УДК 574.587+578.087(282.247.431.2)+595.7(476)

ДИСПЕРСИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ PAST 4.04

Г.Г. Сушко

Учреждение образования «Витебский государственный университет имени П.М. Машерова»

Дисперсионный анализ является одним из основных инструментов для изучения связей многомерных данных. Часто как у молодых исследователей, так и у многих опытных ученых возникают проблемы с выбором метода анализа и особенно с выбором соответствующего пакета анализа данных.

Цель работы – продемонстрировать возможности приложения PAST для анализа экологических данных и, в частности, сравнения нескольких выборок с использованием дисперсионного анализа.

Материал и методы. *Материалом для демонстрации выполнения дисперсионного анализа в PAST 4.04 послужили данные, составляющие выборки, соответствующие и не соответствующие закону нормального распределения. Предложены краткое описание дисперсионного анализа в синэкологических исследованиях и методология использования one-way ANOVA и Kruskal–Wallis H-test.*

Результаты и их обсуждение. *На двух примерах для данных, не соответствующих и соответствующих закону нормального распределения, продемонстрированы методики разведочного анализа данных, собственно анализа, апостериорных сравнений и визуализации результатов.*

Заключение. *Приложение PAST 4.04 обладает всеми необходимыми возможностями для параметрического и непараметрического однофакторного дисперсионного анализа и может быть рекомендовано для выполнения научно-исследовательской работы студентов, магистрантов и аспирантов.*

Ключевые слова: *PAST 4.04, дисперсионный анализ, one-way ANOVA, Kruskal–Wallis H-test.*

ANALYSIS OF VARIANCE USING PAST 4.04 SOFTWARE

G.G. Sushko

Educational Establishment “Vitebsk State P.M. Masherov University”

Analysis of variance is one of the main tools for analyzing relationships of multidimensional data. Often, both young researchers and many experienced scientists have problems with the choice of the analysis method, and especially with the selection of the appropriate data analysis software.

The aim of this work is to demonstrate the capabilities of the PAST software for analyzing of environmental data and, in particular, comparing several samples using analysis of variance.

Material and methods. *Material for demonstrating the performance of analysis of variance in PAST 4.04, the data constituting the sample are appropriate and do not correspond to the law of normal distribution. A brief description of the analysis of variance in synecological studies and the methodology for using one-way ANOVA and Kruskal–Wallis H-test are given.*

Findings and their discussion. *On two examples, for data that do not correspond and those that correspond the law of normal distribution, the technique of exploratory data analysis, analysis of variance, post-hoc tests and visualization of results were demonstrated.*

Conclusion. *The PAST 4.04 software has all the necessary capabilities for parametric and nonparametric analysis of variance and can be recommended for performing research work of students, undergraduates and postgraduate students.*

Key words: *PAST 4.04, analysis of variance, one-way ANOVA, Kruskal–Wallis H-test.*

Современные экологические исследования должны базироваться на четко доказательной основе, которая формируется с использованием статистических методов анализа данных, так как от этого зависят правильность выявленных закономерностей и правомерность сформулированных выводов. Наиболее частыми ошибками являются применение параметрических методов анализа к данным, не соответствующим закону нормального распределения, использование методик, предусматривающих линейные зависимости при их отсутствии. Нужно помнить, что при выполнении стати-

стических анализов имеется ряд условий, которые должны выполняться, и с ними необходимо ознакомиться в специальной литературе. В частности, при проведении дисперсионного анализа и построении регрессионных моделей требуется так называемый разведочный анализ, который может занимать больше времени, чем собственно сам анализ данных.

По ряду причин у большинства молодых исследователей (магистрантов, аспирантов), да и у многих опытных ученых возникают проблемы с выбором метода анализа и особенно с выбором соответствующего пакета анализа данных. В первом случае трудности возникают в связи с недостаточностью теоретической базы, так как большинство учебников написаны на английском языке или требуют достаточно высокого уровня подготовки для восприятия материала [1–2]. Классические пособия по биометрии на русском языке в основном содержат теоретические основы статистического анализа с подробным описанием их математического аппарата [3–5]. Но, к сожалению, в них отсутствует описание инструментов для реализации расчетов с применением современных компьютерных технологий. Во втором случае имеющееся лицензионное программное обеспечение, в частности STATISTICA и SPSS, дорогостоящее, а для отдельных пакетов требуются навыки программирования. В частности для работы в наиболее популярной в настоящее время статистической среде R нужно писать программный код для соответствующего типа анализа.

PAST (PAlaeontological STatistics) – это бесплатное программное обеспечение для анализа научных данных с функциями для обработки данных, построения графиков, одномерной и многомерной статистики, анализа биоразнообразия, анализа временных рядов, морфометрии и стратиграфии и др. К сожалению, приложение не имеет поддержки на русском языке, но это компенсируется тем, что оно не требует установки (можно работать даже с флэш-накопителя), быстро запускается и достаточно компактно. Программное обеспечение разработал сотрудник Музея естественной истории (Natural History Museum) города Осло (Норвегия) Øyvind Hammer [6]. Несмотря на то, что изначально приложение было предназначено для палеонтологических исследований, оно получило широкую популярность во всем мире среди экологов. На сегодняшний день используется версия 2020 года PAST 4.04, значительно расширенная и дополненная новыми возможностями.

В настоящее время дисперсионный анализ является одним из основных инструментов для анализа связей многомерных данных. В частности, в синэкологических исследованиях возникает необходимость выявить достоверные различия между несколькими выборками, которые позволяют продемонстрировать закономерности распределения видов в пространстве (градиенте местообитаний) и во времени, различия видового состава, численности разнообразия, морфологии, организовать правильные подходы к охране редких видов и ценных местообитаний и многое другое.

В связи с вышесказанным цель данной работы – продемонстрировать возможности приложения PAST для анализа экологических данных и, в частности, сравнения нескольких выборок с использованием дисперсионного анализа.

