федеральное государственное бюджетное образовательное учреждение высшего образования «Красноярский государственный медицинский университет имени профессора В.Ф. Войно-Ясенецкого» Министерства здравоохранения Российской Федерации

УТВЕРЖДАЮ
Проректор по учебной работе, д.м.н., доцент
И.А. Соловьева
«20» октября 2021 г.

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ
по дисциплине
Программное обеспечение статистической и интеллектуальной обработки данных
для подготовки обучающихся по направлению подготовки 38.04.02 Менеджмент, направленность (профиль) «Управление в здравоохранении на основе интеллектуального анализа данных»

Красноярск
2021

Практическое занятие №1

Тема: Визуальный анализ данных (В интерактивной форме).

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.1.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для внеаудиторной работы по теме

	ВСЕГО	90	

Аннотация (краткое содержание темы):

Функция ЧАСТОТА в Microsoft Excel

Описание. Вычисляет частоту появления значений в интервале значений и возвращает массив чисел. Функцией ЧАСТОТА можно воспользоваться, например, для подсчета количества результатов тестирования, попадающих в интервалы результатов. Поскольку данная функция возвращает массив, ее необходимо вводить как формулу массива.

Синтаксис.

ЧАСТОТА(массив_данных; массив_интервалов)

Массив_данных — обязательный аргумент. Массив или ссылка на множество значений, для которых вычисляются частоты. Если аргумент "массив_данных" не содержит значений, функция ЧАСТОТА возвращает массив нулей.

Массив_интервалов — обязательный аргумент. Массив или ссылка на множество интервалов, в которые группируются значения аргумента "массив_данных". Если аргумент "массив_интервалов" не содержит значений, функция ЧАСТОТА возвращает количество элементов в аргументе "массив_данных".

Примечания

Функция ЧАСТОТА вводится как формула массива после выделения диапазона смежных ячеек, в которые требуется вернуть полученный массив распределения.

Функция ЧАСТОТА пропускает пустые ячейки и текст.

Формулы, возвращающие массивы, необходимо вводить как формулы массива.

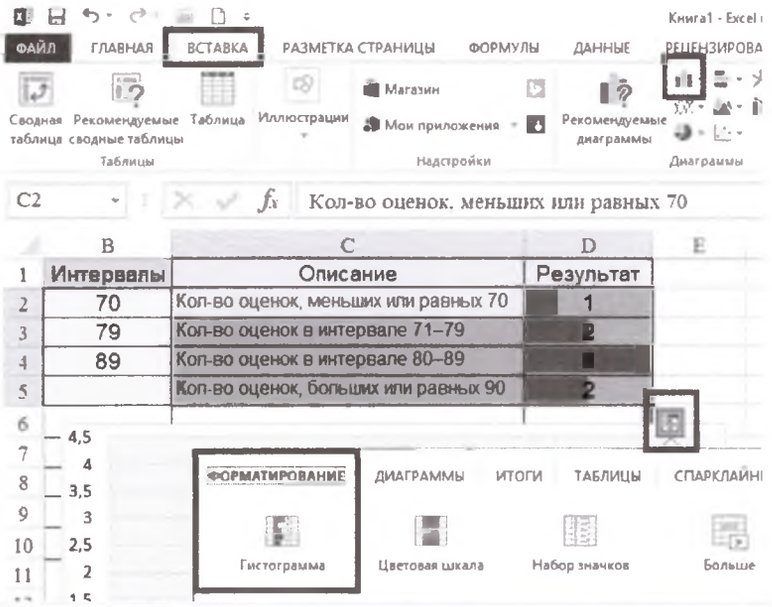
Количество элементов в возвращаемом массиве на единицу больше числа элементов в массиве "массив_интервалов". Дополнительный элемент в возвращаемом массиве содержит количество значений, превышающих верхнюю границу интервала, содержащего наибольшие значения. Например, при подсчете трех диапазонов значений (интервалов), введенных в три ячейки, убедитесь в том, что функция ЧАСТОТА возвращает значения в четырех ячейках. Дополнительная ячейка возвращает число значений в аргументе "массив_данных", превышающих значение верхней границы третьего интервала.

