \quad (3)$$

Все прочие подсчеты проводятся как в случае равных чисел наблюдений в сериях опыта. Отличие состоит лишь в нахождении числа степеней свободы. Оно будет следующим: $N - m$.

Действие каждого из факторов будем рассматривать отдельно. Чтобы это сделать, необходимо вносить данные наблюдений ячейки в таблицы, которые соответствуют различным уровням данного фактора (с учетом того, что других факторов попросту нет). Затем, после произведенных действий, воспользуемся формулами (1) – (3).

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Для того, чтобы проанализировать полученные результаты и получить информацию о значимости, как каждого фактора в отдельности, так и факторов взаимодействия, необходимо рассматривать экспериментальные наблюдения, соответствующие всевозможным уровням. И если имеется пара факторов, один из которых имеет *m* уровней, а другой r уровней, то создается еще один новый фактор, характеризующий их взаимодействие. Затем снова применяются формулы (1) – (3) [10].

Программная реализация

Разработанная в среде программирования Delphi программа предназначена для выявления значимости влияния факторов по данным наблюдений.

Входными данными программы служат результаты проведенных опытов, которые записываются в матрицу, а также количество факторов. Количество задаваемых опытов варьируется от 3 до 10. Что касается факторов, их можно задавать от 2 до 4. На выходе, программа предоставляет информацию о значимости влияния факторов, как по отдельности, так и их взаимодействия.

Программа разработана на основе представленного выше алгоритма, реализующего метод многофакторного дисперсионного анализа, с помощью критерия Фишера.

На первом шаге, в построенной таблице, при заданном количестве факторов и количестве опытов, вводятся данные наблюдений. Затем, по этим наблюдениям организуется цикл, производящий разбиение сумм. С полученными отсортированными результатами, программа организует следующий цикл, суть которого заключается в оценке и вычислении дисперсий. После произведенных расчетов, программа вычисляет значение критерия Фишера и на последнем шаге, сравнивает его со стандартным табличным значением,

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

при полученных степенях свободы и при заданном уровне значимости. На основании данного сравнения, программа выводит результат, в котором содержится главный итог – оценка значимости факторов и их взаимодействия.

Окно ввода исходных данных (для удобства реализована возможность загрузки данных из файла) выглядит следующим образом:

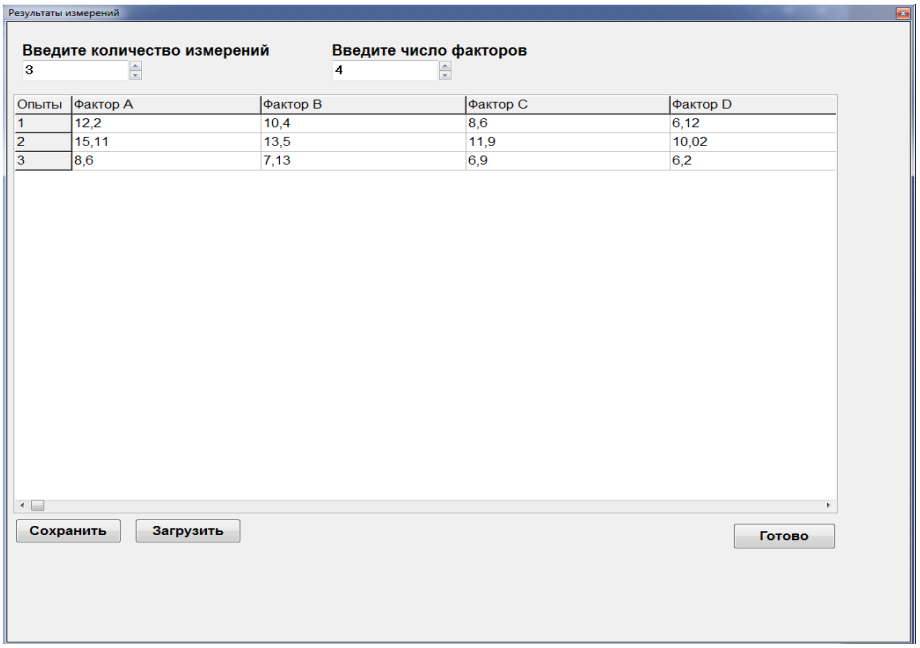

Рис. 1. – Окно ввода данных

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Итоги работы программы:

