

Федеральное бюджетное учреждение науки
«ГОСУДАРСТВЕННЫЙ НАУЧНЫЙ ЦЕНТР ВИРУСОЛОГИИ И
БИОТЕХНОЛОГИИ «ВЕКТОР»
Федеральной службы по надзору в сфере защиты прав потребителей и
благополучия человека
(ФБУН ГНЦ ВБ «Вектор» Роспотребнадзора)



УТВЕРЖДАЮ
Генеральный директор
ФБУН ГНЦ ВБ «Вектор»
Роспотребнадзора
Р.А. Максютов

« 16 » августа 2022 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
«СТАТИСТИКА И АНАЛИЗ ДАННЫХ»

Группа научных специальностей: 1.5. Биологические науки

Научная специальность: 1.5.3. Молекулярная биология

1.5.6. Биотехнология

1.5.10. Вирусология

Кольцово 2022

Составитель:

А.А. Черевко, канд. физ.-мат. наук, доцент

Согласовано:

Заведующий отделом аспирантуры



Т.Ю. Болдырева

Программа рассмотрена и одобрена на заседании Ученого совета ФБУН ГНЦ ВБ «Вектор»
Роспотребнадзора.

Протокол «10» 05.10.2022 г. № ____

Содержание	
1. Общие положения.....	4
2. Требования к планируемым результатам освоения дисциплины	4
3. Содержание дисциплины	4
3.1. Трудоемкость освоения дисциплины.....	4
3.2. Тематический план дисциплины.....	5
3.3. Содержание разделов и тем дисциплины.....	5
4. Оценка качества освоения дисциплины	6
4.1. Текущий контроль успеваемости.....	6
4.2. Промежуточная аттестация	7
4.3. Перечень вопросов для подготовки к зачету	7
4.4. Критерии оценивания качества освоения дисциплины	8
5. Учебно-методическое обеспечение и материально-техническое обеспечение дисциплины	8
5.1 Основная учебная литература	8
5.2 Дополнительная учебная литература.....	8
5.3. Научная и справочно-библиографическая литература	8
5.4. Ресурсы сети «Интернет»	8
5.5. Информационные технологии (программные продукты)	9
5.6. Материально-техническое обеспечение.....	9

1. Общие положения

Цель дисциплины - обучить основным принципам математической статистики для анализа данных биологических исследований, выбору подходящего метода анализа в зависимости от имеющихся данных и использования его, научить основам анализа данных с использованием специализированных компьютерных программ.

Дисциплина отнесена к обязательным дисциплинам образовательного компонента программы подготовки научных и научно-педагогических кадров в аспирантуре по группе научных специальностей 1.5 «Биологические науки». Изучается на 2 курсе.

2. Требования к планируемым результатам освоения дисциплины

В результате освоения дисциплины «Статистика и анализ данных» аспирант должен:

Знать	Основы описательной статистики. Случайные события, свойства вероятности, независимые и зависимые события. Основные распределение вероятностей и их свойства. Основы теории проверки статистических гипотез. p -уровень значимости, доверительные интервалы. Ложноположительные и ложноотрицательные заключения (ошибки первого и второго рода). Метод главных компонент. Установление статистических зависимостей на основе корреляционного и регрессионного анализа.
Уметь	Вычислять меры центральной тенденции и меры изменчивости. Выделять и обрабатывать выбросы. Пользоваться средствами визуализации данных в R на основе библиотеки ggplot2. Проверять данные на нормальность распределения графическими методами и тестами. Использовать критерии различий для двух и нескольких групп, в случае связанных и несвязанных выборок, для данных имеющих как нормальное, так и не нормальное распределение. Использовать метод главных компонент. Строить линейную и логистическую регрессию. Проводить графический анализ качества регрессии.
Владеть	Навыками: <ul style="list-style-type: none">• преобразования данных для приведения распределения к нормальному виду (Вох-Сох преобразование),• определения мощности критериев проверки гипотез, и оценки необходимого количества опытов;• проверки попарных различий между многими группами с помощью post hoc критериев;• проверки условия применимости линейной регрессии – на основе теоремы Гаусса-Маркова.

