Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Начальник отдела лицензирования и аккредитации
_____ Чаленко К.Н.
«01» 06 2020 г.

Рабочая программа дисциплины
Статистический анализ данных в RStudio

по профессионально-образовательной программе направление 01.03.05 "Статистика"
профиль 01.03.05.01 "Анализ больших данных"

Для набора 2020 года

Квалификация
Бакалавр

КАФЕДРА **Статистики, эконометрики и оценки рисков**

Семестр (<Курс>.<Семестр на курсе>)	4 (2.2)		Итого	
Недель	16			
Вид занятий	УП	РП	УП	РП
Лабораторные	64	64	64	64
Итого ауд.	64	64	64	64
Контактная работа	64	64	64	64
Сам. работа	188	188	188	188
Часы на контроль	36	36	36	36
Итого	288	288	288	288

ОСНОВАНИЕ

Учебный план утвержден учёным советом вуза от 25.02.2020 протокол № 8.

Программу составил(и): к.э.н., доцент Кракашова О.А. _____

Зав. кафедрой: д.э.н., проф. Ниворожкина Л.И. _____

Методическим советом направления: к.э.н., доцент Кислая И.А. _____

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	научить обучающихся самостоятельно ставить и решать задачи статистического анализа lfyys[в среде RStudio

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ПК-2:способностью самостоятельно осуществлять постановку задачи статистического анализа и оценивания в избранной предметной области, выбор и применение статистического инструментария и программных средств

ПК-3:способностью самостоятельно осваивать новые методы прикладной и математической статистики для их использования в аналитической работе

ПК-4:способностью осознанно применять методы математической и дескриптивной статистики для анализа количественных данных, содержательно интерпретировать полученные результаты

ПК-8:способностью формировать входные массивы статистических данных в соответствии с заданными признаками и процедурами

ПК-9:способностью осуществлять расчет сводных и производных показателей в соответствии с утвержденными методиками, в том числе с применением необходимой вычислительной техники и стандартных компьютерных программ

В результате освоения дисциплины обучающийся должен:

Знать:

основные методы статистического анализа больших данных, возможности их применения в RStudio;
современные методы анализа и моделирования больших данных;
методы математической и дескриптивной статистики для анализа больших данных;
способы формирования массивов статистических данных, требования, предъявляемые к статистической информации, источники больших данных
методику расчета сводных и производных показателей, в том числе в RStudio

Уметь:

ставить задачи по статистическому анализу больших данных, выбирать инструменты анализа в RStudio;
проводить расчеты и интерпретировать статистические показатели по современным методикам;
применять методы статистического анализа в прикладных исследованиях, содержательно интерпретировать полученные результаты;
моделировать и проектировать структуру баз больших данных;
рассчитывать и интерпретировать основные статистические показатели, в том числе в RStudio

Владеть:

навыками анализа больших данных, в том числе с помощью RStudio;
современной методикой анализа больших данных;
прикладными методами анализа больших данных, средствами содержательной интерпретации полученных результатов;
навыками формирования массивов больших данных в соответствии с заданными признаками и процедурами;
методикой расчета сводных и обобщающих показателей рядов больших данных, моделирования и прогнозирования, в том числе с помощью RStudio

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Компетен-ции	Литература
	Раздел 1. Введение в язык программирования R и пакет RStudio				
1.1	Тема "Начало работы и получение справочной информации в R. Пакет RStudio". Особенности программирования на языке R. Специфические особенности пакета RStudio. Интерфейс RStudio, разметка по умолчанию. Начало работы в RStudio. Ввод команд. Выход или прерывание расчетов. Просмотр прилагаемой документации. Определение рабочей папки. Создание нового проекта RStudio. Получение справки по функции. Получение справки по пакету. Поиск справки в прилагаемой документации и Интернете. R как калькулятор: полезные операторы. Создание и удаление переменных: присвоение. Рабочее пространство. История команд. Типы данных. Скрипты. Пакеты: установка и активация. Поиск соответствующих функций и пакетов. /Лаб/	4	4	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5

1.2	Тема "Начало работы и получение справочной информации в R. Пакет RStudio". Особенности программирования на языке R. Специфические особенности пакета RStudio. Интерфейс RStudio, разметка по умолчанию. Начало работы в RStudio. Ввод команд. Выход или прерывание расчетов. Просмотр прилагаемой документации. Определение рабочей папки. Получение справки по функции. Получение справки по пакету. Поиск справки в прилагаемой документации и Интернете. R как калькулятор: полезные операторы. Создание и удаление переменных: присвоение. Рабочее пространство. История команд. Типы данных. Скрипты. Пакеты: установка и активация. Поиск соответствующих функций и пакетов. /Ср/	4	2	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
1.3	Тема "Ввод и вывод данных в RStudio". Ввод данных различных типов с клавиатуры. Чтение данных из файлов. Структуры данных. Векторы, действия с векторами. Последовательности. Матрицы, массивы, таблицы: общая информация, создание, структура, выбор элементов, арифметические операции. Доступ к встроенным наборам данных. Преобразование данных. Перенаправление вывода в файл. Список файлов. /Лаб/	4	6	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
1.4	Тема "Ввод и вывод данных в RStudio". Ввод данных различных типов с клавиатуры. Чтение данных из файлов. Структуры данных. Векторы, действия с векторами. Последовательности. Матрицы, массивы, таблицы: общая информация, создание, структура, выбор элементов, арифметические операции. Доступ к встроенным наборам данных. Преобразование данных. Перенаправление вывода в файл. Список файлов. /Ср/	4	10	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
1.5	Тема "Списки, циклы и функции в RStudio". Факторы. Работа со списками. Циклы в RStudio. Функции. /Лаб/	4	4	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
1.6	Тема "Списки, циклы и функции в RStudio". Факторы. Работа со списками. Циклы в RStudio. Функции. /Ср/	4	18	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
1.7	Тема "Временные ряды. Ввод и подготовка данных к анализу". Импорт временных рядов из внешнего источника. Алгоритм работы с временными рядами. Пропуски в данных. Трансформация таблиц в специальный формат временных рядов. /Лаб/	4	6	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
1.8	Тема "Временные ряды. Ввод и подготовка данных к анализу". Импорт временных рядов из внешнего источника. Алгооритм работы с временными рядами. Пропуски в данных. Трансформация таблиц в специальный формат временных рядов. /Ср/	4	14	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
1.9	Тема «R Markdown и публикации». Создание нового документа. Добавление заголовка, автора или даты. Форматирование текста документа. Вставка заголовков документов. Вставка списка. Вывод результатов из кода R. Контролируем, какой код и результаты отображаются. Вставка графика. Вставка таблицы. Вставка таблицы данных. Вставка математических уравнений. Генерация вывода HTML. Генерация вывода в формате PDF. Генерация вывода в формате Microsoft Word. Генерация выходных данных презентации. Создание параметризованного отчета. Организация рабочего процесса в R Markdown. /Лаб/	4	6	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5

1.10	Тема «R Markdown и публикации». Создание нового документа. Добавление заголовка, автора или даты. Форматирование текста документа. Вставка заголовков документов. Вставка списка. Вывод результатов из кода R. Контролируем, какой код и результаты отображаются. Вставка графика. Вставка таблицы. Вставка таблицы данных. Вставка математических уравнений. Генерация вывода HTML. Генерация вывода в формате PDF. Генерация вывода в формате Microsoft Word. Генерация выходных данных презентации. Создание параметризованного отчета. Организация рабочего процесса в R Markdown. /Ср/	4	18	ПК-3 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
	Раздел 2. Анализ данных в RStudio				
2.1	Тема "Описательная статистика в R (RStudio)". Основные числовые характеристики в RStudio. Сводка данных. Расчет относительных частот. Представление данных в виде таблицы. Создание таблиц сопряженности. Проверка категориальных переменных на независимость. Расчет квантилей (и квартилей) набора данных. Инвертирование квантиля. Стандартизация. Проверка распределения на нормальность. Тест последовательностей. Сравнение законов распределения двух выборок. /Лаб/	4	6	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
2.2	Тема "Описательная статистика в R (RStudio)". Основные числовые характеристики в RStudio. Сводка данных. Расчет относительных частот. Представление данных в виде таблицы. Создание таблиц сопряженности. Проверка категориальных переменных на независимость. Расчет квантилей (и квартилей) набора данных. Инвертирование квантиля. Стандартизация. Проверка распределения на нормальность. Тест последовательностей. Сравнение законов распределения двух выборок. /Ср/	4	24	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
2.3	Тема " Визуализация данных в R и RStudio". Графические пакеты и их основные функции. Функции и аргументы. Диаграмма рассеяния. Выделение подмножеств цветом и размером. Несколько ситем координат. Слои. Гистограмма распределения. Диаграммы и графики. "Ящик с усами", гистограмма частот, график функции плотности нормального распределения. Добавление заголовка и меток. Добавление (или удаление) координатной сетки. Применение темы к графику ggplot. Добавление (или удаление) условных обозначений. Построение регрессионной линии точечной диаграммы. Создание гистограммы. Добавление доверительных интервалов в гистограмму. Раскраска гистограммы. Построение линии из точек x и y. Изменение типа, ширины или цвета линии. Построение нескольких наборов данных. Добавление вертикальных или горизонтальных линий. Создание диаграммы размаха. Создание диаграммы размаха для каждого уровня фактора. Создание гистограммы. Добавление оценки плотности к гистограмме. Создание графиков квантиль-квантиль. Построение переменной в нескольких цветах. График функции. Отображение нескольких графиков на одной странице. /Лаб/	4	4	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5