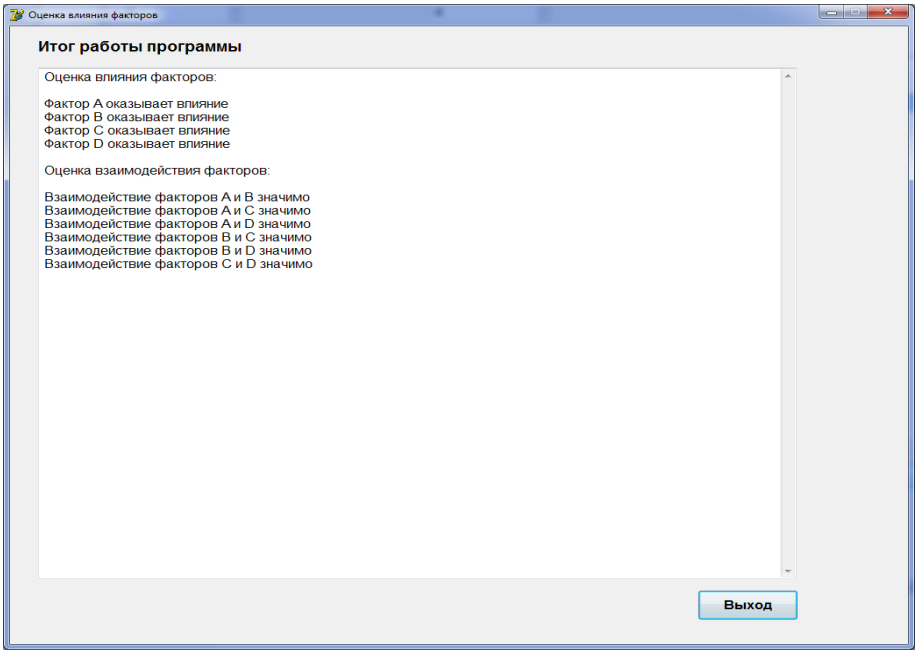

Рис. 2. – результаты наблюдений

Заключение

Дисперсионный анализ является довольно важным и часто используемым статистическим методом, охватывающим задачи биологии, медицины, экономики и других наук. Благодаря современным программным продуктам (таким как Statistica, SPSS, в частности), процесс реализации данного метода занимает небольшое количество времени и усилий. Именно в этом, и в своем большом охвате всевозможных исследований, и состоит его востребованность.

В данной работе была рассмотрена и изучена теоретическая составляющая дисперсионного анализа. На ее основе была разработана модель общего случая и подход к реализации многофакторного анализа в среде программирования Delphi. Созданный алгоритм и интерфейс программы, способен дать нам ответ относительно влияния факторов, а также их взаимодействия, на исследуемые признаки.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

Литература

1. Орлова А.И. Математика случая: Вероятность и статистика – основные факты. Издательство МЗ-Пресс, 2009. 234 с.

2. Ветров А.А., Ломовацкий Г.И. Дисперсионный анализ в экономике. Издательство: Статистика, 2007. 138 с.

3. Сидоренко Е.В. Методы математической обработки в психологии. Издательство: Речь, 2011. 304 с.

4. Ермолаев О.Ю. Математическая статистика для психологов. Издательство: МПСИ Флинта, 2009. 411 с.

5. Яковлев М. Я. Янгирова А. В. Метод и результаты численной оценки эффективных механических свойств резинокордных композитов для случая двухслойного материала. Инженерный вестник Дона. 2013. №2. URL: ivdon.ru/ru/magazine/archive/n2y2013/1639.

6. Уилкс С. Математическая статистика. Издательство: Наука, 2006. 374 с.

7. Magomedov I. A. Mezhieva A.I. Suleymanova M.A. Inženernyj vestnik Dona (Rus). 2018. №4. URL: ivdon.ru/ru/magazine/archive/n4y2018/5334.

8. Rick Turner. Introduction to Analysis of Variance: Design, Analyis & Interpretation. SAGE Publications, Inc, 2001. 192 p.

9. Puntanen, Simo, Styan, George P. H., Isotalo, Jarkko. Matrix Tricks for Linear Statistical Models. Springer. 2011. 208 p.

10. Patrick Doncaster, Andrew Davey. Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences. Cambridge University Press. 2007. 288 p.

References

1. Orlova A.I. Matematika sluchaya: Veroyatnost` i statistika – osnovny`e fakty` [Mathematic of value: Probability and Statistics - Basic Facts]. Izdatel`stvo MZ-Press, 2009. 234 p.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019

2. Vetrov A.A., Lomovaczkij G.I. Dispersionny`j analiz v e`konomike [Analysis of variance in economics]. Izda-tel`stvo: Statistika, 2007. 138 p.

3. Sidorenko E.V. Metody` matematicheskoj obrabotki v psixologii [Methods of mathematical processing in psychology]. Izda-tel`stvo: Rech`, 2011. 304 p.

4. Ermolaev O.Yu. Matematicheskaya statistika dlya psixologov [Mathematical statistics for psychologists]. Izdatel`stvo: MPSI Flinta, 2009. 411 p.

5. Yakovlev M. Y. Yakovleva M. Y. Inženernyj vestnik Dona (Rus). 2012. №2. URL: ivdon.ru/ru/magazine/archive/n2y2013/1639.

6. Uilks S. Matematicheskaya statistika [Mathematical statistics]. Izdatel`stvo: Nauka, 2006. 374 p.

7. Magomedov I. A. Mezhieva A.I. Suleymanova M.A. Inženernyj vestnik Dona (Rus). 2018. №4. URL: ivdon.ru/ru/magazine/archive/n4y2018/5334.

8. Rick Turner. Introduction to Analysis of Variance: Design, Analyis & Interpretation. SAGE Publications, Inc, 2001. 192 p.

9. Puntanen, Simo, Styan, George P. H., Isotalo, Jarkko. Matrix Tricks for Linear Statistical Models. Springer. 2011. 208 p.

10. Patrick Doncaster, Andrew Davey. Analysis of Variance and Covari-ance: How to Choose and Construct Models for the Life Sciences. Cambridge University Press. 2007. 288 p.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2019