

Федеральное государственное бюджетное учреждение науки
Институт проблем экологии и эволюции им. А.Н. Северцова Российской академии наук



РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Статистический анализ данных в биологии»
(наименование дисциплины)

Направление подготовки:

06.06.01 – Биологические науки

(указывается код и наименование направления подготовки)

Направленность (профиль) подготовки:

«Зоология», «Энтомология», «Ихтиология», «Экология»
(указывается наименование направленности)

«Гидробиология», «Паразитология»

Квалификация: **Исследователь. Преподаватель-исследователь.**

Москва, 2015 г.

Программа составлена в соответствии с утвержденным Федеральным государственным образовательным стандартом высшего образования по направлению подготовки 06.06.01 Биологические науки (уровень подготовки кадров высшей квалификации) – Приказ Министерства образования и науки Российской Федерации № 871 от 30.07.2014 г., зарегистрировано в Минюсте Российской Федерации 20 августа 2014 года № 33686.

Автор: к.б.н. Васильева Н.А.

Программа одобрена на заседании Ученого совета ИПЭЭ РАН, протокол №9 от 5 ноября 2015 года.

Согласовано:

Зам. директора ИПЭЭ РАН по научной работе

А.В. Суров

Отв. за аспирантуру

М.В. Кропоткина

Аннотация

Дисциплина «Статистический анализ данных в биологии» реализуется в рамках основной профессиональной образовательной программы высшего образования - программы подготовки научно-педагогических кадров в аспирантуре Федерального государственного бюджетного учреждения науки Института проблем экологии и эволюции им. А.Н. Северцова Российской академии наук (ИПЭЭ РАН) по направлению подготовки 06.06.01 Биологические науки, по направленностям (профилям) программы Зоология, Энтомология, Ихтиология, Экология, Гидробиология, Паразитология аспирантам очной и заочной форм обучения. Рабочая программа разработана с учетом требований ФГОС по направлению подготовки 06.06.01 Биологические науки, утвержденного приказом Министерства образования и науки Российской Федерации № 871 от 30.07.2014 г., зарегистрировано в Минюсте Российской Федерации 20 августа 2014 года № 33686.

Основным источником материалов для формирования программы являются: учебники, монографические издания, публикации, материалы конференций, симпозиумов, семинаров, интернет-ресурсы. Общая трудоемкость дисциплины по учебному плану составляет 3 зачетные единицы (108 академических часа), из них лекций –36 часов, практических занятий – 36 часов и 34 часа самостоятельной работы (выполнение домашней работы, выполнение учебного проекта по анализу симулированных данных). Дисциплина реализуется на 2 году обучения. Текущая аттестация проводится не менее 2 раз в соответствии с заданиями и формами контроля, предусмотренные настоящей программой. Промежуточная оценка знания осуществляется в форме экзамена (4 академических часа).

Место дисциплины в структуре основной профессиональной образовательной программы высшего образования - программы подготовки научно-педагогических кадров в аспирантуре: Дисциплина «Статистический анализ данных в биологии» является обязательной дисциплиной вариативной части ООП. Ее целью является овладение теоретическими основами и инструментарием современного математического анализа данных, в частности, в пакете программ Statistica (StatSoft, Tulsa, OK, USA) и в программной среде для обработки данных R (R Development Core Team).

Компетенции обучающегося, формируемые в результате освоения дисциплины

универсальные компетенции:

способность к критическому анализу и оценке современных научных достижений, в частности, в отношении валидности результатов публикуемых исследований; умение генерировать новые идеи при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);

общеобразовательные компетенции:

способность самостоятельно осуществлять научно- исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследований и информационно- коммуникационных технологий (ОПК-1);

профессиональные компетенции:

способность к критической оценке опубликованных данных в области зоологии и смежных дисциплин (ПК-2)

способность проводить анализ научных фактов в области зоологии, самостоятельно ставить задачу исследования для решения актуальных проблем зоологии и реализовывать исследовательские протоколы на практике (ПК-3)

способность к критической оценке опубликованных данных в области энтомологии и смежных дисциплин (ПК-5)