2.4	Тема " Визуализация данных в R и RStudio". Графические пакеты и их основные функции. Функции и аргументы. Диаграмма рассеяния. Выделение подмножеств цветом и размером. Несколько ситем координат. Слои. Гистограмма распределения. Диаграммы и графики. "Ящик с усами", гистограмма частот, график функции плотности нормального распределения. Добавление заголовка и меток. Добавление (или удаление) координатной сетки. Применение темы к графику ggplot. Добавление (или удаление) условных обозначений. Построение регрессионной линии точечной диаграммы. Создание гистограммы. Добавление доверительных интервалов в гистограмму. Раскраска гистограммы. Построение линии из точек x и y. Изменение типа, ширины или цвета линии. Построение нескольких наборов данных. Добавление вертикальных или горизонтальных линий. Создание диаграммы размаха. Создание диаграммы размаха для каждого уровня фактора. Создание гистограммы. Добавление оценки плотности к гистограмме. Создание графиков квантиль-квантиль. Построение переменной в нескольких цветах. График функции. Отображение нескольких графиков на одной странице. /Ср/	4	18	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
2.5	Тема "Анализ вариационных рядов". Построение дискретного и интервального вариационного ряда. Расчет числовых характеристик вариационного ряда. Эмпирическая функция распределения. Построение графиков: полигон, гистограмма, кумулята и огива. Правило сложения дисперсии. Эмпирическое корреляционное отношение и коэффициент детерминации. Расчет начальных и центральных моментов вариационного ряда. Расчет коэффициентов асимметрии и эксцесса. Расчет коэффициента корреляции Пирсона. Интерпретация полученных результатов. /Лаб/	4	4	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
2.6	Тема "Анализ вариационных рядов". Построение дискретного и интервального вариационного ряда. Расчет числовых характеристик вариационного ряда. Эмпирическая функция распределения. Построение графиков: полигон, гистограмма, кумулята и огива. Правило сложения дисперсии. Эмпирическое корреляционное отношение и коэффициент детерминации. Расчет начальных и центральных моментов вариационного ряда. Расчет коэффициентов асимметрии и эксцесса. Расчет коэффициента корреляции Пирсона. Интерпретация полученных результатов. /Ср/	4	8	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
2.7	Тема «Анализ данных, измеренных на номинальной и порядковой шкалах». Номинальные и порядковые данные. Расчет коэффициентов ассоциации и контингенции, коэффициент взаимной сопряженности К.Пирсона, ранговых коэффициентов корреляции Спирмена и Кендалла. Особенности их вычисления при наличии связанных рангов. Коэффициент конкордации как характеристика связи между несколькими признаками, измеренными на порядковой шкале. Интерпретация полученных результатов. /Лаб/	4	6	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
2.8	Тема «Анализ данных, измеренных на номинальной и порядковой шкалах». Номинальные и порядковые данные. Расчет коэффициентов ассоциации и контингенции, коэффициент взаимной сопряженности К.Пирсона, ранговых коэффициентов корреляции Спирмена и Кендалла. Особенности их вычисления при наличии связанных рангов. Коэффициент конкордации как характеристика связи между несколькими признаками, измеренными на порядковой шкале. Интерпретация полученных результатов. /Ср/	4	22	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5

2.9	Тема «Проверка статистических гипотез». Проверка гипотез о числовых значениях параметров. Проверка гипотез о равенстве средних двух и более совокупностей. Проверка гипотез о равенстве долей двух и более совокупностей. Проверка гипотез о равенстве дисперсий двух и более совокупностей. Проверка гипотез о законе распределения. Сравнение двух вероятностей биномиальных распределений. Проверка гипотезы о значимости выборочного коэффициента корреляции. Выборочный коэффициент ранговой корреляции Спирмена и проверка гипотезы о его значимости. Выборочный коэффициент корреляции Кендалла и проверка гипотезы о его значимости. Критерий Вилкоксона и проверка об однородности двух выборок. /Лаб/	4	8	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
2.10	Тема «Проверка статистических гипотез». Проверка гипотез о числовых значениях параметров. Проверка гипотез о равенстве средних двух и более совокупностей. Проверка гипотез о равенстве долей двух и более совокупностей. Проверка гипотез о равенстве дисперсий двух и более совокупностей. Проверка гипотез о законе распределения. Сравнение двух вероятностей биномиальных распределений. Проверка гипотезы о значимости выборочного коэффициента корреляции. Выборочный коэффициент ранговой корреляции Спирмена и проверка гипотезы о его значимости. Выборочный коэффициент корреляции Кендалла и проверка гипотезы о его значимости. Критерий Вилкоксона и проверка об однородности двух выборок. /Ср/	4	20	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
2.11	Тема «Основы анализа и моделирования тенденции развития рядов динамики». Применение методов преобразования рядов динамики. Расчет аналитических показателей изменения уровней рядов динамики. Анализ компонент ряда динамики. Выявление основной тенденции (тренда) в рядах динамики методами укрупнения интервалов,скользящего среднего и аналитического выравнивания ряда динамики. Экстраполяция ряда динамики. Построение индексов сезонности. Интерпретация полученных результатов. /Лаб/	4	6	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
2.12	Тема «Основы анализа и моделирования тенденции развития рядов динамики». Применение методов преобразования рядов динамики. Расчет аналитических показателей изменения уровней рядов динамики. Анализ компонент ряда динамики. Выявление основной тенденции (тренда) в рядах динамики методами укрупнения интервалов,скользящего среднего и аналитического выравнивания ряда динамики. Экстраполяция ряда динамики. Построение индексов сезонности. Интерпретация полученных результатов. /Ср/	4	20	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
2.13	Тема "Контент-анализ текстов в RStudio". Анализ текстовой информации в R(RStudio). Контент-анализ текстового фрагмента. Построение облака тегов. /Лаб/	4	4	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.4 Л2.5
2.14	Тема "Контент-анализ текстов в RStudio". Анализ текстовой информации в R(RStudio). Контент-анализ текстового фрагмента. Построение облака тегов. /Ср/	4	14	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5
2.15	/Экзамен/	4	36	ПК-2 ПК-9 ПК-3 ПК-4 ПК-8	Л1.1 Л1.2 Л1.3Л2.1 Л2.2 Л2.3 Л2.4 Л2.5

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Структура и содержание фонда оценочных средств для проведения текущей и промежуточной аттестации представлены в Приложении 1 к рабочей программе дисциплины.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л1.1	Ниворожкина Л. И.	Статистические методы анализа данных: учеб.	М.: РИО, 2016	105
Л1.2	Минашкин В. Г., Садовникова Н. А., Шмойлова Р. А.	Бизнес-статистика и прогнозирование: учебно-практическое пособие: учебное пособие	Москва: Евразийский открытый институт, 2010	http://biblioclub.ru/index.php?page=book&id=90810 неограниченный доступ для зарегистрированных пользователей
Л1.3	Лемешко, Б. Ю., Лемешко, С. Б., Постовалов, С. Н., Чимитова, Е. В.	Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: монография	Новосибирск: Новосибирский государственный технический университет, 2011	http://www.iprbookshop.ru/47719.html неограниченный доступ для зарегистрированных пользователей