Пример

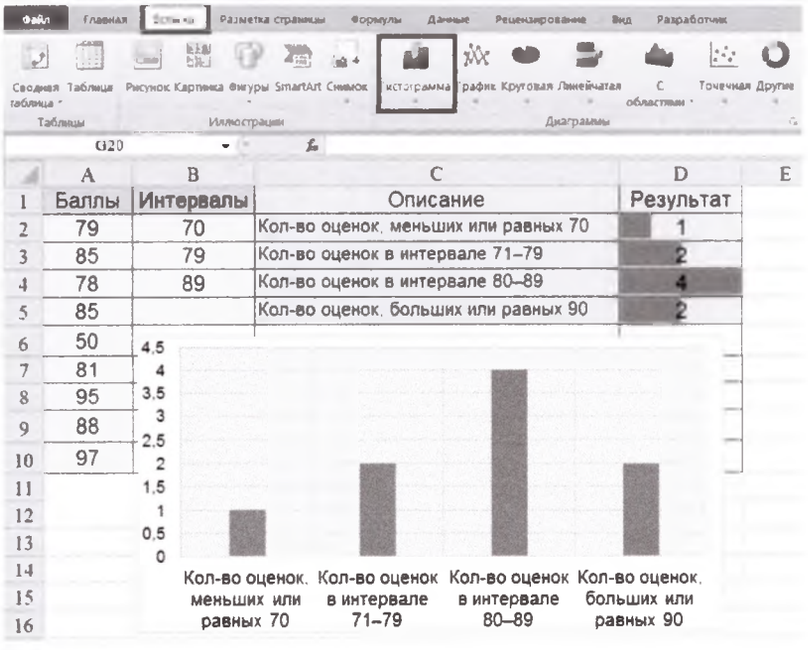
	A	B	C	D
1	Баллы	Интервалы	Описание	Результат
2	79	70	Кол-во оценок, меньших или равных 70	1
3	85	79	Кол-во оценок в интервале 71–79	2
4	78	89	Кол-во оценок в интервале 80–89	4
5	85		Кол-во оценок, больших или равных 90	2
6	50			
7	81		Формула	
8	95		=ЧАСТОТА(A2:A10;B2:B4)	
9	88			
10	97			

Примечание. Формула в ячейке C8 является формулой массива. Чтобы эта функция возвращала значения в ячейки D2, D3, D4 и D5, выделите диапазон с этими ячейками, введите формулу, как в C8, нажмите клавиши CTRL+SHIFT+Enter. Иначе будет возвращено только значение в ячейке D2.

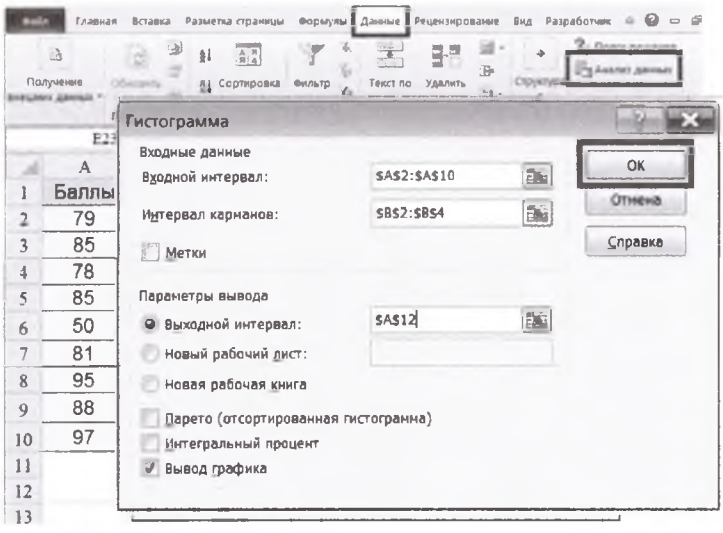
интерфейс и форматирование гистограмм MS Excel 2010