3. Содержание дисциплины

3.1 Трудоемкость освоения дисциплины

Вид учебной работы	Всего часов/ З.Е.
Контактная работа обучающихся с преподавателем (всего)	27
в том числе: лекции	9

Вид учебной работы	Всего часов/ З.Е.
практические (семинарские) занятия	18
Самостоятельная работа обучающихся	9
Общая трудоемкость	36/1

3.2. Тематический план дисциплины

Наименование разделов и тем	Всего часов	Лекции	Практические (семинарские) занятия	Самостоятельная работа
Тема 1. Основы описательной статистики	3	1	2	
Тема 2. Программы для обработки данных	2	1		1
Тема 3. Распределение вероятностей и основы графики	6	1	4	1
Тема 4. Проверка статистических гипотез	2	1		1
Тема 5. Критерии различий для двух несвязанных выборок	4	1	2	1
Тема 6. Сравнение нескольких выборок	5	2	2	1
Тема 7. Критерии различий для связанных выборок	3		2	1
Тема 8. Анализ многомерных данных. Метод главных компонент	2	1		1
Тема 9. Корреляционный анализ	3		2	1
Тема 10. Основы регрессионного анализа	6	1	4	1
Итого по дисциплине	36	9	18	9

3.3. Содержание разделов и тем дисциплины

Тема 1. Основы описательной статистики. Случайные события, свойства вероятности, независимые и зависимые события. Меры центральной тенденции и меры изменчивости. Выбросы и их обработка.

Тема 2. Программы для обработки данных. Программы для обработки данных. Microsoft Excel, пакет Statistica, среда R+RStudio

Тема 3. Распределение вероятностей и основы графики. Распределение вероятностей. Нормальное распределение, его характеристики. Распределения отличные от нормальных. Логнормальное распределение. Преобразование данных для приведения распределения к нормальному виду (Вох-Сох преобразование). Проверка на нормальность. Графические методы. Тест Шапиро-Уилка, тест Колмогорова-Смирнова и его модификация Лиллиефорса. Средства визуализации данных в R на основе библиотеки ggplot2: гистограммы и полигоны частот, круговые диаграммы, диаграммы рассеяния и пузырьковые диаграммы, boxplot («ящик с усами»).

Тема 4. Проверка статистических гипотез. Проверка статистических гипотез. р-уровень значимости, доверительные интервалы. Ложноположительные и ложноотрицательные заключения (ошибки первого и второго рода), мощность критериев проверки гипотез, оценка необходимого количества опытов.

Тема 5. Критерии различий для двух несвязанных выборок. t-критерий Стьюдента и U-критерий Манна-Уитни для средних значений. F-критерий равенства дисперсий Фишера, критерий Хи-квадрат Пирсона для не количественных данных.

Тема 6. Сравнение нескольких выборок. Сравнение нескольких выборок: основы дисперсионного анализа (ANOVA). H-критерием Краскела-Уоллеса для проверки равенства медиан нескольких выборок. Проблема множественных сравнений при попарном сравнении более чем двух групп, апостериорные (post hoc) критерии. t-критерий Стьюдента с поправкой Бонферрони, критерий Тьюки (для групп с равными дисперсиями); критерий Даннета, критерий Геймса-Хоуэлла (для групп с различными дисперсиями).

Тема 7. Критерии различий для связанных выборок. t-критерий Стьюдента для связанных (зависимых) выборок, T-критерий Вилкоксона, критерий Мак-Немара (вместо Хи-квадрат). Обработка результатов эксперимента с контрольной группой и двумя замерами (в начале и в конце эксперимента).

Тема 8. Анализ многомерных данных. Метод главных компонент. Основы метода главных компонент

Тема 9. Корреляционный анализ. Установление статистических зависимостей между двумя и более переменными.

Тема 10. Основы регрессионного анализа. Линейная регрессия для одной и нескольких переменных. Коэффициент детерминации R^2 , уровни значимости коэффициентов регрессии и уровень значимости регрессии в целом. Условия применимости линейной регрессии – остатки регрессии и теорема Гаусса-Маркова. Графический анализ качества регрессии. Логистическая регрессия

Темы практических занятий с обучающимися

Вычисление показателей описательной статистики. Основы работы в R и RStudio.