5.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л2.1	Ниворожкина Л. И., Морозова З. А., Гурьянова И. Э., Ниворожкина Л. И.	Математическая статистика с элементами теории вероятностей в задачах с решениями: учеб. пособие для студентов вузов, обучающихся по напр. подгот. "Экономика", "Менеджмент", "Упр. персоналом", "Гос. и муницип. упр.", "Бизнес-информатика" (квалификация (степень) "бакалавр")	М.: Дашков и К, 2016	251
Л2.2	Марков А. С., Лисовский К. Ю.	Базы данных. Введение в теорию и методологию: учеб.	М.: Финансы и статистика, 2006	50
Л2.3		Журнал "Вопросы статистики"	,	1
Л2.4	Чубукова, И. А.	Data Mining: учебное пособие	Москва, Саратов: Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020	http://www.iprbookshop.ru/89404.html неограниченный доступ для зарегистрированных пользователей
Л2.5	Каган Е. С.	Прикладной статистический анализ данных: учебное пособие	Кемерово: Кемеровский государственный университет, 2018	http://biblioclub.ru/index.php?page=book&id=573550 неограниченный доступ для зарегистрированных пользователей

5.3 Профессиональные базы данных и информационные справочные системы

Статистика Центрального банка Российской Федерации. http://www.cbr.ru/statistics/

Статистика Федеральной службы государственной статистики https://rosstat.gov.ru/statistic

Единая межведомственная информационно – статистическая система (ЕМИСС) https://fedstat.ru/

База данных показателей муниципальных образований https://rosstat.gov.ru/storage/mediabank/munst.htm

СПС «Консультант Плюс»

5.4. Перечень программного обеспечения

RStudio

R язык программирования

5.5. Учебно-методические материалы для студентов с ограниченными возможностями здоровья

При необходимости по заявлению обучающегося с ограниченными возможностями здоровья учебно-методические материалы предоставляются в формах, адаптированных к ограничениям здоровья и восприятия информации. Для лиц с нарушениями зрения: в форме аудиофайла; в печатной форме увеличенным шрифтом. Для лиц с нарушениями слуха: в форме электронного документа; в печатной форме. Для лиц с нарушениями опорно-двигательного аппарата: в форме электронного документа; в печатной форме.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Помещения для проведения всех видов работ, предусмотренных учебным планом, укомплектованы необходимой специализированной учебной мебелью и техническими средствами обучения. Лабораторные занятия проводятся в компьютерных классах, рабочие места в которых оборудованы необходимыми лицензионными программными средствами и выходом в Интернет.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)
Методические указания по освоению дисциплины представлены в Приложении 2 к рабочей программе дисциплины.

Приложение 1

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

1.1 Показатели и критерии оценивания компетенций:

ЗУН, составляющие компетенцию	Показатели оценивания	Критерии оценивания	Средства оценивания
ПК-2: способностью самостоятельно осуществлять постановку задачи статистического анализа и оценивания в избранной предметной области, выбор и применение статистического инструментария и программных средств			
Знать: основные методы статистического анализа больших данных, возможности их применения в RStudio	Формулирует ответы на поставленные вопросы; решает тестовое задание в части методов анализа больших данных	Полнота и содержательность ответа; умение приводить примеры	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 1-48) Т – тест (Т 1-16), О – опрос (О 1-115)
Уметь: ставить задачи по статистическому анализу больших данных, выбирать инструменты анализа в RStudio	Выполняет контрольные задания, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе с использованием RStudio	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
Владеть: навыками анализа больших данных, в том числе с помощью RStudio	Выполняет контрольные задания, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе с RStudio	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
ПК-3: способностью самостоятельно осваивать новые методы прикладной и математической статистики для их использования в аналитической работе			
Знать: современные методы анализа и моделирования больших данных	Формулирует ответы на поставленные вопросы; решает тестовое задание в части методов анализа больших данных	Полнота и содержательность ответа; умение приводить примеры	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 1-48) Т – тест (Т 1-16), О – опрос (О 1-115)
Уметь: проводить расчеты и интерпретировать статистические показатели по современным методикам	Выполняет контрольные задания, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе с RStudio в части расчета статистических показателей	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
Владеть: современной методикой анализа больших данных	Выполняет контрольные задания, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе с RStudio в части методики расчета статистических показателей по большим данным	Полнота и правильность решений; обоснованность обращения к базам данных и выборе методики; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
ПК-4: способностью осознанно применять методы математической и дескриптивной статистики для анализа количественных данных, содержательно интерпретировать полученные результаты			
Знать: методы математической и дескриптивной статистики для анализа больших данных	Формулирует ответы на поставленные вопросы; решает тестовое задание в части методов математической и дескриптивной статистики.	Полнота и содержательность ответа; умение приводить примеры	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 1-48) Т – тест (Т 1-16), О – опрос (О 1-115)

Уметь: применять методы статистического анализа в прикладных исследованиях, содержательно интерпретировать полученные результаты	Выполняет контрольные задания, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе в части использования методов статистического анализа	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
Владеть: прикладными методами анализа больших данных, средствами содержательной интерпретации полученных результатов	Выполняет контрольные задания, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе в RStudio в части использования методов статистического анализа	Полнота и правильность решений; обоснованность обращения к базам данных и выборе методов анализа; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
ПК-8: способностью формировать входные массивы статистических данных в соответствии с заданными признаками и процедурами			
Знать: способы формирования массивов статистических данных, требования, предъявляемые к статистической информации, источники больших данных	Формулирует ответы на поставленные вопросы; решает тестовое задание в части способов формирования массивов статистических данных	Полнота и содержательность ответа; умение приводить примеры	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 1-48) Т – тест (Т 1-16), О – опрос (О 1-115)
Уметь: моделировать и проектировать структуру баз больших данных	Моделирует и проектирует структуру баз данных при выполнении задания к лабораторной работе	Обоснованность решений при построении баз данных	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
Владеть: навыками формирования массивов больших данных в соответствии с заданными признаками и процедурами	Моделирует и проектирует структуру баз данных при выполнении задания к лабораторной работе в соответствии с заданными признаками и процедурами	Обоснованность решений при формировании массивов больших данных	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
ПК-9: способностью осуществлять расчет сводных и производных показателей в соответствии с утвержденными методиками, в том числе с применением необходимой вычислительной техники и стандартных компьютерных программ			
Знать: методику расчета сводных и производных показателей, в том числе в RStudio	Формулирует ответы на поставленные вопросы; решает тестовое задание в части расчета сводных и производных показателей больших данных	Полнота и содержательность ответа; умение приводить примеры	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 1-48) Т – тест (Т 1-16), О – опрос (О 1-115)
Уметь: рассчитывать и интерпретировать основные статистические показатели, в том числе в RStudio	Выполняет контрольные задания в части расчета основных статистических показателей, в том числе в RStudio, формирует отчет по заданию к лабораторной работе	Полнота и правильность решений; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)
Владеть: методикой расчета сводных и обобщающих показателей рядов больших данных, моделирования и прогнозирования, в том числе с помощью RStudio	Выполняет контрольные задания, используя стандартные методики расчета сводных и обобщающих показателей по большим данным, анализирует и интерпретирует полученные результаты, формирует отчет по заданию к лабораторной работе, выполненных в RStudio	Полнота и правильность решений; обоснованность выбора методики; содержательность выводов и интерпретации полученных результатов	ВЗЭ – вопросы и практико-ориентированные задания к экзамену (ВЗЭ 49-63) ЛР – задание к лабораторной работе (ЛР 1-9), КЗ – контрольное задание (З 1-15)

1.2 Шкалы оценивания:

Текущий контроль успеваемости и промежуточная аттестация осуществляется в рамках накопительной балльно-рейтинговой системы в 100-балльной шкале:

84-100 баллов (оценка «отлично»)
67-83 баллов (оценка «хорошо»)
50-66 баллов (оценка «удовлетворительно»)
0-49 баллов (оценка «неудовлетворительно»)

2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Вопросы и практико-ориентированные задания к экзамену