способность проводить анализ научных фактов в области энтомологии, самостоятельно ставить задачу исследования для решения актуальных проблем энтомологии и реализовывать исследовательские протоколы на практике (ПК-6)

способность к критической оценке опубликованных данных в области ихтиологии и смежных дисциплин (ПК-8)

способность проводить анализ научных фактов в области ихтиологии, самостоятельно ставить задачу исследования для решения актуальных проблем ихтиологии и способность реализовывать исследовательские протоколы на практике (ПК-9)

способность к критической оценке опубликованных данных в области экологии и смежных дисциплин (ПК-11)

способность проводить анализ научных фактов в области экологии, самостоятельно ставить задачу исследования для решения актуальных проблем экологии и способность реализовывать исследовательские протоколы на практике (ПК-12)

способность к критической оценке опубликованных данных в области гидробиологии и смежных дисциплин (ПК-14)

способность проводить анализ научных фактов в области гидробиологии, самостоятельно ставить задачу исследования для решения актуальных проблем гидробиологии и способность реализовывать исследовательские протоколы на практике (ПК-15)

способность к критической оценке опубликованных данных в области паразитологии и смежных дисциплин (ПК-17)

способность проводить анализ научных фактов в области паразитологии, самостоятельно ставить задачу исследования для решения актуальных проблем паразитологии и реализовывать исследовательские протоколы на практике (ПК-18)

способность к комплексному и систематическому анализу полученных научно-исследовательских результатов для формирования и развития собственной тематики исследований и представления их в современных рейтинговых формах (ПК-19)

В результате изучения дисциплины «Статистический анализ данных в биологии» аспирант должен достичь следующих результатов обучения:

Знать:

требования к выборкам и ограничения, которые задаются предполагаемыми методами анализа данных, основные теоретические принципы анализа данных, тестирования статистических гипотез, принципы построения аналитических моделей, подбора оптимальных моделей

Уметь:

подготавливать данные для последующего анализа, формировать выборки под конкретные исследовательские задачи, производить анализ данных с использованием специализированного программного обеспечения.

Владеть:

представлениями о статистических методах и критериях, используемых в современных биологических исследованиях, об их назначении, ограничениях и об интерпретации результатов статистических тестов; современными методами статистического анализа данных.

Структура дисциплины:

Вид занятий	Количество часов
Лекции	36
Семинары	
Лабораторно-практические занятия	36
Самостоятельная работа	34
Экзамен	4
ИТОГО	108

Содержание дисциплины:

№	Наименование темы (раздела)	Краткое содержание темы (раздела)	Объем темы (раздела), ак.ч.				
			Л	С	ПЗ	СР	Итого
1	Основные понятия в анализе биологических данных. Описательная статистика.	Статистика как инструмент для количественного анализа и интерпретации данных. Описательная и индуктивная статистика. Понятия выборки и измерения. Принципы сбора данных для статистического анализа. Свойства переменных. Частотное распределение переменной. Графическое представление частотного распределения. Характеристики распределения (центр, разброс и форма распределения). Нормальное распределение. Площадь нормального распределения. Распределение выборочных средних.	2				2
2	Тестирование гипотез в статистике. Критерии Стьюдента	Понятие статистики критерия. Понятие статистической гипотезы, принципы формулирования взаимоисключающих гипотез. Процедура тестирования гипотез в статистике, статистические ошибки. Критерии Стьюдента: одновыборочный, двухвыборочный для независимых выборок, двухвыборочный для связанных выборок.	2		2	2	6
3	Дисперсионный анализ (ANOVA). Мощність	Дисперсионный анализ. Связь с двухвыборочным t-критерием Стьюдента. Рекомендации и требования к выборкам.	4		4	4	12