интерфейс MS Excel 2007

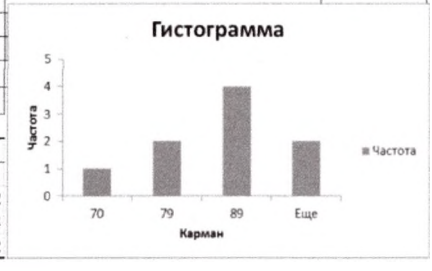


пакет Анализ данных



пакет Анализ данных

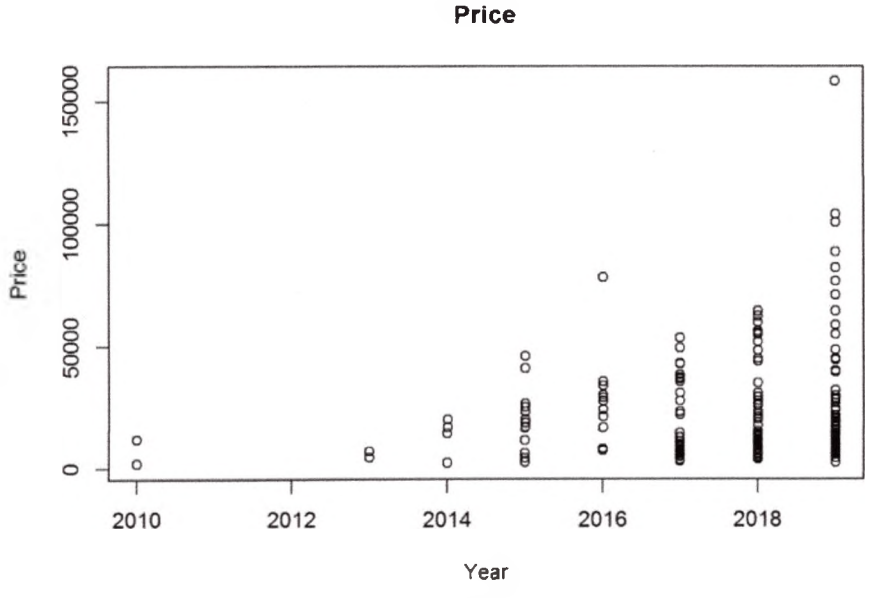
	A	B	C	D
1	Баллы	Интервалы	Описание	Результат
2	79	70	Кол-во оценок, меньших или равных 70	1
3	85	79	Кол-во оценок в интервале 71–79	2
4	78	89	Кол-во оценок в интервале 80–89	4
5	85		Кол-во оценок, больших или равных 90	2
6	50			
7	81			
8	95			
9	88			
10	97			
11				
12	*Карман*	*Частота*		
13	70	1		
14	79	2		
15	89	4		
16	Еще	2		



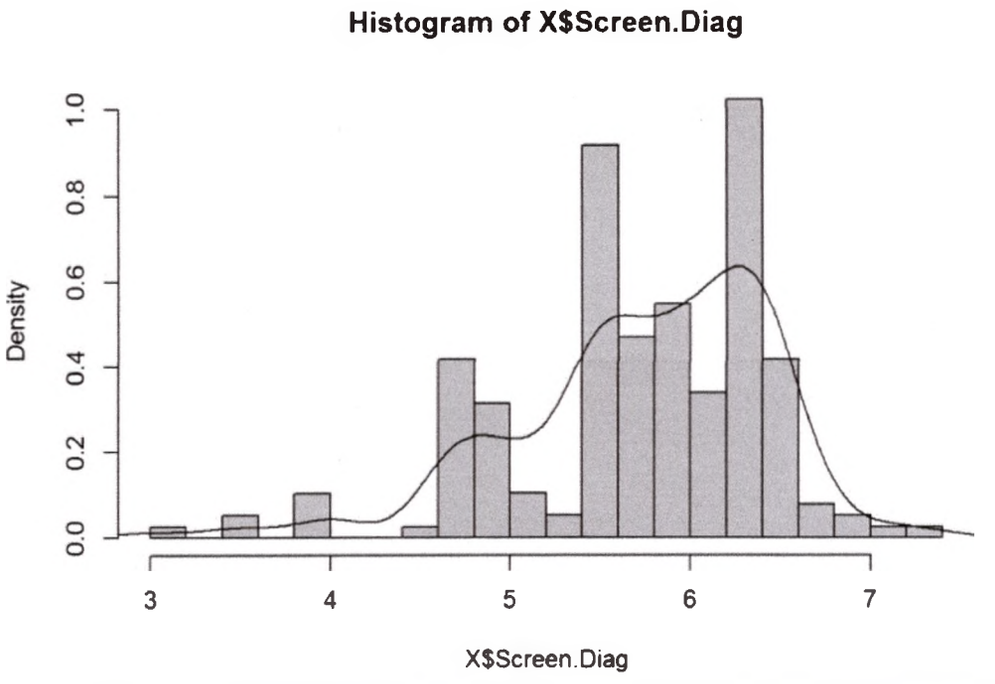
Базовая графика R (R-studio)