Материал и методы. Материалом для демонстрации выполнения дисперсионного анализа в PAST 4.04 послужили данные, составляющие выборки, соответствующие и не соответствующие закону нормального распределения. В частности, измерения высоты трех видов кустарничков (багульника, хамедафны и голубики), выполненные автором, и число видов животных, нахождение которых гипотетически предполагается в трех условных биотопах. В таблице для первого примера исходные данные приводятся в столбцах, где указана высота трех видов кустарничков (багульника, хамедафны и голубики), измеренная в 20-кратной повторности. Во втором примере в таблице исходных данных, также в столбцах, указано число видов гипотетических животных, зарегистрированных в трех биотопах в результате 10 учетов в каждом из них.

Краткое описание возможностей дисперсионного анализа в синэкологических исследованиях. В англоязычной литературе дисперсионный анализ обычно называется анализом вариации (Analysis of variance). В его основе лежит анализ отклонений всех единиц исследуемой совокупности от среднего арифметического. В качестве меры отклонений берется дисперсия или средний квадрат отклонений. Отклонения, вызываемые воздействием данного фактора, сравниваются с величиной отклонений, вызываемых случайными факторами (т.е. теми, которые не учтены исследователем, но тоже могут оказывать влияние). Если отклонения, вызываемые исследуемым

фактором, более существенны, чем случайные отклонения, то считается, что этот фактор оказывает значимое влияние [1].

Если рассматривается более двух независимых выборок, то обычно используется однофакторный дисперсионный анализ (one-way ANOVA) в случае соответствия данных нормальному распределению или его непараметрические аналоги (тесты Краскелла–Уоллеса, Фридмана и др.) – в противном случае. При этом нулевая гипотеза утверждает, что между выборками различий нет, тогда как альтернативная – различия есть хотя бы между двумя выборками [1; 2].

Как правило, возникает вопрос: почему анализ однофакторный, а сравнивается более двух выборок? Если имеется только одна зависимая переменная, например, число видов в более чем двух биотопах, мы имеем дело с однофакторным дисперсионным анализом (one-way ANOVA). Так, зависимая переменная – число видов, а несколько независимых – типы биотопов, в которых они отмечены (допустим три). Для исследования различий при влиянии двух одновременно действующих независимых переменных на зависимую переменную применяется двухфакторный дисперсионный анализ (two-way ANOVA). Например, если мы имеем дело с такими переменными, как температура и влажность, при однофакторном анализе оценивается их влияние на зависимую переменную по отдельности, тогда как при двухфакторном – еще и их совместное влияние. Многомерный дисперсионный анализ (multivariate analysis of variance, MANOVA) выполняется при наличии более чем одной зависимой переменной [2].

Условия применения параметрического однофакторного дисперсионного анализа (one-way ANOVA). Первое необходимое условие – соответствие данных закону нормального распределения. Вторым необходимым условием однофакторного дисперсионного анализа является однородность (гомоскедастичность) групповых дисперсий (homogeneity of variance, или homoscedasticity of variance) [7]. То есть кроме нормального распределения в каждой группе, значения зависимой переменной должны также иметь одинаковую степень разброса. Необходимость выполнения этого условия определяется способом вычисления внутри- и межгрупповых дисперсий, применяемым в дисперсионном анализе: при значительно различающихся групповых дисперсиях используемые формулы просто не будут работать корректно. Если эти условия не выполняются, нужно применять непараметрические аналоги ANOVA.

Способы проверки однородности групповых дисперсий включают как графические, так и формальные методы. Визуальную оценку однородности групповых дисперсий можно выполнить при помощи диаграмм размахов по размерам интерквартильных размахов (если высота «ящиков» сильно различается, то дисперсия неоднородна, и наоборот). Условие однородности дисперсий можно проверить формально при помощи соответствующих тестов. Одним из наиболее часто используемых является тест Левене (Levene's test). В случае гомогенности дисперсий p-уровень теста Левене должен быть больше 0,05. Если статистика Левене является значимой на уровне 0,05, нулевая гипотеза о том, что группы имеют одинаковую дисперсию, отвергается.

Кроме того, выборки должны быть независимыми между собой, а также независимыми должны быть и наблюдения в каждой из выборок.

Важный этап – анализ остатков (residuals). Остатки характеризуют внутригрупповую дисперсию (ее еще называют флуктуирующей или остаточной дисперсией), которая не может быть объяснена влиянием анализируемого экспериментального фактора и является следствием случайной флуктуации данных или действием другого фактора [7].

Интерпретация итоговой таблицы дисперсионного анализа. На начальном этапе анализа, независимо, в каком пакете он выполнен, мы получаем таблицу данных, в которой в первую очередь нужно обратить на p-уровень. Если его значение < 0,05, анализируемые выборки различаются. В таблице будет ряд показателей, которые мы охарактеризуем ниже.

SS (Sum of Squares) – сумма квадратов (отклонений) внутри группы, MS (Mean of Squares) – оценки дисперсии между группами, F – значение критерия Фишера (внутри- и межгрупповые дисперсии сравниваются при помощи F-критерия), *df* (degrees of freedom) – число степеней свободы, p – p-уровень [1; 2; 5].

Апостериорные сравнения (post-hoc tests). Поскольку дисперсионный анализ показывает наличие или отсутствие различий между сравниваемыми переменными, с его помощью нельзя узнать, какие именно группы признаков различаются. Для этого предусмотрены множественные попарные сравнения средних величин. Одним из наиболее популярных методов таких сравнений является критерий Тьюки, или критерий достоверно значимой разности Тьюки (Tukey's honestly significant difference test, или просто Tukey's HSD test). Если между парами сравниваемых групп есть достоверные различия, значения p-уровня будут меньше 0,05. Апостериорные сравнения можно продублировать графически. Полученные результаты теста хорошо согласуются с визуальной оценкой различий на диаграммах размахов [1].