1. Особенности программирования на языке R. Специфические особенности пакета RStudio. Интерфейс RStudio.
2. Типы и структуры данных.
3. Скрипты. Основные пакеты.
4. Векторы, действия с векторами.
5. Последовательности.
6. Матрицы, массивы, таблицы.
7. Списки.
8. Циклы в RStudio.
9. Функции.
10. Создание нового документа. Форматирование текста документа. Вставка заголовков документов. Вставка списка.
11. Таблицы и графики.
12. Вывод данных в разных форматах (HTML, PDF, Microsoft Word, CSV…).
13. Организация рабочего процесса в R Markdown.
14. Основные числовые характеристики в RStudio.
15. Сводка данных.
16. Расчет квантилей (и квартилей) набора данных. Инвертирование квантиля.
17. Стандартизация.
18. Проверка распределения на нормальность.
19. Сравнение законов распределения двух выборок.
20. Построение дискретного и интервального вариационного ряда.
21. Расчет числовых характеристик вариационного ряда.
22. Эмпирическая функция распределения.
23. Построение графиков: полигон, гистограмма, кумулята и огива.
24. Правило сложения дисперсий.
25. Эмпирическое корреляционное отношение и коэффициент детерминации.
26. Расчет начальных и центральных моментов вариационного ряда.
27. Расчет коэффициентов асимметрии и эксцесса.
28. Расчет коэффициента корреляции Пирсона. Проверка гипотезы о значимости выборочного коэффициента корреляции.
29. Выборочный коэффициент ранговой корреляции Спирмена и проверка гипотезы о его значимости.
30. Выборочный коэффициент корреляции Кендалла и проверка гипотезы о его значимости.
31. Критерий Вилкоксона и проверка об однородности двух выборок.
32. Расчет коэффициентов ассоциации и контингенции, коэффициент взаимной сопряженности К.Пирсона,
33. Коэффициент конкордации как характеристика связи между несколькими признаками, измеренными на порядковой шкале.
34. Проверка гипотезы о средней.
35. Определение границ доверительных интервалов для средней и медианы.
36. Проверка гипотезы о доле.
37. Определение границ доверительного интервала для доли.
38. Проверка гипотез о равенстве средних двух и более совокупностей.

39. Проверка гипотез о равенстве долей двух и более совокупностей.
40. Проверка гипотез о равенстве дисперсий двух и более совокупностей.
41. Проверка гипотез о законе распределения.
42. Преобразования рядов динамики.
43. Расчет аналитических показателей изменения уровней рядов динамики.
44. Анализ компонент ряда динамики.
45. Выявление основной тенденции (тренда) в рядах динамики методами укрупнения интервалов, скользящего среднего и аналитического выравнивания ряда динамики.
46. Экстраполяция ряда динамики.
47. Построение индексов сезонности.
48. Контент-анализ текстового фрагмента. Облако тегов.
49. Используйте массив данных *rivers*, в котором представлены длины (в милях) 141 основных рек в Северной Америке. Подгрузите этот массив с помощью команды *data(rivers)*. Чему равна средняя длина этих рек?
50. Используйте массив данных *rivers*, в котором представлены длины (в милях) 141 основных рек в Северной Америке. Подгрузите этот массив с помощью команды *data(rivers)*. Во сколько раз наибольшее значение длины реки превышает наименьшее значение длины реки?
51. Чему равна вероятность того, что случайная величина X, которая распределена N(78, 144), будет лежать в промежутке от 24 до 85?
52. Сколько миль налетали пассажиры в Америке, на примере встроенного массива данных *airmiles*. За какой год есть первые наблюдения?
53. Чему равняется *length(c(7,7,7))*
54. Что вернет выражение *sum(1:3<2)*?
55. American Community Survey предоставляет скачиваемые данные из различных обследований общества в Соединенных Штатах. С помощью команды download.file () скачать данные из опроса о жилье в штате Айдахо в 2006 г. с сайта : https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv
Загрузите эти данные в R. Книга кодирования, описывающая имена переменных находится по адресу: https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf
Сколько категорий стоимостью $ 1 млн или больше?
56. Используйте данные из задания 1. Рассмотрим переменную FES. Какой из принципов "аккуратных данных" (tidy data) нарушаются в этой переменной?
57. Скачать Excel таблицу из данных Natural Gas Aquisition Program по адресу: 26https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx (original data source: http://catalog.data.gov/dataset/natural-gas-acquisition-program) Прочитайте строки 18-23 и столбцы 7-15 в R и присвойте результат переменной с именем dat
Чему равно значение выражения sum(dat$Zip*dat$Ext,na.rm=T)?
58. Прочитайте XML данные о ресторанах г.Балтимора с сайта: https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml
Сколько ресторанов имеют zipcode 21231?
59. Скачайте данные опроса 2006 г. о жилье для штата Айдахо с помощью команды download.file () по адресу:
https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv
Используя команду fread() загрузите данные в R, назовите объект DT
Что из перечисленного ниже является самым быстрым способом для расчета средних значений переменной pwgtp15 для мужчин и женщин с использованием пакета data.table?
60. Рассмотрим набор переменных: Имя; Год рождения; Телефон; Кол-во сестер(братьев); Годовой доход. Например, Анна; 1975; 8929223; 0; 66000.
Какие переменные качественные, а какие - количественные?
61. В социальном обследовании, проводимом ежегодно в Соединенных Штатах, спрашивается, сколько друзей у людей (number of friends) и как они оценивают свой уровень счастья (very happy, pretty happy, not too happy).
Для того чтобы оценить связь между этими двумя переменными исследователь вычисляет среднее количество друзей для людей, которые классифицировали себя как очень счастлив, довольно счастлив, и не слишком счастлив.
Какие переменные независимые? зависимые ?
62. В исследовании, опубликованном в 2011 PNAS USA, 120 пожилых мужчин и женщин (средний возраст около 65 лет), которые добровольно согласились участвовать в этом исследовании, были случайным образом распределены в две группы.
В первой группе добровольцы ходили по дорожке в парке три раза в неделю; в другой - делали множество менее аэробных упражнений, в том числе йогу и тренировки с отягощением.
Через год сканирование мозга показало, что у «пешеходов» гиппокамп (часть мозга, отвечающая за формирование воспоминаний) увеличился в объеме в среднем примерно на 2%; в другой группе объем гиппокампа снизился на 1,4%.
Что из перечисленного ниже ложно?
63. В одном американском городе был проведен опрос о жилье, чтобы определить цену типичного дома в городе, в котором проживает в основном средний класс, но есть очень дорогой пригород. Средняя стоимость дома в этом городе примерно $ 650 000.
Верно ли, что большинство домов в этом городе стоят более $ 650 000?

Критерии оценивания:

- 84-100 баллов, оценка «отлично» выставляется, если ответы обучающегося на оба теоретических вопроса фактически верны, проявлены глубокие исчерпывающие знания в объеме пройденной программы дисциплины в соответствии с поставленными программой курса целями и задачами обучения; успешно решена задача, дана содержательная интерпретация полученных при решении задачи результатов; изложение материала при ответе - грамотное и логически стройное;
- 67-83 балла, оценка «хорошо» выставляется, если при ответах на оба теоретических вопроса обучающимся проявлено наличие твердых и достаточно полных знаний в объеме пройденной программы дисциплины в соответствии с целями обучения, в целом успешно решена задача, дана содержательная интерпретация полученных при решении задачи результатов; материал изложен четко, допускаются отдельные логические и стилистические погрешности;
- 50-66 баллов, оценка удовлетворительно выставляется, если при ответах на оба теоретических вопроса обучающимся проявлено наличие твердых знаний в объеме пройденного курса в соответствии с целями обучения, ответы изложены с отдельными ошибками, уверенно исправленными после дополнительных вопросов; ход решения задачи в целом – правильный, допускаются незначительные погрешности в интерпретации полученных результатов, незначительные ошибки в решении, в целом не повлиявшие на результат, уверенно исправленные после дополнительных вопросов;
- 0-49 баллов, оценка неудовлетворительно выставляется, если при ответах на оба теоретических вопроса обучающимся допущены грубые ошибки, проявлено непонимание сущности излагаемого вопроса, не решена или не полностью решена задача, ответы на дополнительные и наводящие вопросы - неуверенны и неточны.

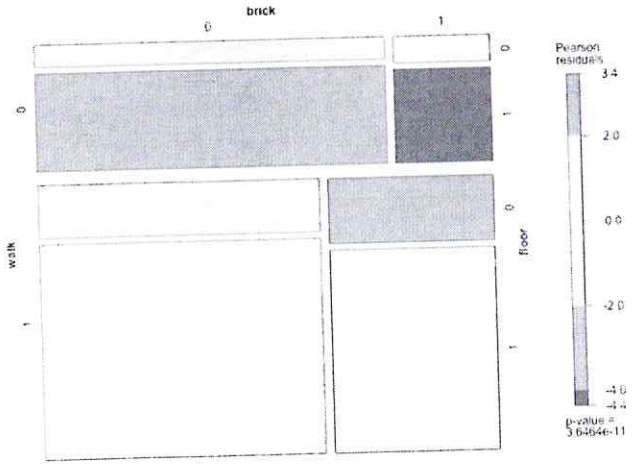
Тесты

1. Используйте встроенный набор данных *iris*. Введите в R команду *data(iris)*. Значение, стоящее в 3-ой строчке 4-ого столбца – это:
a) 4.6; b) 0.2; c) 0.4; d) 1.3.