	статистического теста. Размер эффекта и практическая значимость результатов.	Однофакторный дисперсионный анализ. Эффект множественных сравнений. Апостериорные тесты (Бонферрони, Тьюки, Ньюмена-Кейлса, Даннета, Шеффе). Анализ контрастов. Дисперсионный анализ для связанных выборок. Многофакторный дисперсионный анализ. Модели дисперсионного анализа. Модель смешанных эффектов. Гнездовой (Nested) дисперсионный анализ. Многомерный дисперсионный анализ (MANOVA). Понятие мощности статистического критерия. Практическая интерпретация результатов (размер эффекта). Оценка необходимого размера выборки для предполагаемого статистического анализа.					
4.	Корреляции. Регрессионный анализ.	Коэффициент корреляции как параметр описательной статистики. Коэффициент корреляции Пирсона. Определение знака и величины коэффициента корреляции. Факторы, влияющие на коэффициент корреляции. Получение матрицы корреляции. Требования к выборке для тестирования гипотезы о коэффициенте корреляции Пирсона. Регрессионный анализ. Уравнение регрессии. Ошибки предсказания и поиск лучшей линии регрессии. Метод наименьших квадратов. «Анализ остатков» (residual analysis). Множественная линейная регрессия и корреляция. Нелинейная регрессия. Анализ ковариаций (ANCOVA).	4	4	4	4	12
5.	Трансформация данных. Непараметрические критерии.	Требования к выборке для проведения параметрических тестов. Случайность измерений. Независимость измерений. Проверка соответствия распределения в выборке нормальному закону. Критерии Колмогорова-Смирнова, Лиллифорса, Шапиро-Уилкса. Гомогенность дисперсии. Тесты на гомогенность. Выбросы (outliers). Биномиальное распределение. Распределение Пуассона. Трансформация данных. Непараметрические критерии. Ранжирование данных. Критерии Манна-Уитни, Вальда-Вольфовица, Вилкоксона, Крускала-Уоллеса, Фридмана, знаковый, медианный тесты. Ранговые корреляции.	4	4	4	4	12
6	Частотный анализ	Критерии согласия. Сравнение наблюдаемых частот с теоретическими. Биномиальный тест. Критерий χ^2 Пирсона.	4	4	4	4	12

		<p>Поправка Йейтса. Таблицы сопряжённости. Четырёхпольные таблицы для независимых выборок. Точный критерий Фишера и критерий χ^2 для независимых выборок. Корреляция между качественными переменными. Повторные измерения бинарных переменных: критерии Мак-Немара и Кохрана. Лог-линейные модели (Log-linear models).</p>						
7.	<p>Основы многомерных методов анализа. Дискриминантный анализ.</p>	<p>Общие представления о методах многомерного анализа. Понятие многомерного распределения, его характеристики. Матрицы дисперсий, ковариаций и корреляций переменных. Собственное значение и собственный вектор матрицы. Матрица расстояний между объектами. Подготовка данных для многомерного анализа. Дискриминантный анализ. Дискриминантные функции, их структура и интерпретация. Функции классификации и Классификационная матрица. Пошаговый дискриминантный анализ. Обобщённый дискриминантный анализ.</p>	4	4	4	4	12	
8.	<p>Факторный анализ. Анализ главных компонент.</p>	<p>Задачи факторного анализа. Различные типы факторного анализа. Подготовка данных к анализу. Анализ главных компонент, его основные этапы. Главные компоненты, их структура и интерпретация. Вращение компонент. Канонический анализ.</p>	4	4	4	4	12	
9.	<p>Многомерное шкалирование, кластерный анализ. Обобщённые линейные модели.</p>	<p>Выявление структуры данных на основе матриц расстояний (dissimilarity) между объектами. Способы оценки расстояний в многомерном пространстве (Евклидовы расстояния, квадрат Евклидова расстояния, Манхэттенские расстояния). Многомерное шкалирование, основные этапы. Диаграмма Шеппарда. Интерпретация результатов многомерного шкалирования. Кластерный анализ. Различные методики кластерного анализа, варианты построения деревьев классификации. Кластеризация методом K средних. Обобщённые линейные модели (generalized linear models). Принципы построения модели. Тестирование гипотез. Поиск оптимальной модели и информативные критерии.</p>	4	4	4	4	12	
10.	<p>Основы анализа данных с помощью языка R.</p>	<p>Язык R как аппарат для анализа биологических данных. Подготовка данных для анализа в R и импорт данных. Графические интерфейсы в R. Основы синтаксиса. Подбор и</p>	4	4	4	4	12	

		загрузка программных пакетов и поиск функций, соответствующих задачам исследования. Описание данных. Основные статистические критерии в R. Преимущества R в построении сложных аналитических моделей, обзор возможностей R, выходящих за пределы возможностей традиционного программного обеспечения.					
11. Экзамен					4	36	34 108

Л – лекции, С – семинары, ЛПЗ – лабораторно-практические занятия, СР – самостоятельная работа

Образовательные технологии

Лекции, семинары, практические занятия, написание рефератов, подготовка презентаций и выступлений.