Условия применения непараметрического однофакторного дисперсионного анализа. Критерий Краскелла–Уоллеса (Kruskal–Wallis H-test) – непараметрический аналог однофакторного дисперсионного анализа, позволяющий выявить различия между несколькими независимыми выборками, если данные не соответствуют закону нормального распределения и (или) не соблюдается однородность групповых дисперсий. Как и в большинстве непараметрических методов для количественных данных, исходные данные преобразуются в ранги и производится обработка уже рангов.

Подобно классическому дисперсионному анализу, тест Краскела–Уоллиса помогает сделать заключение о том, что сравниваемые группы статистически значимо различаются (при $p < 0,05$) либо статистически значимых различий между группами нет ($p > 0,05$). Чтобы выяснить, между какими группами есть отличия, также выполняют апостериорные тесты. Наиболее часто используют непараметрический критерий Данна (Dunn post hoc test, Dunn's multiple comparison test). Критерий применим для независимых групп как равной, так и различной величины [1].

Апостериорные сравнения, как правило, проводятся с помощью критерия Стьюдента для независимых выборок. Это может вызвать недоумение, так как главное ограничение t-критерия перед дисперсионным анализом состоит в том, что он используется для парных сравнений (когда есть только 2 группы).

Однако в отличие от простых попарных сравнений с помощью t-критерия при апостериорных сравнениях рассчитываются новые критические p-уровни для предотвращения ошибки 1 типа более 5%. Наиболее частый способ коррекции ошибки 1 типа – поправка Бонферрони (Bonferroni correction). При ее использовании уровень ошибки 1 типа делится на количество сравнений различных пар признаков. Например, если имеется 3 сравнения, то новый критический уровень должен быть получен делением 0,05 на 3 и составляет 0,017.

Критерий Фридмана. Данный критерий является непараметрической альтернативой однофакторному дисперсионному анализу для зависимых выборок с повторными измерениями. Например, если сравнивается несколько групп данных, полученных на одних и тех же объектах спустя определенный промежуток времени [2].

Результаты и их обсуждение. ***Однофакторный дисперсионный анализ (one-way ANOVA) для параметрических данных.*** В таблице исходных данных, которые приводятся в столбцах, указана высота трех видов кустарничков (багульника, хамедафны и голубики), измеренная в 20-кратной повторности. Нужно выявить различие или сходство высоты этих растений одной жизненной формы, произрастающих в одном типе леса.

Разведочный анализ данных. Загружаем таблицу с данными из файла MS Exsel в PAST 4.04 и выполняем проверку на нормальность распределения с использованием вкладки **Univariate/Normality tests.** При проверке на нормальность оказалось, что отдельные данные близки, но не соответствуют закону нормального распределения. Для приведения данных в соответствие с условиями анализа выполняем логарифмирование. После логарифмирования (вкладка Transform/Log) выполняется условие соответствия закону нормального распределения, о чем свидетельствуют значения соответствующих тестов и, в частности, теста Шапиро–Уилка ($p > 0,05$) (рис. 1).

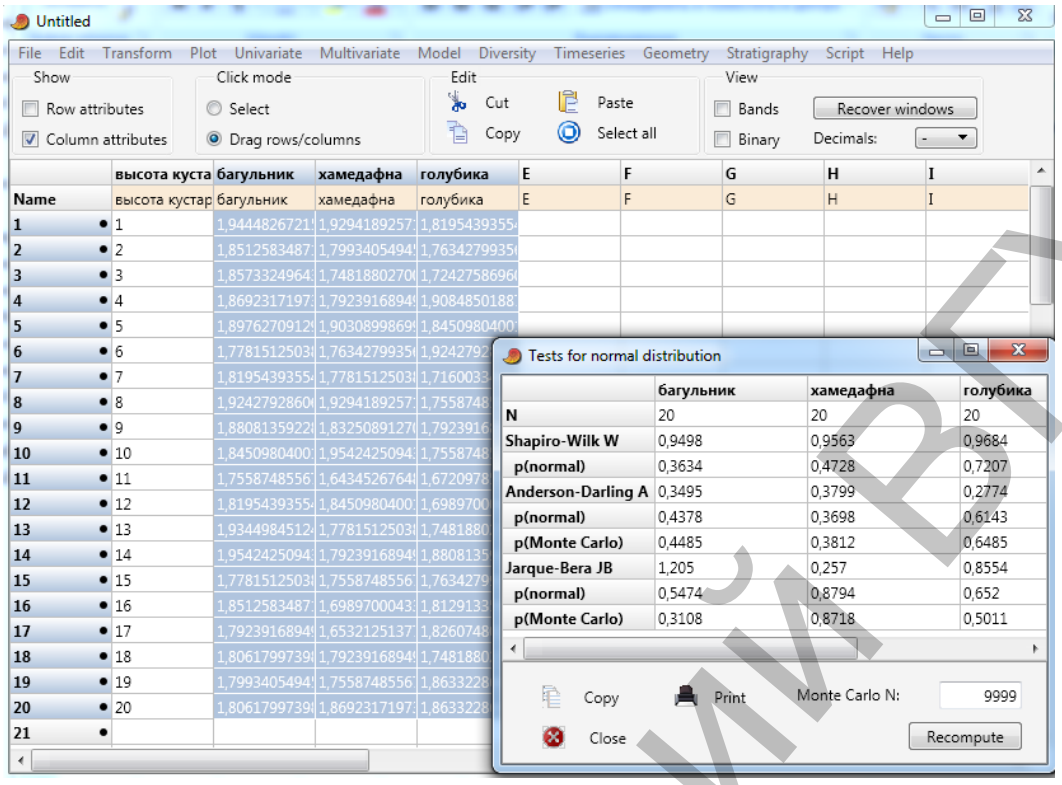


Рис. 1. **Проверка данных на соответствие закону нормального распределения**

Далее нужно убедиться, что выполняется условие однородности дисперсии. Для этого нужно использовать вкладку **Univariate/ANOVA etc. (several samples)/Several sample tests (ANOVA, Kruskal–Wallis)/ANOVA** (собственно, это и есть панель дисперсионного анализа) и обратить внимание на значения теста Левене (Levene's test). В нашем примере p-уровень данного теста превышает 0,05 (p=0,256). Следовательно, и второе условие выполняется (рис. 2). Поэтому мы можем провести параметрический дисперсионный анализ (one-way ANOVA).