2. Чтобы начать работать с функциями из какого-либо определённого пакета (например, *ggplot2*), необходимо:
a) Каждый раз устанавливать пакет через *Tools → Install pacages*, но один раз подгрузить библиотеку *library(ggplot2)*;
b) Один раз установить пакет через *Tools → Install pacages* и каждый раз при начале работы подгружать библиотеку *library(ggplot2)*;
c) Каждый раз устанавливать пакет через *Tools→Install pacages* и каждый раз при начале работы подгружать библиотеку *library(ggplot2)*;
d) Один раз установить пакет через *Tools → Install pacages* и один раз подгрузить библиотеку *library(ggplot2)*.

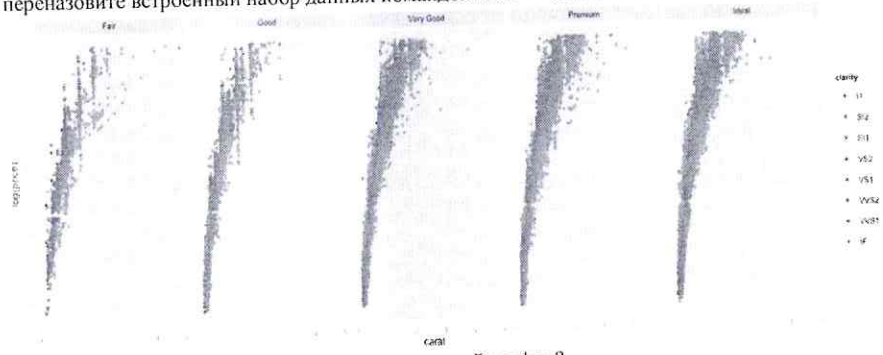
3. С помощью каких двух команд можно посмотреть некоторые строчки набора данных?
a) *head(…)*; b) *nrow(…)*; c) *help(…)*; d) *glimpse(…)*.

4. Ниже представлен мозаичный график по данным о квартирах, где *walk* обозначает, в пешей доступности ли квартира от метро или нет, *brick* обозначает, в кирпичном доме ли квартира или нет, а *floor* обозначает, находится ли квартира на первом или последнем этаже или нет.



Проинтерпретируйте данный график.
a) Квартир, которые находятся в пешей доступности от метро, являются кирпичными и не находятся на первом или последнем этажах, - меньше, чем в любой другой группе.
b) Квартир, которые находятся не в пешей доступности от метро, являются кирпичными и не находятся на первом или последнем этажах, - меньше, чем в любой другой группе.
c) Квартир, которые находятся не в пешей доступности от метро, являются кирпичными и находятся на первом или последнем этажах, - меньше, чем в любой другой группе.
d) Квартир, которые находятся в пешей доступности от метро, не являются кирпичными и не находятся на первом или последнем этажах, - меньше, чем в любой другой группе.

5. График построен по данным diamonds из пакета *ggplot2*. Подключив пакет *library(ggplot2)*, переназовите встроенный набор данных командой *diam <- diamonds*.



С помощью какой команды можно получить такой график?
a) *ggplot(data=diam, aes(log(carat),log(price))) + geom_point(aes(color=cut)) + facet_wrap(~cut)*;
b) *ggplot(data=diam, aes(carat,log(price))) + geom_point(aes(color=clarity)) + facet_grid(~cut)*;
c) *ggplot(data=diam, aes(log(carat),log(price))) + geom_point(aes(color=cut)) + facet_grid(~cut)*;
d) *ggplot(data=diam, aes(log(price),carat)) + geom_point(aes(color=clarity)) + facet_wrap(~cut)*.

6. Какая из предложенных команд поможет построить диаграмму рассеяния (scatterplot) по данным *rivers*?
a) *qplot(df=rivers, rivers)*;
b) *qplot(df=rivers)*;
c) *plot(rivers)*;
d) *plot(df=rivers)*.

7. Для каких данных подходит мозаичный график: *mosaic(...)* ?
a) Для данных, где есть только переменные с типом *numeric*.
b) Для данных, где много переменных с типом *factor*.
c) Для данных, где нет переменных с типом *factor*.

8. Для количественной переменной можно построить
a) только мозаичный график;
b) график плотности, но не гистограмму;
c) график плотности и гистограмму;
d) гистограмму, но не график плотности.

9. Какой именно день недели был 13.04.1940?
Подсказка: используйте функцию *wday()* из пакета *library(lubridate)*. Можно включить опцию
label = TRUE. Не пятница 13? :)
a) Понедельник;
b) Пятница;
c) Суббота;
d) Воскресенье.

10. С помощью какой функции можно реализовать линейную аппроксимацию для пропущенных наблюдений?

a) *na.locf()*;
b) *na.true()*;
c) *na.approx()*;
d) *na.omit()*.

11. Команда *vcovHAC(...)* позволяет получить, ковариационную матрицу, которая…
a) имеет поправку только на мультиколлинеарность;
b) имеет поправку только на автокорреляцию;
c) имеет поправку на гетероскедастичность и автокорреляцию;
d) имеет поправку только на гетероскедастичность.

12. Какие из вариантов присвоения не вызовут ошибки:
a) x -> 3;
b) x <- 3;
c) 3 <- x;
d) 3 -> x.

13. Как сделать вектор из трех чисел?
a) (7,7,7);
b) [7,7,7];
c) {7,7,7};
d) c(7,7,7)

14. Для каких аргументов функция is.finite вернет true?
a) 1;
b) NA;
c) NaN;
d) (+Inf).

15. Что вернет выражение *sum(1:3>2)*?
a) 0; b) 1; c) 2; d) 3. Ошибка.

16. Каким образом можно сформировать вектор (FALSE, FALSE, TRUE)?
a) c(FALSE, FALSE, TRUE);
b) 2:4>3;
c) 2:4<3;
d) -c(TRUE, TRUE, FALSE).

Критерии оценки:
Максимальное количество баллов – 10.
Из имеющегося банка тестов в каждом семестре формируется тестовое задание, содержащее 10 тестов для соответствующего семестра.
Правильный ответ на каждый тест оценивается в 1 балл, неправильный – 0 баллов.

Вопросы для опроса

1. Особенности программирования на языке R.
2. Специфические особенности пакета RStudio.
3. Интерфейс RStudio, разметка по умолчанию.
4. Начало работы в RStudio. Ввод команд. Выход или прерывание расчетов.
5. Просмотр прилагаемой документации. Определение рабочей папки.
6. Создание нового проекта RStudio.
7. Получение справки по функции. Получение справки по пакету. Поиск справки в прилагаемой документации и Интернете.
8. R как калькулятор: полезные операторы.
9. Создание и удаление переменных: присвоение.
10. Рабочее пространство. История команд.
11. Типы данных.
12. Скрипты.
13. Пакеты: установка и активация.
14. Поиск соответствующих функций и пакетов.
15. Ввод данных различных типов с клавиатуры.
16. Чтение данных из файлов.
17. Структуры данных.
18. Векторы, действия с векторами.
19. Последовательности.
20. Матрицы, массивы, таблицы: общая информация, создание, структура, выбор элементов, арифметические операции.
21. Доступ к встроенным наборам данных. Преобразование данных.
22. Перенаправление вывода в файл. Список файлов.
23. Факторы.
24. Работа со списками.
25. Циклы в RStudio.
26. Функции.
27. Импорт временных рядов из внешнего источника.
28. Алгоритм работы с временными рядами.
29. Пропуски в данных.
30. Трансформация таблиц в специальный формат временных рядов.
31. Создание нового документа. Добавление заголовка, автора или даты.
32. Форматирование текста документа. Вставка заголовков документов. Вставка списка.
33. Вывод результатов из кода R.
34. Вставка графика.
35. Вставка таблицы. Вставка таблицы данных.
36. Вставка математических уравнений.
37. Генерация вывода HTML.
38. Генерация вывода в формате PDF.
39. Генерация вывода в формате Microsoft Word.
40. Генерация выходных данных презентации. Создание параметризованного отчета.
41. Организация рабочего процесса в R Markdown.
42. Основные числовые характеристики в RStudio.
43. Сводка данных.
44. Расчет относительных частот.
45. Представление данных в виде таблицы.
46. Создание таблиц сопряженности.
47. Проверка категориальных переменных на независимость.
48. Расчет квантилей (и квартилей) набора данных. Инвертирование квантиля.