Текущая и промежуточная аттестация.

Текущая аттестация аспирантов проводится в соответствии с локальным актом ИПЭЭ РАН - Положением о текущем контроле, промежуточной и итоговой аттестации аспирантов ИПЭЭ РАН по программам высшего образования – программам подготовки научно-педагогических кадров в аспирантуре и является обязательной. Текущая аттестация по дисциплине проводится в форме решения задач по данной дисциплине.

Объектами оценивания выступают: активность на занятиях, своевременность выполнения различных видов заданий, посещаемость занятий; степень усвоения теоретических знаний и уровень овладения практическими умениями и навыками по всем видам учебной работы, проводимых в рамках семинаров, практических занятий и самостоятельной работы.

Промежуточная аттестация аспирантов по дисциплине проводится в соответствии с локальным актом ИПЭЭ РАН - Положением о текущем контроле, промежуточной и итоговой аттестации аспирантов ИПЭЭ РАН по программам высшего образования – программам подготовки научно-педагогических кадров в аспирантуре и является обязательной. Промежуточная аттестация по дисциплине осуществляется в форме экзамена. Аспирант допускается к экзамену в случае выполнения всех учебных заданий и мероприятий, предусмотренных настоящей программой. В случае наличия учебной задолженности (пропущенных занятий и (или) невыполненных заданий) аспирант обрабатывает пропущенные занятия и выполняет задания.

Оценивание обучающегося на промежуточной аттестации осуществляется на экзамене с использованием нормативных оценок по 5-х балльной системе (5-отлично, 4- хорошо, 3-удовлетворительно, 2-неудовлетворительно).

Оценивание аспиранта на промежуточной аттестации в форме экзамена

Оценка	Требования к знаниям и критерии выставления оценок
2, неудовлетворительно	Аспирант при ответе демонстрирует плохое знание значительной части основного материала в области статистического анализа данных. Не информирован или слабо разбирается в проблемах и/или не в состоянии наметить пути их решения, не может статистически проанализировать данные в соответствии с поставленной

	задачей.
3, удовлетворительно	Аспирант при ответе демонстрирует знания только основного материала в области статистического анализа данных, фрагментарно понимает назначение основных статистических методов, их ограничения и интерпретацию. Не всегда может подобрать корректный метод анализа данных для заданной ситуации, провести соответствующий анализ с использованием специализированного программного обеспечения, интерпретировать результаты и изложить их в соответствии с общепринятыми стандартами
4, хорошо	Поступающий при ответе демонстрирует хорошие знания в области статистического анализа данных, владеет основными принципами статистического анализа данных, понимает назначение основных статистических методов, их ограничения и интерпретацию. Может подобрать корректный метод анализа данных для заданной ситуации, провести соответствующий анализ с использованием специализированного программного обеспечения, но не всегда в состоянии интерпретировать результаты и изложить их в соответствии с общепринятыми стандартами.
5, отлично	Поступающий при ответе демонстрирует глубокое и прочное владение и использование знаний в области статистического анализа данных, владеет принципами статистического анализа данных, понимает назначение статистических методов, их ограничения и интерпретацию. Информирован и способен делать анализ проблем и намечать пути их решения. Может подобрать корректный метод анализа данных для заданной ситуации, провести соответствующий анализ с использованием специализированного программного обеспечения, интерпретировать результаты и изложить их в соответствии с общепринятыми стандартами.