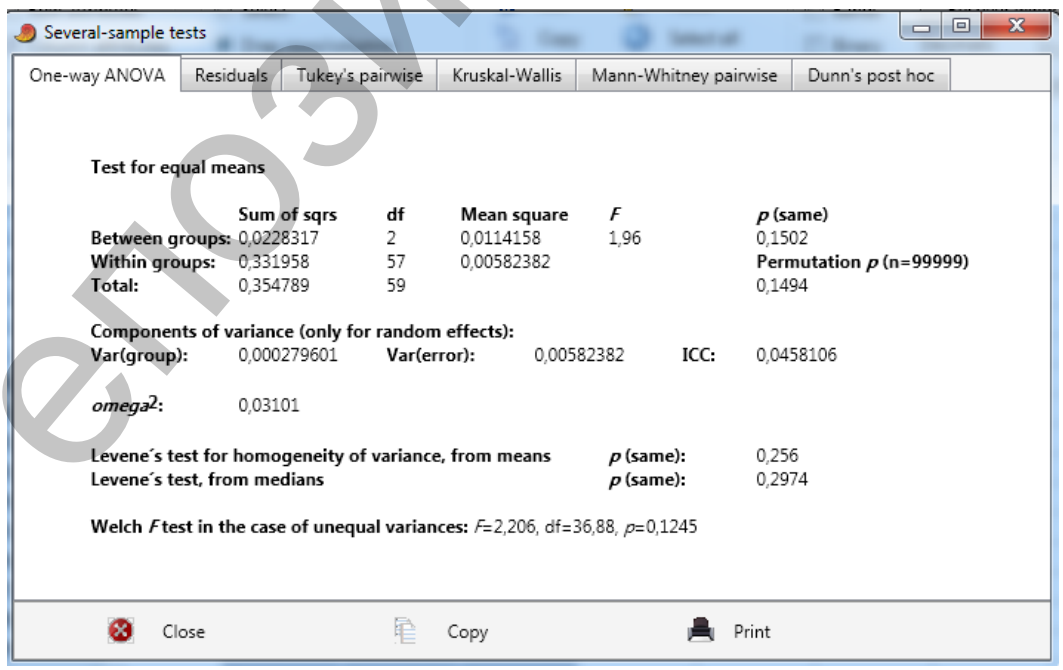


Рис. 2. **Результаты дисперсионного анализа (one-way ANOVA)**

Дисперсионный анализ. Используя ту же вкладку **Univariate/ANOVA etc. (several samples)/Several sample tests (ANOVA, Kruskal–Wallis)/ANOVA,** анализируем данные таблицы результатов Test for equal means. Как видно из рис. 2, F=1,96, p=0,150. Следовательно, между высотой кустарничков достоверные отличия отсутствуют. Перестановочный тест (Permutation test) подтверждает эти результаты (p=0,149).

Апостериорные сравнения. Выполняем с применением вкладки **Univariate/ANOVA etc. (several samples)/Several sample tests (ANOVA, Kruskal–Wallis)/Tukey's pairwise.** Тест Тьюки также не выявил значимых различий между значениями конкретных переменных при их попарных сравнениях: p>0,05 (рис. 3).

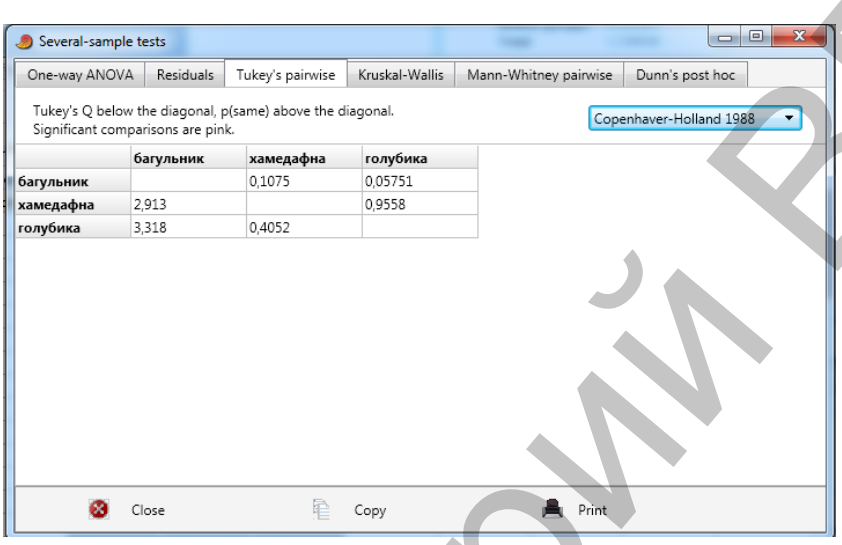


Рис. 3. **Результаты апостериорных сравнений дисперсионного анализа (one-way ANOVA)**

Визуализация результатов анализа. Поскольку дисперсионный анализ показывает только наличие или отсутствие различий между переменными, но не отражает их величину, разнородность переменных можно отобразить на графике. В большинстве случаев используют диаграммы размаха («ящики с усами»). Их можно построить с помощью вкладки Plot/Barchart/Box plot (рис. 4).

