Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
Московский государственный университет имени М.В. Ломоносова
Факультет биоинженерии и биоинформатики

УТВЕРЖДАЮ
Декан
факультета биоинженерии
и биоинформатики,
академик
_______/В.П. Скулачев /

«____»_____20__г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины:

Методы прикладной статистики в биологии

Уровень высшего образования:
специалитет

Направление подготовки (специальность):

06.05.01 Биоинженерия и биоинформатика

Форма обучения:
очная

Рабочая программа рассмотрена и одобрена

Ученым советом факультета
(протокол №_____,_____)

Москва 20___

Рабочая программа дисциплины разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по специальности 06.05.01 «Биоинженерия и биоинформатика» (программы специалитета) в редакции приказа МГУ от 30 декабря 2016 г.

Год (годы) приема на обучение – 2016, 2017, 2018, 2019.

Цель и задачи дисциплины

Цель курса - изучение основных понятий математической статистики в приложении к задачам биомедицинской тематики

Задачи курса: сформировать у студентов представление о базовых статистических принципах, таких как генеральная совокупность и выборки, проверка гипотез и построение доверительных интервалов, а также научить студентов решать задачи из повседневной жизни составлять статистические отчеты.

1. Место дисциплины в структуре ОПОП ВО – вариативная часть, профессиональный цикл, курс IV – семестр 7.

2. Входные требования для освоения дисциплины, предварительные условия (если есть): *освоение дисциплины «Теория вероятностей»*

3. Планируемые результаты обучения по дисциплине:

Знать:

Ожидается общее понимание базовых принципов статистики, таких как получение выборки, проверка гипотез, построение доверительных интервалов, оценок, разложение дисперсии и т.д.

Уметь

Ожидается, что студенты будут способны решать задачи из повседневной жизни, составлять статистические отчеты и делать выводы (это умение не ограничивается биомедицинской тематикой). Студенты будут уметь работать в группах и приобретут хорошие практики управления и обмена данными, а также разовьют способность делать устные и письменные презентации.

Владеть:

Студенты будут способны читать и понимать статистические задачи, выполнять статистические тесты, выбирать статистические тесты для своих частных задач, графически представлять данные, планировать статистические исследования.

4. Формат обучения – лекционные занятия.

5. Объем дисциплины составляет 2 з.е., в том числе 28 академических часов, отведенных на контактную работу обучающихся с преподавателем, 44 академических часа на самостоятельную работу обучающихся.

6. Краткое содержание дисциплины (аннотация):

Этот вводный курс является продолжением университетского курса элементарной теории вероятностей, а также соответствующего ему англоязычного курса Advanced Placement Statistics, с расширением до более широкого спектра тем и приложениями к задачам из повседневной жизни и, в особенности, к биомедицинским задачам. Материал курса предлагается в виде пяти последовательных модулей, каждый из которых состоит из лекции и практических занятий. В практикуме используется язык R-statistics и приложение R-studio. Тем не менее, курс направлен на изучение предмета статистики, а не программного обеспечения, поэтому языки программирования в нем задействованы лишь инструментально.

Наименование и краткое содержание разделов и тем дисциплины, **Форма промежуточной аттестации по дисциплине**	**Всего (часы)**	В том числе	
		Контактная работа (работа во взаимодействии с преподавателем) Виды контактной работы, часы	**Самостоятельная работа обучающегося, часы** *(виды самостоятельной работы – эссе, реферат, контрольная работа и пр. – указываются при необходимости)*

		Занятия лекционного типа	Занятия семинарского типа	**Всего**	
Планирование статистического эксперимента. Эксперимент и наблюдательное исследование. Простая случайная выборка. Отклонения в выборках и их классификация. Описательные статистики. Способы графического изображения выборок: гистограмма, бокс-плот и пр.	6	2		2	Домашняя работа, Квиз, 4 часа
Обзор элементарных сведений из теории вероятностей. Условная вероятность. Формула Байеса. Формула полной вероятности. Случайные величины. Математическое ожидание, дисперсия. Ковариация и корреляция. Геометрический смысл.	2	2		2	
Дискретные распределения. Равномерное, биномиальное, геометрическое и Пуассоновское распределения. Примеры дискретных распределений в биологических задачах.	7	3		3	Домашняя работа, Квиз, 4 часа