49. Стандартизация.
50. Проверка гипотезы о средней (t- критерий).
51. Определение границ доверительных интервалов для средней и медианы.
52. Проверка гипотезы о доле.
53. Определение границ доверительного интервала для доли.
54. Проверка распределения на нормальность.
55. Тест последовательностей.
56. Проверка гипотезы о равенстве двух средних.
57. Непараметрическое сравнение местоположений двух выборок.
58. Проверка значимости коэффициента корреляции.
59. Проверка гипотез о равенстве пропорций.
60. Парные сравнения между средними значениями групп.
61. Сравнение законов распределения двух выборок.
62. Графические пакеты и их основные функции.
63. Функции и аргументы.
64. Диаграмма рассеяния.
65. Выделение подмножеств цветом и размером.
66. Несколько систем координат.
67. Слои.
68. Гистограмма распределения.
69. Диаграммы и графики.
70. "Ящик с усами", гистограмма частот, график функции плотности нормального распределения.
71. Добавление заголовка и меток. Добавление (или удаление) координатной сетки.
72. Применение темы к графику ggplot. Добавление (или удаление) условных обозначений.
73. Построение регрессионной линии точечной диаграммы.
74. Создание гистограммы.
75. Добавление доверительных интервалов в гистограмму. Раскраска гистограммы.
76. Построение линии из точек x и y. Изменение типа, ширины или цвета линии.
77. Построение нескольких наборов данных.
78. Добавление вертикальных или горизонтальных линий. Создание диаграммы размаха.
79. Создание диаграммы размаха для каждого уровня фактора.
80. Добавление оценки плотности к гистограмме.
81. Создание графиков квантиль-квантиль.
82. Построение переменной в нескольких цветах.
83. График функции.
84. Отображение нескольких графиков на одной странице.
85. Построение дискретного и интервального вариационного ряда.
86. Расчет числовых характеристик вариационного ряда.
87. Эмпирическая функция распределения.
88. Построение графиков: полигон, гистограмма, кумулята и огива.
89. Правило сложения дисперсий.
90. Эмпирическое корреляционное отношение и коэффициент детерминации.
91. Расчет начальных и центральных моментов вариационного ряда.
92. Расчет коэффициентов асимметрии и эксцесса.
93. Расчет коэффициента корреляции Пирсона.
94. Номинальные и порядковые данные.
95. Расчет коэффициентов ассоциации и контингенции, коэффициент взаимной сопряженности К.Пирсона,
96. Коэффициент конкордации как характеристика связи между несколькими признаками, измеренными на порядковой шкале.
97. Проверка гипотез о числовых значениях параметров.
98. Проверка гипотез о равенстве средних двух и более совокупностей.
99. Проверка гипотез о равенстве долей двух и более совокупностей.
100. Проверка гипотез о равенстве дисперсий двух и более совокупностей.
101. Проверка гипотез о законе распределения.
102. Сравнение двух вероятностей биномиальных распределений.
103. Проверка гипотезы о значимости выборочного коэффициента корреляции.
104. Выборочный коэффициент ранговой корреляции Спирмена и проверка гипотезы о его значимости.
105. Выборочный коэффициент корреляции Кендалла и проверка гипотезы о его значимости.
106. Критерий Вилкоксона и проверка об однородности двух выборок.
107. Применение методов преобразования рядов динамики.
108. Расчет аналитических показателей изменения уровней рядов динамики.
109. Анализ компонент ряда динамики.
110. Выявление основной тенденции (тренда) в рядах динамики методами укрупнения интервалов, скользящего среднего и аналитического выравнивания ряда динамики.
111. Экстраполяция ряда динамики.
112. Построение индексов сезонности.
113. Анализ текстовой информации в R(RStudio).
114. Контент-анализ текстового фрагмента.
115. Построение облака тегов.

Критерии оценивания:

Максимальный балл – 12.

В течение семестра студент отвечает на 12 вопросов. Ответ на каждый вопрос оценивается максимум в 1 балл.

Критерии оценивания 1 вопроса:

0,84-1,0 балла выставляется студенту, если изложенный материал фактически верен, продемонстрированы глубокие исчерпывающие знания в объеме пройденной программы в соответствии с поставленными программой курса целями и задачами обучения, изложение материала при ответе - грамотное и логически стройное;

0,67-0,83 балла выставляется студенту, если продемонстрированы твердые и достаточно полные знания в объеме пройденной программы дисциплины в соответствии с целями обучения; материал изложен достаточно полно с отдельными логическими и стилистическими погрешностями;

0,5-0,66 балла выставляется студенту, если продемонстрированы твердые знания в объеме пройденного курса в соответствие с целями обучения, ответ содержит отдельные ошибки, уверенно исправленные после дополнительных вопросов;

0-0,49 балла выставляется студенту, если ответ не связан с вопросом, допущены грубые ошибки в ответе, продемонстрированы непонимание сущности излагаемого вопроса, неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Задания к лабораторным работам

Задание к лабораторной работе №1
«Начало работы и получение справочной информации в R. Пакет RStudio»

1) Запустите RStudio.
2) Зайдите в раздел Tools — Global options.
3) В разделе General: а) уберите галочку у Restore .Rdata into workspace in startup; б) выберите —Neverl у Save workspace to .Rdata on exit.
4) В разделе Sweave: "Weave .Rnw files using" выберите knitr.
5) В разделе Code - Diagnostics: выставьте все галочки.
6) Установите свежую версию Rtools.
7) Создайте папку для установки пакетов без русских букв и пробелов, например, C:/Rlib.
8) Выполните в консоли команду: system ("setx R_LIBS C:/ Rlib ") Вместо C:/Rlib должно быть имя папки, созданной для установки пакетов.
9) Перезапустите RStudio.
10) Проверьте, что R знает, куда ему ставить пакеты. Для этого выполните в консоли RStudio команду: .libPaths(). Она должна указать путь к папке C:/Rlib. После этого все пакеты будут ставиться в папку C:/Rlib.
11) Установите все необходимые пакеты R анализа данных. Чтобы увидеть установленные пакеты на пользовательской инсталляции, введите команду: library().

Составьте отчет.

Задание к лабораторной работе №2
«Ввод и вывод данных в RStudio»

1. Операторы присваивания

Воспользуйтесь оператором присваивания

```
> aa<-5;aa
[1] 5
> 6->bb;bb
[1] 6
> c=7;c
[1] 7
> cc<-dd<-8;cc
[1] 8 > dd
[1] 8
>
```

Объясните действия и полученные результаты

2. Команды

Воспользуйтесь командами:

```
> help(ls)
> help.search("norm")
> apropos("norm")
> getwd()
```

Объясните действия и полученные результаты

Самостоятельно опробуйте действие команд:

ls(),
objects(),
rm(имя_объекта)
example(имя_функции)
history(n)
setwd("имя_новой_рабочей_директории")
dir()
source('имя_файла.R')
sink('имя_файла.расширение')

3. Создайте переменные x и y типов double и integer соответственно.

```
> x=double(1)
> x=5
> y=integer(1)
> y=7
> is.integer(x)
[1] FALSE
> is.double(x)
[1] TRUE
> is.integer(y)
[1] FALSE
> is.double(y)
[1] TRUE
> is.numeric(x)
[1] TRUE
> is.numeric(y)
[1] TRUE 22
> y=integer(1)
> is.integer(y)
[1] TRUE
```

Объясните действия и полученные результаты

4. Создайте последовательности от 2 до 10 и от 10 до 2.

5. Создайте:
- Вектор (vector)
- Матрицу (matrix)
- Массив (array)
- Фактор (factor)
- Список (list)
- Таблица данных (data.frame).

6. Опробуйте:
- Логические операции сравнения
- Простейшие математические операции
- Логарифмические и экспоненциальные функции
- Функции округления
- Модуль и квадратный корень
- Специальные функции
- Тригонометрические функции

7. Опробуйте функции scan(), read.table(), read.csv(), write(), cat(), write.table(), write.csv(), write.csv().

Составьте отчет.