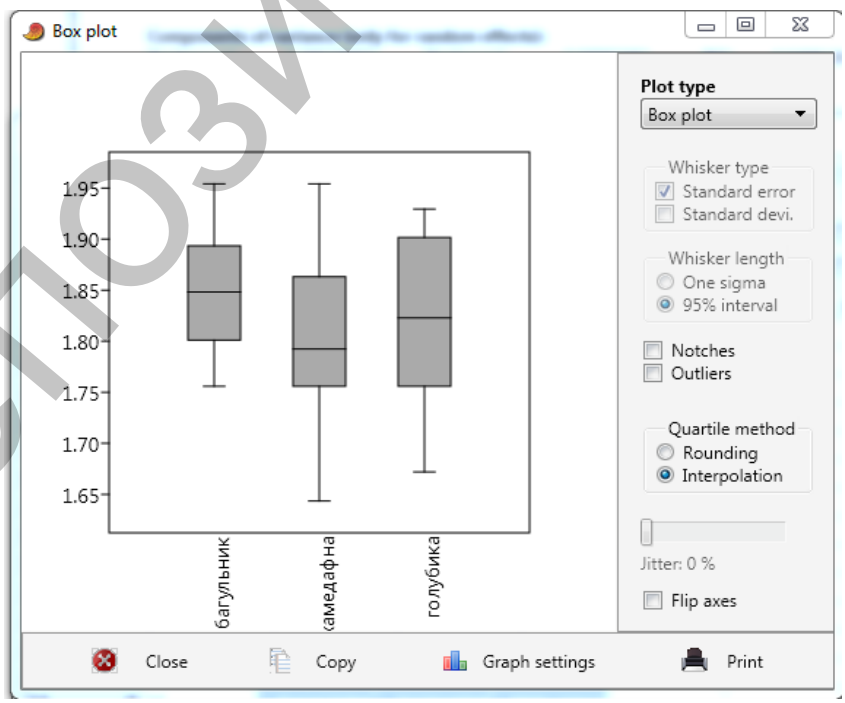


Рис. 4. **Демонстрация различий между переменными**

Как видно из рис. 4, на первый взгляд, различия видны, по крайней мере, по размерам «ящиков», однако они статистически не достоверны. Кроме того, если смотреть на «усы», то они расположены примерно на одном уровне, что также свидетельствует об отсутствии различий.

Анализ остатков. Для оценки достоверности выполненного анализа исследуем остатки модели, которые должны соответствовать закону нормального распределения с помощью вкладки Residuals. Как видно из диаграммы рассеяния и значения теста Шапиро–Уилка (p>0,05), данное условие выполняется и можно утверждать о достоверности результатов дисперсионного анализа (рис. 5).

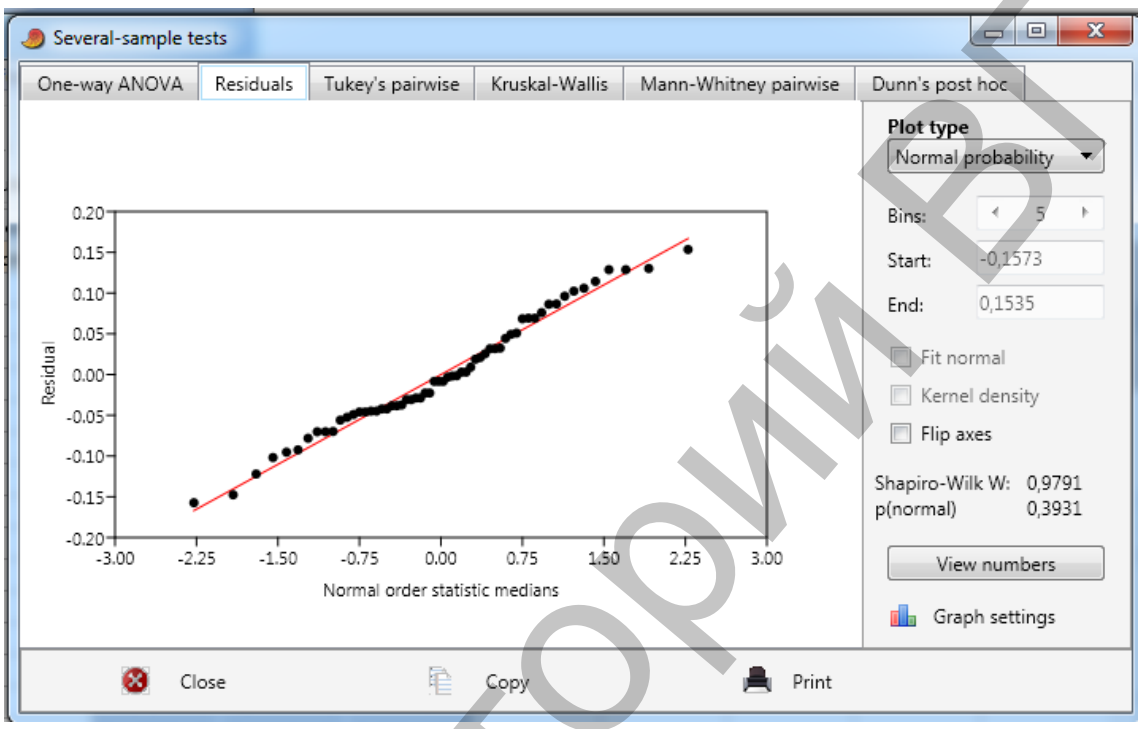


Рис. 5. **Визуализация остатков**

Однофакторный дисперсионный анализ (Kruskal–Wallis test) для непараметрических данных. В таблице исходных данных, которые приводятся в столбцах, указано число видов гипотетических животных, зарегистрированных в трех биотопах в результате 10 учетов в каждом из них. Нужно установить различие или сходство видового богатства этих 3 местообитаний.