Средства визуализации данных в R на основе библиотеки ggplot2: гистограммы и полигоны частот, круговые диаграммы, диаграммы рассеяния и пузырьковые диаграммы, boxplot («ящик с усами»).

Проверка на нормальность. Графические методы. Тест Шапиро-Уилка, тест Колмогорова-Смирнова и его модификация Лиллиефорса.

Критерии различий для двух несвязанных выборок. F-критерий равенства дисперсий Фишера, критерий Хи-квадрат Пирсона для не количественных данных.

Проблема множественных сравнений при попарном сравнении более чем двух групп, апостериорные (post hoc) критерии. t-критерий Стьюдента с поправкой Бонферрони, критерий Тьюки (для групп с равными дисперсиями); критерий Даннета, критерий Геймса-Хоуэлла (для групп с различными дисперсиями).

Критерии различий для связанных выборок. t-критерий Стьюдента для связанных (зависимых) выборок, T-критерий Вилкоксона, критерий Мак-Немара (вместо Хи-квадрат). Обработка результатов эксперимента с контрольной группой и двумя замерами (в начале и в конце эксперимента).

Установление статистических зависимостей. Корреляционный анализ.

Условия применимости линейной регрессии – остатки регрессии и теорема Гаусса-Маркова. Графический анализ качества регрессии.

4. Оценка качества освоения дисциплины

4.1. Текущий контроль успеваемости

Текущий контроль освоения материала дисциплины для аспирантов проводится по итогам выполнения заданий самостоятельной работы. Для текущего контроля успеваемости аспирантов используется проведение устного группового опроса по одной или нескольким важным темам курса. Предусмотрено выполнение контрольных заданий. Успешное выполнение контрольных заданий является обязательным. Аспиранты, не выполнившие в полном объеме все эти работы, не допускаются к сдаче зачета

4.2. Промежуточная аттестация

Для контроля качества освоения дисциплины учебным планом предусмотрен зачет. Форма проведения экзамена зачета – в форме ответов на вопрос преподавателя и выполнения практического задания в среде R.

4.3. Перечень вопросов для подготовки к зачету

- Случайные события.
 - Основные свойства вероятности.
 - Независимые и зависимые события.
 - Определения основных мер центральной тенденции и изменчивости.
 - Что такое выбросы и как их обрабатывать?
 - Что такое нормальное распределение, его характеристики?
 - Как осуществляется преобразование данных для приведения распределения к нормальному виду?
 - Как проверить на нормальность графическим методом?
 - Как проводятся тесты Шапиро-Уилка, Колмогорова-Смирнова и его модификация Лиллиефорса?
 - Что такое р-уровень значимости?
 - Что такое доверительные интервалы?
 - Что такое мощность критерия проверки гипотез?
 - Как проводятся t-критерий Стьюдента и U-критерий Манна-Уитни для средних значений? Условия их применимости.
 - Как проводятся F-критерий равенства дисперсий Фишера, критерий Хи-квадрат Пирсона? Условия их применимости.
 - Общая схема сравнения нескольких выборок
 - Как проводится дисперсионный анализ? Условия его применимости.
 - Как проводится H-критерием Краскела-Уоллеса для проверки равенства медиан нескольких выборок? Условия его применимости.
 - Что такое проблема множественных сравнений при попарном сравнении более чем двух групп?
 - Что такое апостериорные (post hoc) критерии?
 - Как проводятся t-критерий Стьюдента с поправкой Бонферрони, критерий Тьюки; критерий Даннета, критерий Геймса-Хоуэлла? В чем между ними различия и каковы их условия применимости.
 - Как провести t-критерий Стьюдента для связанных (зависимых) выборок, T-критерий Вилкоксона? Условия их применимости.
 - Как провести критерий Мак-Немара?
 - Как провести обработку результатов эксперимента с контрольной группой и двумя мерами (в начале и в конце эксперимента)?
 - Схема применения метода главных компонент.
- Варианты практических заданий**
- Как установить статистическую зависимость между переменными?
 - Как построить линейную регрессию для одной или нескольких переменных?
 - Что такое коэффициент детерминации R²?