Приближение биномиального распределения Пуассоновским. Выборки с возвращением и без возвращения. Гипергеометрическое распределение	6	2		2	Домашняя работа, Квиз, 4 часа
Непрерывные распределения. Нормальное распределение. Распределение пропорций и его связь с биномиальным распределением. Центральная предельная теорема. Аппроксимация биномиального распределения нормальным и поправка на непрерывность. Контрольная работа 2 часа	6	2		2	Домашняя работа, Квиз, 4 часа
Теория точечного оценивания. Несмещенность и эффективность оценок. Среднеквадратичное отклонение. Теорема Штейнера. Математическое ожидание и дисперсия выборочного среднего. Примеры несмещенных и эффективных оценок.	6	2		2	Домашняя работа, Квиз, 4 часа
Интервальное оценивание. Доверительные интервалы и их интерпретация. Уровень доверия. Стандартная ошибка. Поправка на конечный размер генеральной совокупности.	7	3		3	Домашняя работа, Квиз, 4 часа
Интервальное оценивание пропорций и разности пропорций. Оценки для среднего и разности средних в случае известных и неизвестных стандартных отклонений. Распределение Стьюдента и условия его применимости. Случай равных дисперсий и оценки дисперсии.	2	2		2	

Проверка гипотез. Ошибки первого и второго рода. Уровень значимости и сила теста. P-значение и его интерпретация. Тестирование гипотез с использованием доверительных интервалов и тестоых статистик. Критические значения. Согласованность гипотез и доверительных интервалов	8	4		4	Домашняя работа, Квиз, 4 часа
Распределение хи-квадрат. Условия применимости в задачах. Критерий согласия Пирсона. Таблицы сопряженности. Выборки без возвращения и точный тест Фишера. Биологические задачи.	6	2		2	Домашняя работа, Квиз, 4 часа
Свойства выборочного стандартного отклонения и интервальное оценивание дисперсии. Проверка гипотез с использованием распределения хи-квадрат. Распределение Фишера. Свойства квантилей. Проверка гипотез с использованием распределения Фишера.	6	2		2	Домашняя работа, Квиз, 4 часа
Однофакторный и двухфакторный дисперсионный анализ (ANOVA). Проверка гипотез в дисперсионном анализе. Предположения. Доверительные интервалы для одновременного оценивания разности средних. Разложение суммы квадратов вычетов.	6	2		2	Домашняя работа, Квиз, 4 часа
Промежуточная аттестация: зачет					4
Итого	72	28		28	44

7. Фонд оценочных средств (ФОС) для оценивания результатов обучения по дисциплине

7.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости.

Пример

I. Домашнее задание

1. Чему равно стандартное отклонение нормального распределения со средним 25, если 18% наблюдений лежат выше 29?

2. Компания заявляет, что ее аспирин избавляет от головной боли быстрее, чем любая другая подобная таблетка. Чтобы проверить это утверждение было выбрано 15 таблеток аспирина и 15 таблеток, произведенных другой компанией. Лекарства давали 30 случайно выбранным людям и измеряли количество минут до исчезновения головной боли. Средние оказались равны 8.5 и 9.1, соответственно, а стандартные отклонения – 4.2 и 4.9.

a. Опишите какие систематические ошибки могут быть в таком эксперименте и как вы будете с ними бороться
b. Протестируйте на 95% уровне доверия гипотезу о том, что аспирин данной компании избавляет от головной боли быстрее, чем аспирин другого производства. Объясните свои предположения.
3. Предложите схему эксперимента для оценки нового химического соединения для защиты кожаных изделий как протектора обуви в зимний сезон. Опишите переменные, мешающие факторы, как вы собираетесь с ними бороться, будете ли использовать ослепление и пр.

II. Практическая работа
1. Таблица CO2 содержит результаты эксперимента по влиянию низких температур на рост растения Echinochloa crusgalli. Используя ggplot2 представьте эти данные графически. Не выполняйте статистические тесты

2. В данной задаче вам нужно при помощи симуляций показать, что сумма n квадратов стандартных нормальных распределений имеет хи-квадрат распределение с n степенями свободы. Симулируйте 1000 выборок объёма 10 из стандартного нормального распределения и сосчитайте сумму квадратов (apply(A^2,1,sum)). Затем постройте эмпирическую функцию плотности вероятности, а также ожидаемую плотность вероятности хи-квадрат распределения. Используя функцию chisq.test с шестью интервалами, протестируйте гипотезу о том, что 1000 наблюдений пришли из хи-квадрат распределения с 10 степенями свободы. Попробуйте другие значения n. Объясните ваши действия.
3. При помощи симуляций продемонстрируйте, что 2-выборочный доверительный интервал для разности средних на самом деле имеет выбранный вами уровень значимости. Представим, что два фармацевтических конвейера производят пищевую добавку. Добавка должна содержать определенное количество соединения А. Чтобы оценить количество продукта, из продукции каждого конвейера взяли выборку размера 12. Выборочные средние оказались равны 127.5 и 120.1 мг соответственно. Допустим стандартное отклонение равно 8.1 мг для каждого конвейера.
а. Постройте и интерпретируйте 95% доверительный интервал для разности средних. Интерпретируйте отдельно уровень значимости. Какие предположения нужны для того, чтобы этот интервал был верным?
б. Симулируйте 1000 пар выборок из соответствующего распределения со средним 125 мг и стандартным отклонением 8.1 мг. Постройте 95% доверительный интервал для каждой пары и сосчитайте в каком проценте случаев доверительный интервал покрывает гипотетическое значение 0. Сколько раз вы ожидали он будет его покрывать? Проведите соответствующий статистический тест.