Задание к лабораторной работе №3
«Списки, циклы и функции в RStudio»

1. Операторы if, ifelse, for, while, repeat, break, next, switch

Пусть x и y являются векторами одинаковой длины (10). Задайте условие: если x не равен y, то берётся отношение x/y. В результате должны получить вектор, чья размерность совпадает с размерностью исходных векторов. Из каких элементов он состоит?

```
> x=1:10; y=10;1
> x
[1] 1 2 3 4 5 6 7 8 9 10
> y
[1] 10 9 8 7 6 5 4 3 2 1
> if (x<y) {x/y}
[1] 0.1000000 0.2222222 0.3750000 0.5714286 0.8333333 1.2000000
[7] 1.7500000 2.6666667 4.5000000 10.0000000
```

Объясните действия и полученные результаты

Создайте два вектора x и y одинаковой длины и присвойте переменной z либо со знаком + номер совпадающих элементов, либо со знаком – номер несовпадающих элементов.

```
> x=c(1,3,1,5,1,7,1,9)
> y=c(2,3,4,5,2,7,1,8)
> z=ifelse(x==y,1:10,(-1):(-10))
> z
[1] -1 2 -3 4 -5 6 7 -8
```

Объясните действия и полученные результаты

```
> x=1:10; y=10:1
> x
[1] 1 2 3 4 5 6 7 8 9 10
> y
[1] 10 9 8 7 6 5 4 3 2 1
> w=vector(length=10,mode='numeric')
> for(i in 1:10)
+ { if (x[i]<y[i]) {w[i]=x[i]/y[i]}
+ else {w[i]=x[i]*y[i]}
+ }
> w
[1] 0.1000000 0.2222222 0.3750000 0.5714286 0.8333333 30.0000000
[7] 28.0000000 24.0000000 18.0000000 10.0000000
```

Объясните действия и полученные результаты

```
> x=-10
> while (x<0) {z=x; x=x+1}
> z
[1] -1
```

Объясните действия и полученные результаты

```
> t=-10
> repeat{
+ if (t>0) break
+ f=log(abs(t))
+ t=t+1}
> f
[1] -Inf
```

Объясните действия и полученные результаты

```
> x=numeric(5)
> for (i in 1:5)
+ { x[i]=switch(i,cos(pi),exp(1),log2(4),log10(0.01),TRUE)}
> x
[1] -1.000000 2.718282 2.000000 -2.000000 1.000000
```

Объясните действия и полученные результаты

2. Функции

Создайте функцию, вычисляющую норму — корень квадратный из скалярного произведения векторов x и y.

```
> norm = function(x,y) sqrt(x%*%y)
> norm(1:4,2:5)
[,1]
[1,] 6.324555
```

Объясните действия и полученные результаты

Создайте функцию, находящую для произвольного числа векторов их минимальные, максимальные и средние значения.

```
fun= function (...) {
data = list(...)
n =length(data)
maxs = numeric(n)
mins = numeric(n)
means<- numeric(n)
for (i in 1:n) {
maxs[i] = max(data[[i]])
mins[i]<-min(data[[i]])
means[i]<-mean(data[[i]])
}
print(maxs)
print(mins)
print(means)
invisible(NULL)
}
```

Объясните действия и полученные результаты

```
ff=function(x){
y=sum(x)
z=cumprod(x)
}
y=ff(1:10); y
[1] 1 2 6 24 120 720 5040 40320
```

[9] 362880 3628800
Объясните действия и полученные результаты

```
ff1=function(x){
y=sum(x); z=cumprod(x)
return(y)}
ff1(1:10)
[1] 55
```

Объясните действия и полученные результаты.

Составьте отчет.

Задание к лабораторной работе №4
«Описательная статистика в R (RStudio)»

В файле CPS85 https://cloud.mail.ru/public/iBPs/tdBv8MnnY содержится 534 наблюдения о случайно выбранных работниках в USA (май 1985 г.).

Переменные:

ED – количество лет образования;
SOUTH – фиктивная переменная, равна 1, если работник проживает на юге, иначе 0;
NONWH – фиктивная переменная, равна 1, если работник не белый, иначе 0;
HISP – фиктивная переменная, равна 1, если работник латиноамериканец, иначе 0;
FE – фиктивная переменная, равна 1 для женщин, и 0 для мужчин;
MARR – фиктивная переменная, равна 1 для семейных, иначе 0;
MARRFE – фиктивная переменная, равна 1 для замужних женщин, иначе 0;
EX – число лет стажа работы (= AGE-ED-6);
EXSQ – квадрат числа лет стажа работы;
UNION – фиктивная переменная, равна 1, если имеется профсоюз на работе, иначе 0;
LNWAGE – логарифм средней часовой зарплаты;
AGE – возраст в годах;
NDEP – число детей до 18 лет в семье;
MANUF – фиктивная переменная, равна 1, если работает в обрабатывающей промышленности, иначе 0;
CONSTR – фиктивная переменная, равна 1, если работает в строительстве, иначе 0;
MANAG – фиктивная переменная, равна 1, если работа управленческая или административная, иначе 0;
SALES – фиктивная переменная, равна 1, если работает в торговле, иначе 0;
CLER – фиктивная переменная, равна 1, если работает чиновником, иначе 0;
SERV – фиктивная переменная, равна 1, если работает в сфере услуг, иначе 0;
PROF – фиктивная переменная, равна 1, если профессионально-технический работник, иначе 0.

Используя средства R (RStudio):

1. Рассчитайте основные числовые характеристики (среднюю, моду, медиану, дисперсию, стандартное отклонение, коэффициенты асимметрии и эксцесса, вариации) всех количественных переменных.
2. Определите количество и доли: мужчин и женщин, белых и небелых, состоящих и не состоящих в браке, членов и не членов профсоюза.
3. Проверьте гипотезу о нормальном распределении всех количественных переменных.
4. Стандартизируйте значения количественных переменных.
5. Сравните законы распределения логарифма заработной платы мужчин и женщин.
6. Составьте таблицу сопряженности переменных пол и членство в профсоюзной организации.
7. Дайте интерпретацию полученных результатов.
8. Составьте отчет.

Задание к лабораторной работе №5
«Визуализация данных в R и RStudio»

В файле CPS85 https://cloud.mail.ru/public/iBPs/tdBv8MnnY содержится 534 наблюдения о случайно выбранных работниках в USA (май 1985 г.).

Описание переменных приводится к заданию к лабораторной работе №4.

Используя средства R (RStudio):

1. Проиллюстрируйте с помощью всех, доступных Вам графических возможностей R (RStudio), распределения переменных, представленных в файле CPS85. Обратите внимание на необходимость демонстрации различных возможностей пакета.
2. Покажите с помощью графиков различия/сходство в распределении заработной платы:
   - мужчин и женщин;
   - членов и не членов профсоюза,
   - белых и не белых,
   - латиноамериканцев и не латиноамериканцев,
   - состоящих и не состоящих в браке.
3. Прокомментируйте полученные результаты, составьте отчет.

Задание к лабораторной работе №6
«Анализ вариационных рядов в R и RStudio»

В файле CPS85 https://cloud.mail.ru/public/iBPs/tdBv8MnnY содержится 534 наблюдения о случайно выбранных работниках в USA (май 1985 г.).

Описание переменных приводится к заданию к лабораторной работе №4.

Используя средства R (RStudio):

1. Постройте интервальные вариационные ряды распределения
   - средней часовой заработной платы,
   - стажа,
   - возраста,
   - числа лет обучения.
2. Рассчитайте основные числовые характеристики (среднюю, моду, медиану, дисперсию, стандартное отклонение, коэффициенты асимметрии и эксцесса, вариации) всех интервальных вариационных рядов.
3. Постройте графики: полигон, гистограмму, кумуляту и огиву для всех интервальных рядов.
4. Постройте матрицу коэффициентов корреляции количественных переменных, проверьте значимость коэффициентов корреляции
5. Прокомментируйте полученные результаты, составьте отчет.

Задание к лабораторной работе №7
«Проверка статистических гипотез в R и RStudio»

В файле CPS85 https://cloud.mail.ru/public/iBPs/tdBv8MnnY содержится 534 наблюдения о случайно выбранных работниках в USA (май 1985 г.).

Описание переменных приводится к заданию к лабораторной работе №4.

