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:

Язык R и его применение в биоинформатике

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 - освоить язык R – полезный инструмент для статистического анализа данных.
Задачи курса:
- освоение базового синтаксиса языка
- построение графиков
- статистический анализ данных
- примеры задач, требующих статистического анализа, из биоинформатики

1. Место дисциплины в структуре ОПОП ВО – вариативная часть, профессиональный цикл, курс III – семестр 5.
2. Входные требования для освоения дисциплины, предварительные условия (если есть): освоение дисциплин «Теория вероятностей» и «Математическая статистика»
3. Планируемые результаты обучения по дисциплине:
Знать:
Базовый синтаксис языка R, подходы к статистической обработке данных и визуализации результатов
Уметь
Производить первичный разведывательный анализ данных, формулировать постановку задачи на основании имеющихся данных,
проводить статистический анализ, визуализировать полученные результаты с использованием кода на языке R
Владеть:
Навыками создания программного конвейера обработки данных на языке R
Иметь опыт
Статистического анализа данных
4. Формат обучения – лекционные занятия
5. Объем дисциплины составляет 2 з.е., в том числе 36 академических часов, отведенных на контактную работу обучающихся с преподавателем, 36 академических часов на самостоятельную работу обучающихся.
6. Краткое содержание дисциплины (аннотация):

Язык R - современный язык программирования для статистического анализа данных. R - векторизованный язык, это позволяет работать с большим набором данных, как с одним числом. Одним из больших достоинств языка является возможность легко строить красивые графики. В R присутствует большое количество встроенных функций для статистического анализа данных, а также большое количество дополнительных пакетов в открытом доступе для более специализированных задач. Целью данного курса является знакомство студентов с базовым синтаксисом языка, линейными моделями, кластеризацией, а также с некоторыми специальными возможностями для решения биоинформатических задач. Кроме того, студенты научатся рисовать информативные графики для представления результатов. Также студенты узнают о том, какие бывают задачи и данные в биоинформатике, познакомятся с пакетами Bioconductor, которые позволяют удобно работать с геномными данными и не только.

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**	
1. Введение. Вектор. Операции с векторами. Способы создания вектора. Срезы. Data frame.	4	2		2	2
2. Встроенные наборы данных. Срезы в data frame. Which. Работа с переменными. Чтение и сохранение данных. Работа с отсутствующими данными.	5	3		3	2
3. Факторы. Матрицы. Списки. Циклы. for, if, break, next, repeat, while. Apply, lapply, sapply	4	2		2	2
4. Статистические тесты. Выборка. Тестирование гипотез. Ошибки первого и второго рода. P-value. Тест на равенство среднего. Одновыборочный и двувыборочный тесты Стьюдента.	5	3		3	2
5. Работа с таблицами. dplyr	4	2		2	2
6. S3-классы. Хи-квадрат тест и точный тест Фишера.	5	2		2	3
7. Контрольная работа	4	2		2	2
8. Манипуляция с данными - Advanced dplyr.	4	2		2	2
9. Графика - Ggplot2	6	4		4	2
10. Корреляция	4	2		2	2
11. Линейная регрессия	6	4		4	2
12. ANOVA	5	2		2	3
13. Контрольная работа	4	2		2	2
14. Bioconductor, дифференциальная экспрессия генов, GO-анализ	4	2		2	2
15. Кластеризация	4	2		2	2
Промежуточная аттестация - зачет	4				*4 (количество часов, отведенных на*

					промежуточную аттестацию)
Итого	72	36		36	36

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

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

Вам дан файл baltimore_preprocessed3.csv

С помощью вашего seed и команды dplyr sample_n выберите 5000 строк без повторения.

Все задания выполняйте на получившемся файле в 5000 строк.

Везде используйте первый результат после установки seed.

Считаем, что каждая строка - это отдельный работник.

На выполнение работы отводится 90 минут.

Обратите внимание на задание № 15!!! Нужно приложить файл вида Your_surname_DZ3.R, где будут приведены все команды, с помощью которых Вы решали задания. Разделяйте команды для каждого задания комментариями с указанием номера задания.

1. Проверьте гипотезу на уровне значимости 0.05 о том, что средняя ANNUAL_RT в Fire Department и в Police Department за 2002 год отличаются. Допускаем равенство дисперсий__
2. Добавьте столбец NEW, заполненный перемешанными значениями столбца ANNUAL_RT. В каждой строке оставьте минимальное значение между данными в столбец ANNUAL_RT и NEW. а) Чему равно минимальное значение этих минимумов? б) Укажите номер строки, на которой находится этот минимум минимумов? (Ответ в виде: а - 123; б - 456)
3. Запишите топ-5 работников по ANNUAL_RT (в порядке уменьшения зарплаты). Если есть работники с зарплатой одинаковой с 5-м местом - выведите их тоже.

7.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации.

1. Создайте список L1, используя свой seed и цикл: for (i in 1:13){ L1[[i]] = sample(letters,26*i,replace=T)}. В каждом элементе L1 посчитайте самую часто встречающуюся букву и запишите результаты в вектор vec (длиной равной количеству элементов списка). Если букв окажется несколько, возьмите случайную (используйте sample). В качестве ответа приведите vec.

2. Создайте таблицу, используя свой seed и команду: data.frame(replicate(10,sample(1:1000,1000,rep=F))). Подсчитайте сумму нечетных максимумов строк получившейся таблицы.

3. Создайте вектор D с 100 случайными уникальными целочисленными значениями из диапазона от 1 до 10000. Посчитайте сумму Vec, в котором отсутствуют значения с индексами из вектора D._____

Шкала и критерии оценивания результатов обучения по дисциплине.

Результаты обучения	«Неудовлетворительно»	«Удовлетворительно»	«Хорошо»	«Отлично»
Знания: базового синтаксиса языка R, подходов к статистической обработке данных и визуализации результатов	Знания отсутствуют	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания

Умения: производить первичный разведывательный анализ данных, формулировать постановку задачи на основании имеющихся данных, проводить статистический анализ, визуализировать полученные результаты с использованием кода на языке R	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
Владения: навыками создания программного конвейера обработки данных на языке R	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. Уикем Хэдли, Гроулмунд Гарретт. Язык R в задачах науки о данных. Импорт, подготовка, обработка, визуализация и моделирование данных
  2. Брюс Эндрю, Брюс Питер. Практическая статистика для специалистов Data Science
  3. С. Э. Мастицкий. Визуализация данных с помощью ggplot2
  4. Шитиков Владимир Кириллович, Мастицкий Сергей Эдуардович. Статистический анализ и визуализация данных с помощью R
- Перечень лицензионного программного обеспечения (при необходимости)
  Язык R
  RStudio
- Перечень профессиональных баз данных и информационных справочных систем
- Перечень ресурсов информационно-телекоммуникационной сети «Интернет» (при необходимости)
- Описание материально-технического обеспечения.