Разведочный анализ данных. Загружаем таблицу с данными из файла MS Exsel в PAST 4.04 и выполняем проверку на нормальность распределения с использованием вкладки **Univariate/Normality tests.** При проверке на нормальность оказалось, что отдельные данные близки, но не соответствуют закону нормального распределения. После логарифмирования картина не изменилась, о чем свидетельствуют значения соответствующих тестов и, в частности, теста Шапиро–Уилка (p<0,05) (рис. 6). Следовательно, необходимо прибегнуть к выполнению непараметрического дисперсионного анализа.

Так как данные не соответствуют закону нормального распределения, проверка на однородность дисперсии не требуется.

Дисперсионный анализ. Для выполнения дисперсионного анализа применена вкладка **Univariate/ANOVA etc. (several samples)/Several sample tests (ANOVA, Kruskal–Wallis)/Kruskal–Wallis**. Как видно из рис. 7, значения теста Краскела–Уоллиса (H=24,82, p<0,05) свидетельствуют о наличии значимых различий видового богатства трех исследуемых биотопов.

Апостериорные сравнения. Выполняем с использованием вкладки **Univariate/ANOVA etc. (several samples)/Several sample tests (ANOVA, Kruskal–Wallis)/Dunn's post hoc.** Тест Данна выявил значимые различия между значениями всех переменных при их попарных сравнениях: p<0,05 (рис. 8).