Кейс1 (импортируем данные из файла Smart.csv, датасет доступен на СДО в одноименном курсе):

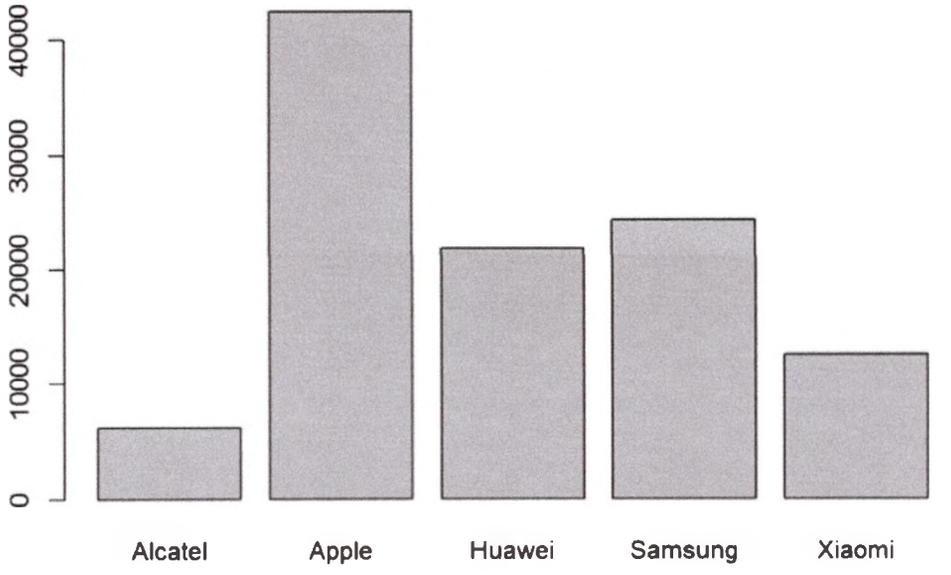
```
library(csv)
X=read.csv("D:/R/lesson1/Smarts.csv",sep=",",header=TRUE)

par(mfrow=c(1,1))
plot( Price.2019 ~ Year, data = X, main = "Price", xlab = "Year", ylab="Price")
```



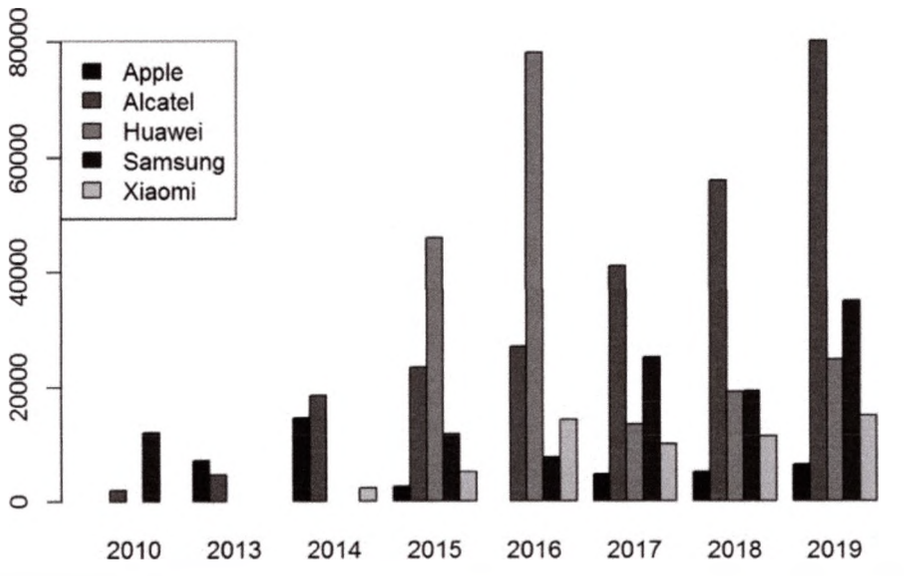
```
hist(X$Screen.Diag, col="grey",nclass=20, probability = TRUE)
lines( density(X$Screen.Diag) )
```