Учебно-методическое и информационное обеспечение дисциплины:

Основная литература

- Sokal, R. R. and F. J. Rohlf. 2012. Biometry: the principles and practice of statistics in biological research. 4th edition. W. H. Freeman and Co.: New York. 937 pp.
- Zar J.K., 2010. Biostatistical analysis. 5th ed. (или 4d ed, 1999) Prentice Hall, New Jersey. 944 P.
- Гланц С, 1998. Медико-Биологическая Статистика McGraw-Hill, 1994; М.: Практика. 459 с.
- Лакин Г. Ф. Биометрия: учебное пособие для биол. спец. вузов /М.: Высшая школа. – 1990. – Т. 352.
- Маслицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R. - М.: ДМК Пресс, 2015. - 496 с. <http://www.ievbras.ru/ecostat/Kiril/>
- Плохинский Н. А. Биометрия. – 1961.
- Плохинский Н. А. Математические методы в биологии: Учебно-методическое пособие для студентов биологических факультетов университетов. – МГУ, 1978.
- Халафян А. А. - Учебник STATISTICA 6 Статистический анализ данных. М. Бином. 2007

Дополнительная литература

- Hurlburt R. T., 2006. Comprehending behavioral statistics. 4rd ed. (или 3rd ed, 2003), Wadsworth/Thomson Learning, Belmont.
- Lehner P.N., 1996. Handbook of ethological methods. Cambridge University press. 672p.
- Боровиков В. Популярное введение в программу Statistica. КомпьютерПресс. 2000.
- Тихомирова М. М. (ред.). Биометрия: учебное пособие. – Изд-во Ленингр. ун-та, 1982.

Базовые журналы:

1. Биометрика
2. Доклады РАН
3. Известия РАН
4. Статистические методы
5. Успехи современной биологии

Библиотечные и Интернет-ресурсы

№	Ссылка на информационный ресурс	Наименование разработки в электронной форме	Доступность (количество точек доступа)
	http://www.nature.com/nature	Nature	64
	http://www.nature.com/methods	Nature Methods	64
	http://www.webofknowledge.com	Web of Science. Библиографическая база данных	64
	http://www.sciencedirect.com/science	ScienceDirect. База журналов издательства Elsevier	64
	http://www.elsevier.com	Elsevier Поисковая система публикаций	64
	http://www.springerlink.com	SpringerLink. База журналов издательства Springer	64
	http://www.springer.com	Springer Поисковая система публикаций	64
	http://www.annualreviews.org	Annual Reviews. База	64
	http://onlinelibrary.wiley.com/	Wiley Электронная библиотека	64
	http://online.sagepub.com/	Sage Journals	64
	http://www.annualreviews.org/	Annual Reviews Sciences Collection	64
	http://www.sciencemag.org/journals	Science/AAS	64

Материально-техническое обеспечение дисциплины.

В ИПЭЭ им. А.Н. Северцова РАН имеется следующее оборудование: компьютеры со специализированным программным обеспечением; проектор, сеть WiFi, ноутбуки, плазменные панели и проекторы для демонстраций

Учебно-методическое обеспечение самостоятельной работы.

Библиотечные и Интернет-ресурсы, консультации с ведущими специалистами Института.

Язык преподавания: русский.

Преподаватель: к.б.н. Васильева Нина Александровна

Контрольные вопросы:

1. Статистика как инструмент для количественного анализа и интерпретации данных. Описательная и индуктивная статистика.
2. Понятия выборки и измерения. Принципы сбора данных для статистического анализа. Свойства переменных. Частотное распределение переменной.
3. Графическое представление частотного распределения.
4. Характеристики распределения.
5. Нормальное распределение. Площадь нормального распределения.
6. Распределение выборочных средних.
7. Понятие статистики критерия. Понятие статистической гипотезы, принципы формулирования взаимоисключающих гипотез.
8. Процедура тестирования гипотез в статистике, статистические ошибки.
9. Критерии Стьюдента: одновыборочный, двухвыборочный для независимых выборок, двухвыборочный для связанных выборок.
10. Дисперсионный анализ. Связь с двухвыборочным t-критерием Стьюдента. Рекомендации и требования к выборкам.
11. Однофакторный дисперсионный анализ.
12. Эффект множественных сравнений. Апостериорные тесты (Бонферрони, Тьюки, Ньюмена-Кейлса, Даннета, Шеффе).
13. Анализ контрастов.
14. Дисперсионный анализ для связанных выборок.
15. Многофакторный дисперсионный анализ.
16. Модели дисперсионного анализа. Модель смешанных эффектов.
17. Гнездовой дисперсионный анализ.
18. Многомерный дисперсионный анализ.
19. Понятие мощности статистического критерия.
20. Практическая интерпретация результатов (размер эффекта).
21. Оценка необходимого размера выборки для предполагаемого статистического анализа.
22. Коэффициент корреляции как параметр описательной статистики. Коэффициент корреляции Пирсона.
23. Определение знака и величины коэффициента корреляции. Факторы, влияющие на коэффициент корреляции.
24. Получение матрицы корреляции. Требования к выборке для тестирования гипотезы о коэффициенте корреляции Пирсона.
25. Регрессионный анализ. Уравнение регрессии.
26. Ошибки предсказания и поиск лучшей линии регрессии. Метод наименьших квадратов.
27. Анализ остатков.
28. Множественная линейная регрессия и корреляция.

29. Нелинейная регрессия.
30. Анализ ковариаций.
31. Требования к выборке для проведения параметрических тестов. Случайность измерений. Независимость измерений.
32. Проверка соответствия распределения в выборке нормальному закону. Критерии Колмогорова-Смирнова, Лиллифорса, Шапиро-Уилкса.
33. Гомогенность дисперсии. Тесты на гомогенность.
34. Выбросы (outliers).
35. Биномиальное распределение. Распределение Пуассона.
36. Трансформация данных.
37. Непараметрические критерии. Ранжирование данных.
38. Критерии Манна-Уитни, Вальда-Вольфовица, Вилкоксона, Крускалла-Уоллеса, Фридмана, знаковый, медианный тесты.
39. Ранговые корреляции.
40. Критерии согласия. Сравнение наблюдаемых частот с теоретическими.
41. Биномиальный тест.
42. Критерий χ^2 Пирсона. Поправка Йейтса.
43. Таблицы сопряжённости. Четырёхпольные таблицы для независимых выборок.
44. Точный критерий Фишера и критерий χ^2 для независимых выборок.
45. Корреляция между качественными переменными.
46. Повторные измерения бинарных переменных: критерии Мак-Немара и Кохрана.
47. Лог-линейные модели.
48. Общие представления о методах многомерного анализа. Понятие многомерного распределения, его характеристики.
49. Матрицы дисперсий, ковариаций и корреляций переменных. Собственное значение и собственный вектор матрицы. Матрица дистанций между объектами.
50. Подготовка данных для многомерного анализа.
51. Дискриминантный анализ. Дискриминантные функции, их структура и интерпретация.
52. Функции классификации и Классификационная матрица.
53. Пошаговый дискриминантный анализ.
54. Обобщённый дискриминантный анализ.
55. Задачи факторного анализа.
56. Различные типы факторного анализа. Подготовка данных к анализу.
57. Анализ главных компонент, его основные этапы.
58. Главные компоненты, их структура и интерпретация.
59. Вращение компонент.
60. Канонический анализ.

61. Выявление структуры данных на основе матриц дистанций между объектами. Способы оценки дистанций в многомерном пространстве.
62. Многомерное шкалирование, основные этапы.
63. Диаграмма Шепарда. Интерпретация результатов многомерного шкалирования.
64. Кластерный анализ. Различные методики кластерного анализа, варианты построения деревьев классификации.
65. Кластеризация методом K средних.
66. Обобщённые линейные модели. Принципы построения модели.
67. Поиск оптимальной модели и информационные критерии.
68. Язык R как аппарат для анализа биологических данных. Подготовка данных для анализа в R и импорт данных.
69. Графические интерфейсы в R.
70. Основы синтаксиса в R.
71. Подбор и загрузка программных пакетов и поиск функций, соответствующих задачам исследования.
72. Описание данных. Основные статистические критерии в R.