Используя средства R (RStudio):

1. Проверьте гипотезы о равенстве дисперсий:
   - средней часовой заработной платы мужчин и женщин,
   - стажа мужчин и женщин,
   - числа лет обучения мужчин и женщин.
2. Проверьте гипотезы о равенстве двух средних:
   - заработной платы мужчин и женщин,
   - стажа мужчин и женщин,
   - числа лет обучения мужчин и женщин.
3. Проверьте гипотез о равенстве пропорций:
   - членов профсоюза среди мужчин и женщин,
   - женщин среди работников различной расы,
4. Прокомментируйте полученные результаты, составьте отчет.

Задание к лабораторной работе №8
«Основы анализа и моделирования тенденции развития рядов динамики»

Используя средства R (RStudio) по имеющимся данным о производстве стали в РФ в 2005-2009 гг.:

1. Постройте аддитивную и мультипликативные модели временного ряда, последовательно выделив сезонную, трендовую и случайную компоненты.
2. Обоснуйте выбор модели тренда.
3. Оцените качество аддитивной и мультипликативной моделей. Выберите из них наилучшую.
4. Используя полученную модель, сделайте краткосрочный точечный прогноз.
5. Оформите отчет. Дайте интерпретацию всех полученных результатов.

	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь
	2005											
Сталь, тыс.тонн	5628	5185	5620	5513	5578	5138	5375	5530	5433	5692	5641	5929
	2006											
Сталь, тыс.тонн	5742	5251	6015	5897	6108	5935	6015	5897	5696	6002	5958	6299
	2007											
Сталь, тыс.тонн	6303	5651	6278	6120	6107	5867	6056	5820	5904	6073	5922	6269
	2008											
Сталь, тыс.тонн	6557	6145	6582	6186	6538	6249	6331	6351	5992	4824	3436	3520
	2009											
Сталь, тыс.тонн	3931	4307	4585	4432	4701	4754	5314	5543	5483	5558	5224	5530

Задание к лабораторной работе №9
«Контент-анализ текстов в RStudio»

Выберите любой понравившийся Вам текст статьи, связанной со статистическим анализом больших данных.

Используя средства R (RStudio):

1. Проведите контент-анализ выбранной статьи.
2. Определите:
   - плотность ключевых слов, процент ключевых фраз;
   - частотность слов;
   - количество стоп-слов;
   - объем текста: количество символов с пробелами и без пробелов;
   - количество слов: уникальных, значимых, всего;
   - водность, процент воды;
   - тошноту текста, классическую и академическую;
   - количество грамматических ошибок.
3. Постройте облако тегов.
4. Существует ли связь между содержанием и названием статьи?
5. Прокомментируйте полученные результаты, составьте отчет.

Критерии оценивания:
Максимальная оценка за все лабораторные работы – 63 балла.

Максимальная оценка по каждой работе - 7 баллов
5,9 – 7,0 балла выставляется, если обучающийся: выполнил работу в полном объеме с соблюдением необходимой последовательности; самостоятельно и рационально выбрал спецификации моделей; грамотно оформил представленный отчет;
4,7 – 5,8 балла выставляется, если обучающийся: выполнил работу в полном объеме с соблюдением необходимой последовательности; самостоятельно и рационально выбрал спецификации моделей; грамотно оформил представленный отчет; дана содержательная интерпретация полученных при решении задач результатов; материал изложен четко; допускаются отдельные логические и стилистические погрешности, уверенно исправленные после дополнительных вопросов;
3,5-4,6 балла выставляется, если обучающийся: выполнил работу в полном объеме с соблюдением необходимой последовательности; самостоятельно и рационально выбрал спецификации моделей; грамотно оформил представленный отчет; дана содержательная интерпретация полученных при решении задач результатов; допускаются отдельные логические и стилистические погрешности; обучающийся может испытывать некоторые затруднения в формулировке суждений;
0-3,5 балла выставляется, если работа не выполнена или выполнена не в полном объеме; обучающийся практически не владеет теоретическим материалом, допуская грубые ошибки, испытывает затруднения в формулировке собственных суждений, неспособен ответить на дополнительные вопросы.

Контрольные задания

1. Используйте массив данных *rivers*, в котором представлены длины (в милях) 141 основных рек в Северной Америке. Подгрузите этот массив с помощью команды *data(rivers)*. Чему равна средняя длина этих рек?

2. Используйте массив данных *rivers*, в котором представлены длины (в милях) 141 основных рек в Северной Америке. Подгрузите этот массив с помощью команды *data(rivers)*. Во сколько раз наибольшее значение длины реки превышает наименьшее значение длины реки?

3. Чему равна вероятность того, что случайная величина X, которая распределена N(78, 144), будет лежать в промежутке от 24 до 85?

4. Сколько миль налетали пассажиры в Америке, на примере встроенного массива данных *airmiles*. За какой год есть первые наблюдения?

5. Чему равняется *length(c(7,7,7))*

6. Что вернет выражение *sum(1:3<2)*?

7. American Community Survey предоставляет скачиваемые данные из различных обследований общества в Соединенных Штатах.
С помощью команды download.file () скачать данные из опроса о жилье в штате Айдахо в 2006 г. с сайта : https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv
Загрузите эти данные в R. Книга кодирования, описывающая имена переменных находится по адресу: https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf
Сколько категорий стоимостью $ 1 млн или больше?

8. Используйте данные из предидущего задания. Рассмотрим переменную FES. Какой из принципов "аккуратных данных" (tidy data) нарушаются в этой переменной?

9. Скачать Excel таблицу из данных Natural Gas Aquisition Program по адресу: 26https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx
(original data source: http://catalog.data.gov/dataset/natural-gas-acquisition-program) Прочитайте строки 18-23 и столбцы 7-15 в R и присвойте результат переменной с именем dat
Чему равно значение выражения sum(dat$Zip*dat$Ext,na.rm=T)?

10 Прочитайте XML данные о ресторанах г.Балтимора с сайта: https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml
Сколько ресторанов имеют zipcode 21231?

11. Скачайте данные опроса 2006 г. о жилье для штата Айдахо с помощью команды download.file () по адресу:
https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv
Используя команду fread() загрузите данные в R, назовите объект DT
Что из перечисленного ниже является самым быстрым способом для расчета средних значений переменной pwgtp15 для мужчин и женщин с использованием пакета data.table?

12. Рассмотрим набор переменных: Имя; Год рождения; Телефон; Кол-во сестер(братьев); Годовой доход. Например, Анна; 1975; 8929223; 0; 66000.
Какие переменные качественные, а какие - количественные?

13. В социальном обследовании, проводимом ежегодно в Соединенных Штатах, спрашивается, сколько друзей у людей (number of friends) и как они оценивают свой уровень счастья (very happy, pretty happy, not too happy).
Для того чтобы оценить связь между этими двумя переменными исследователь вычисляет среднее количество друзей для людей, которые классифицировали себя как очень счастлив, довольно счастлив, и не слишком счастлив.
Какие переменные независимые? зависимые ?

14. В исследовании, опубликованном в 2011 PNAS USA, 120 пожилых мужчин и женщин (средний возраст около 65 лет), которые добровольно согласились участвовать в этом исследовании, были случайным образом распределены в две группы.
В первой группе добровольцы ходили по дорожке в парке три раза в неделю; в другой - делали множество менее аэробных упражнений, в том числе йогу и тренировки с отягощением.
Через год сканирование мозга показало, что у «пешеходов» гиппокамп (часть мозга, отвечающая за формирование воспоминаний) увеличился в объеме в среднем примерно на 2%; в другой группе объем гиппокампа снизился на 1,4%.
Что из перечисленного ниже ложно?

15. В одном американском городе был проведен опрос о жилье, чтобы определить цену типичного дома в городе, в котором проживает в основном средний класс, но есть очень дорогой пригород. Средняя стоимость дома в этом городе примерно $ 650 000.
Верно ли, что большинство домов в этом городе стоят более $ 650 000?

Критерии оценивания:
Максимальный балл - 15

Каждое задание оценивается максимум в 1 балл. Критерии оценивания 1 задания:
0,84-1,0 балла выставляется, если задание выполнено полностью, в представленном решении обоснованно получены правильные ответы, проведен анализ, дана грамотная интерпретация полученных результатов, сделаны выводы.
0,67-0,83 балла выставляется, если задание выполнено полностью, но при анализе и интерпретации полученных результатов допущены незначительные ошибки, выводы – достаточно обоснованы, но неполны.
0,5-0,66 балла выставляется, если задание выполнено частично, анализ и интерпретация полученных результатов не вполне верны, выводы верны частично.
0-0,49 балла выставляется, если решение неверно или отсутствует.