7.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации.
III. Итоговый экзамен (свободный ответ и тест)
1. Макдоналд и Крейтман в 1991 году секвенировали ген алкогольдегидрогеназы в нескольких особях трех видов дрозофил. Вариативные позиции
классифицировались как синонимичные (последовательность аминокислот не изменялась) и несинонимичные, а также полиморфные (различаются между особями одного вида) и фиксированные. В отсутствие естественного отбора, отношение числа синонимичных замен к несинонимичным должно быть приблизительно одинаковым в полиморфных и фиксированных сайтах. В наблюдаемой ими выборке было 43 синонимичных полиморфных, 2 несинонимичных полиморфных, 17 синонимичных фиксированных и 7 несинонимичных фиксированных замен.
a. (1 балл) Чему равны ожидаемые частоты синонимичных и несинонимичных, полиморфных и фиксированных замен? Назовите тест, который нельзя использовать для этой таблицы сопряженности.
b. (2 балла) Протестируйте на 1% уровне значимости есть ли ассоциация между

синонимичностью и полиморфностью.
c. (1 балл) Если тест Макдональда-Крейтмана применить к 100 генам, включая алкоголь дегидрогеназу, чему должно быть равно п- значение с учетом поправки на множественное тестирование? Назовите поправку на множественное тестирование.
2. Кастер и Галли на легкомоторном самолете проследили за голубыми и белыми цаплями от мест гнездования до мест кормления на озере Пелтье
штат Миннесота и записали на каких поверхностях садилась каждая птица. Достаточно ли оснований считать на 5% уровне значимости, что два вида
птиц различаются по своим предпочтениям?
a. Да, так как п-значение<0.01
b. Нет, так как п-значение<0.01
c. Да, так как п-значение<0.02
d. Нет, так как п-значение<0.02
e. Ни одно из перечисленного

Шкала и критерии оценивания результатов обучения по дисциплине.

Результаты обучения	«Неудовлетворительно»	«Удовлетворительно»	«Хорошо»	«Отлично»
Знания: Ожидается общее понимание базовых принципов статистики, таких как получение выборки, проверка гипотез, построение доверительных интервалов, оценок, разложение дисперсии и т.д.	Знания отсутствуют	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения: Ожидается, что студенты будут способны решать задачи из повседневной жизни, составлять статистические отчеты и делать выводы (это умение не ограничивается биомедицинской тематикой). Студенты будут уметь работать в группах и приобретут хорошие практики управления и обмена данными, а также разовьют способность делать	Умения отсутствуют	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение

устные и письменные презентации.				
Владения: Студенты будут способны читать и понимать статистические задачи, выполнять статистические тесты, выбирать статистические тесты для своих частных задач, графически представлять данные, планировать статистические исследования.	Навыки владения отсутствуют	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач

8. Ресурсное обеспечение:

- Перечень основной и дополнительной литературы

1. Samuels, Witmer, and Schaffner. Statistics for Life sciences, Fourth Edition.
2. Peck, Olsen, Devore, Introduction to Statistics and Data Analysis, Third Edition.
3. Barrons AP Statistics, 8th Edition
4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. Высшая школа 1988
5. Shiryaev A.N. Probability MZNMO.

- Перечень лицензионного программного обеспечения (при необходимости)
- Перечень профессиональных баз данных и информационных справочных систем
- Перечень ресурсов информационно-телекоммуникационной сети «Интернет» (при необходимости)

1. AP statistics http://apcentral.collegeboard.com/apc/members/exam/exam_information/8357.html
2. Statistica https://www.statsoft.com/
3. Quick-R http://www.statmethods.net/
4. R by example http://www.mayin.org/ajayshah/KB/R/
5. ggplot2 cookbook http://ggplot2.org/

- Описание материально-технического обеспечения.