Что такое уровни значимости коэффициентов регрессии и уровень значимости регрессии в целом?

Каковы условия применимости линейной регрессии?

Как осуществить графический анализ качества регрессии?

Как построить логистическую регрессию?

4.4. Критерии оценивания качества освоения дисциплины

Оценка	Критерии оценки
Зачтено	Выставляется обучающемуся, показавшему систематизированные знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач.
Не зачтено	Выставляется обучающемуся, который не знает большей части основного содержания учебной программы дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

5. Учебно-методическое обеспечение и материально-техническое обеспечение дисциплины

5.1 Основная учебная литература

1. Гланц, С. Медико-биологическая статистика / Стентон Гланц ; пер. с англ. : Юлий Данилов, Н. Бузикашвили, Д. Самойлов — Москва : Практика, 1998. — 459 с. — Текст : электронный — URL: <http://medstatistic.ru/articles/glantz.pdf>
2. Лонг, Дж. Д. Книга рецептов : Проверенные рецепты для статистики, анализа и визуализации данных / Дж. Д. Лонг, Пол Титор ; пер. с англ. Д. А. Беликова. — Москва : ДМК Пресс, 2020. — 510 с. — ISBN 978-5-97060-835-7.
3. Мастицкий, С. Э. Визуализация данных с помощью ggplot2 : практическое руководство / С. Э. Мастицкий. — Москва : ДМК Пресс, 2017. — 222 с. — ISBN 978-5-97060-470-0.
4. Кабаков, Р. И. R в действии. Анализ и визуализация данных в программе R / Р.И. Кабаков ; пер. с англ. Полины А. Волковой. — Москва : ДМК Пресс, 2014. — 588 с.
5. Faraway, Julian James. Linear models with R. Second edition. CRC Press, 2015.

5.2 Дополнительная учебная литература

1. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. Москва : ФИЗМАТЛИТ, 2006.
2. Шитиков, В. К. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R / В.К. Шитиков, Г.С. Розенберг — Тольятти : Кассандра, 2014. — 314 с. — URL: <http://www.ievbras.ru/ecostat/Kiril>

5.3. Научная и справочно-библиографическая литература

1. Лагутин, М. Б. Наглядная математическая статистика: учебное пособие / М.Б. Лагутин — Москва : БИНОМ. Лаборатория знаний, 2015.

5.4. Ресурсы сети «Интернет»

1. Российское образование (федеральный портал) www.edu.ru
2. Государственная публичная научно-техническая библиотека www.gpntb.ru.
3. Научная электронная библиотека eLIBRARY www.elibrary.ru

5.5. Информационные технологии (программные продукты)

Webinar.ru – Лицензионный договор

«R» - свободная программа для статистических вычислений и графики

5.6. Материально-техническое обеспечение

Учебная аудитория, оснащенная мультимедийным оборудованием (компьютер, проектор).

**Лист визирования программы
для исполнения в очередном учебном году**

Программа пересмотрена и одобрена для исполнения в 20____-20____ учебном году
с изменениями / дополнениями, перечень которых прилагается (Приложение № ____)
без изменений

Протокол № _____ заседания ученого совета от ____ . ____ .20 _____

**Лист визирования программы
для исполнения в очередном учебном году**

Программа пересмотрена и одобрена для исполнения в 20____-20____ учебном году
с изменениями / дополнениями, перечень которых прилагается (Приложение № ____)
без изменений

Протокол № _____ заседания ученого совета от ____ . ____ .20 _____

**Лист визирования программы
для исполнения в очередном учебном году**

Программа пересмотрена и одобрена для исполнения в 20____-20____ учебном году
с изменениями / дополнениями, перечень которых прилагается (Приложение № ____)
без изменений

Протокол № _____ заседания ученого совета от ____ . ____ .20 _____

**Лист визирования программы
для исполнения в очередном учебном году**

Программа пересмотрена и одобрена для исполнения в 20____-20____ учебном году
с изменениями / дополнениями, перечень которых прилагается (Приложение № ____)
без изменений

Протокол № _____ заседания ученого совета от ____ . ____ .20 _____