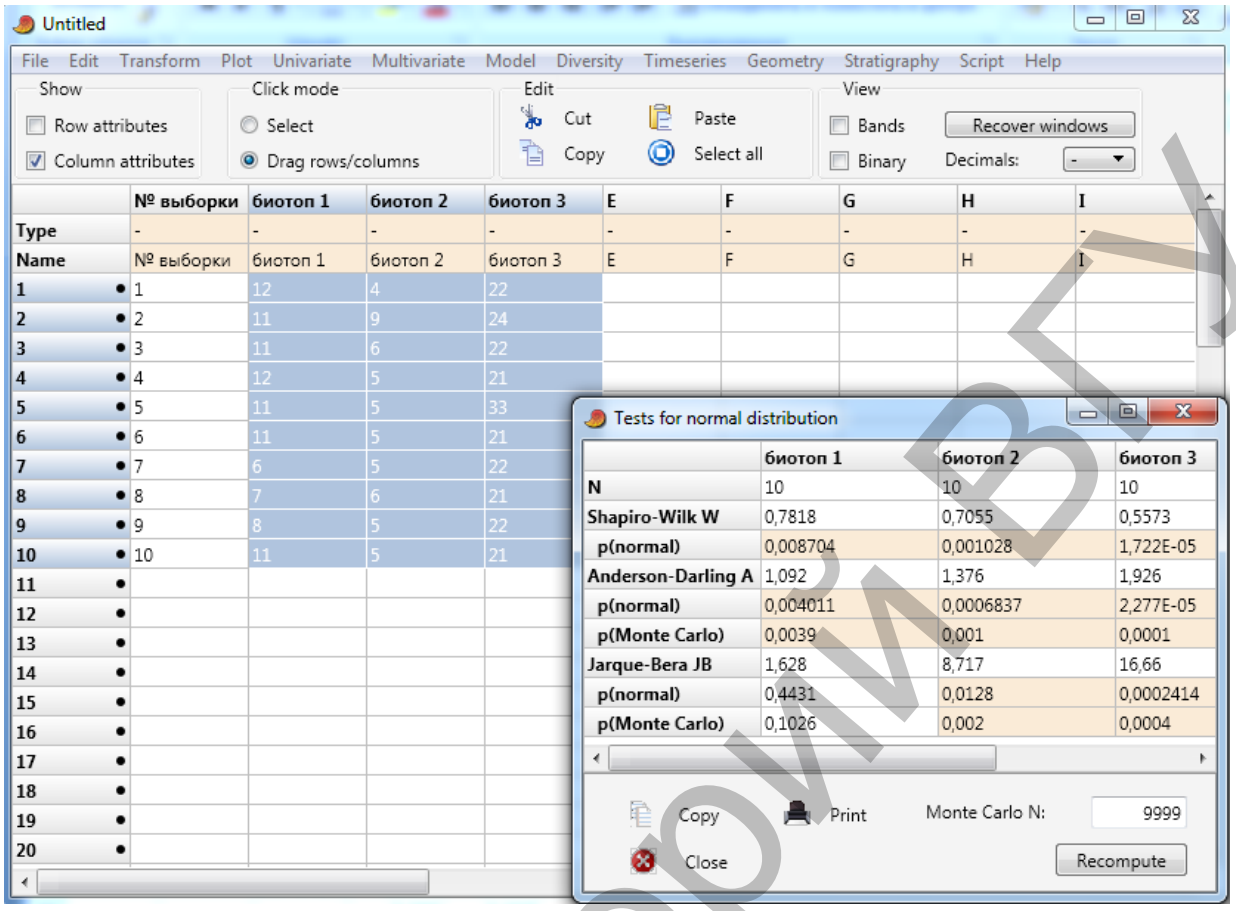


Рис. 6. **Проверка данных на соответствие закону нормального распределения**

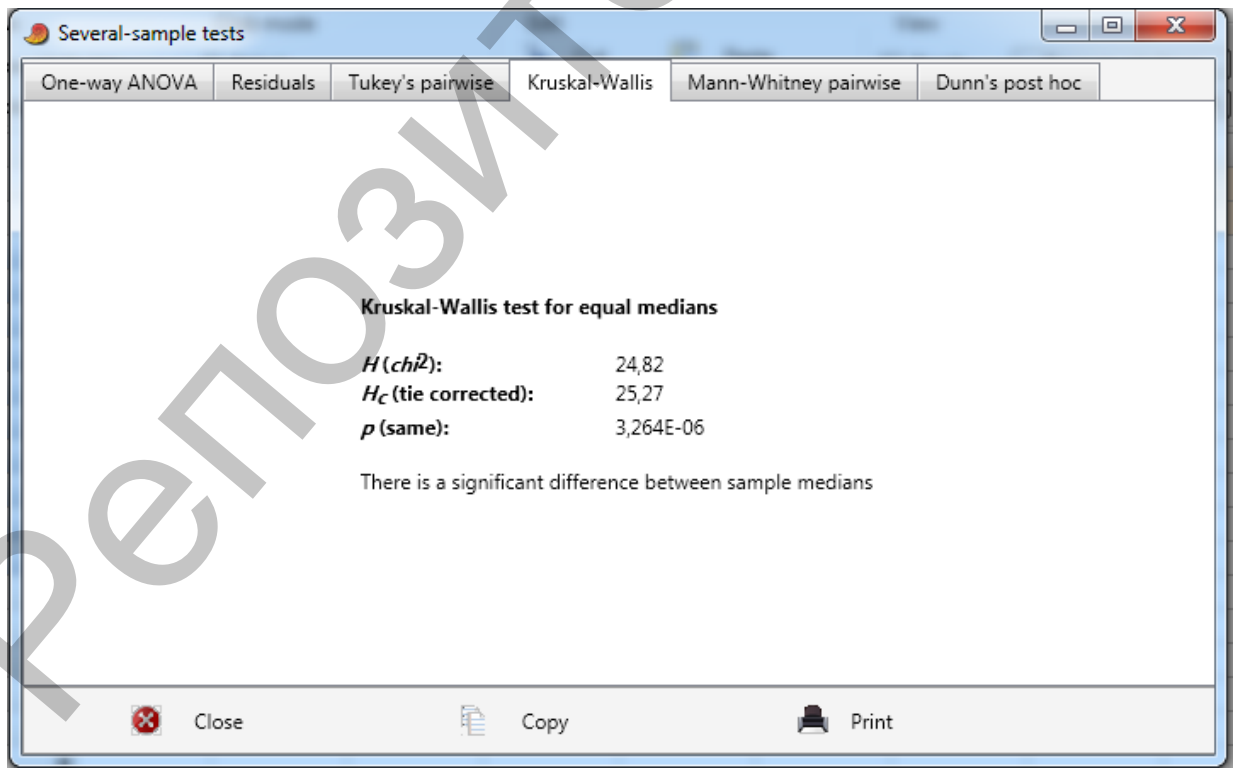
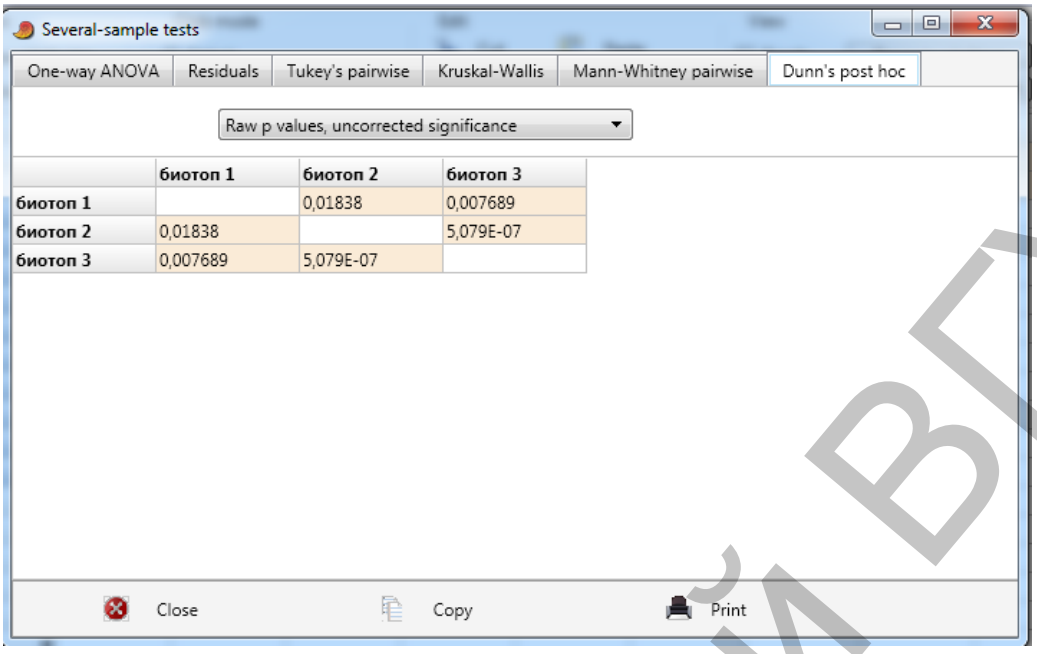


Рис. 7. **Результаты дисперсионного анализа (Kruskal–Wallis test)**



	биотоп 1	биотоп 2	биотоп 3
биотоп 1		0,01838	0,007689
биотоп 2	0,01838		5,079E-07
биотоп 3	0,007689	5,079E-07	

Рис. 8. **Результаты апостериорных сравнений дисперсионного анализа (Kruskal–Wallis test)**

Визуализация результатов анализа. Поскольку дисперсионный анализ показывает только наличие или отсутствие различий между переменными, но не отражает их величину, разнородность переменных можно отобразить на диаграммах размаха, построив их с помощью вкладки Plot/Barchart (рис. 9).

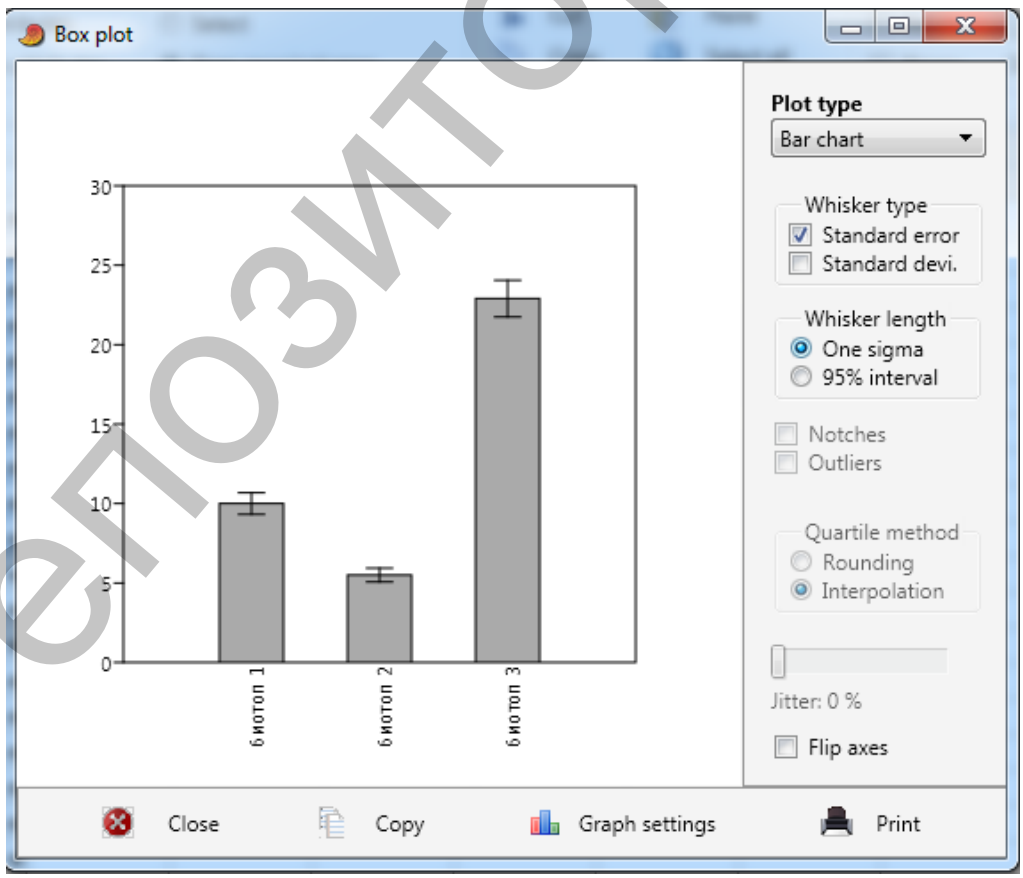


Рис. 9. **Демонстрация различий между переменными**

Как видно из рис. 9, различия установлены и указывают на достоверно более высокое число видов в биотопе 3 и минимальное – в биотопе 2.

Заключение. Таким образом, приложение PAST 4.04 обладает всеми необходимыми возможностями для параметрического и непараметрического однофакторного дисперсионного анализа и может быть рекомендовано для выполнения научно-исследовательской работы студентов, магистрантов и аспирантов.

ЛИТЕРАТУРА

1. McCune, B. Analysis of ecological communities / B. McCune, J.B. Grace. – Gleneden Beach: MjMSoftware Design, 2002. – 304 p.
2. Borcard, D. Numerical Ecology with R / D. Borcard, F. Gillet, P. Legendre. – Wiena: Springer Nature, 2018. – 435 p.
3. Плохинский, Н.А. Биометрия / Н.А. Плохинский. – 2-е изд. – М.: Изд-во МГУ, 1970. – 367 с.
4. Рокицкий, П.Ф. Биологическая статистика / П.Ф. Рокицкий. – Минск: Вышэйшая школа, 1973. – 348 с.
5. Лакин, С.Ф. Биометрия: учеб. пособие для биологических вузов / С.Ф. Лакин. – М.: Высшая школа, 1990. – 352 с.
6. Hammer, Ø. PAST: Paleontological Statistics Software Package for Education and Data Analysis / Ø. Hammer, D.A.T. Harpe, R.D. Ryan // Palaeontologica Electronica. – 2001. – Vol. 4, № 1. – P. 9.
7. Zuur, A.F. A protocol for data exploration to avoid common statistical problems / A.F. Zuur, E.N. Ieno, C.S. Elphick. – Methods of Ecology and Evolution. – 2010. – Vol. 1. – P. 3–14.

REFERENCES

1. McCune, B. Analysis of ecological communities / B. McCune, J.B. Grace. – Gleneden Beach: MjMSoftware Design, 2002. – 304 p.
2. Borcard, D. Numerical Ecology with R / D. Borcard, F. Gillet, P. Legendre. – Wiena: Springer Nature, 2018. – 435 p.
3. Plokhinsky N.A. *Biometriya. 2-ye izdaniye* [Biometrics. 2nd Edition], Moscow, MGU, 1970, 367 p.
4. Rokitsky P.F. *Biologicheskaya statistika* [Biological Statistics], Minsk, Vysheyshaya shkola, 1973, 348 p.
5. Lakin S.F. *Biometriya: uchebnoye posobiye dlia biologicheskikh vuzov* [Biometrics: a Textbook for Biological Universities], Moscow, Vysshaya shkola, 1990, 352 p.
6. Hammer, Ø. PAST: Paleontological Statistics Software Package for Education and Data Analysis / Ø. Hammer, D.A.T. Harpe, R.D. Ryan // Palaeontologica Electronica. – 2001. – Vol. 4, № 1. – P. 9.
7. Zuur, A.F. A protocol for data exploration to avoid common statistical problems / A.F. Zuur, E.N. Ieno, C. Elphick. – Methods of Ecology and Evolution. – 2010. – Vol. 1. – P. 3–14.

Поступила в редакцию 23.12.2020

Адрес для корреспонденции: e-mail: gennadis@rambler.ru – Сушко Г.Г.