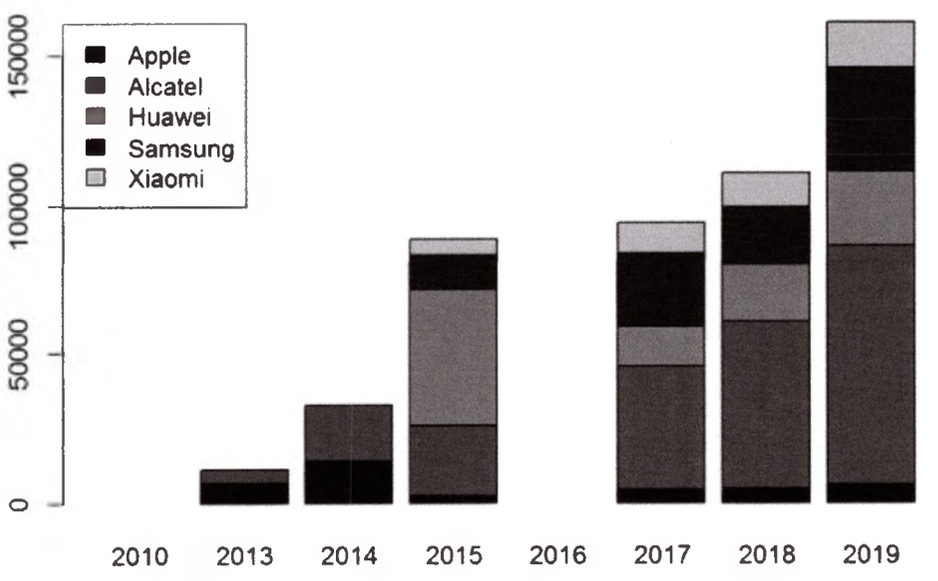

```
mean.prices=tapply(X$Price.2019, X$Company, mean)
barplot(mean.prices)
```



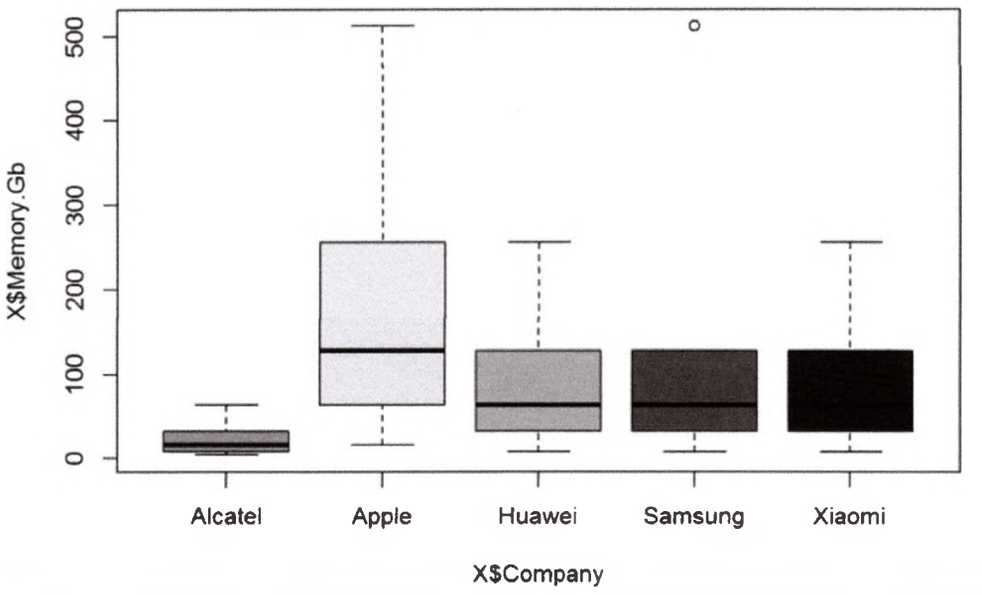
```
mean.prices2=tapply(X$Price.2019, list(X$Company,X$Year), mean)
barplot(mean.prices2, beside=TRUE, col=1:5)
legend("topleft",legend=c("Apple","Alcatel","Huawei","Samsung","Xiaomi"), fill=1:5)
```



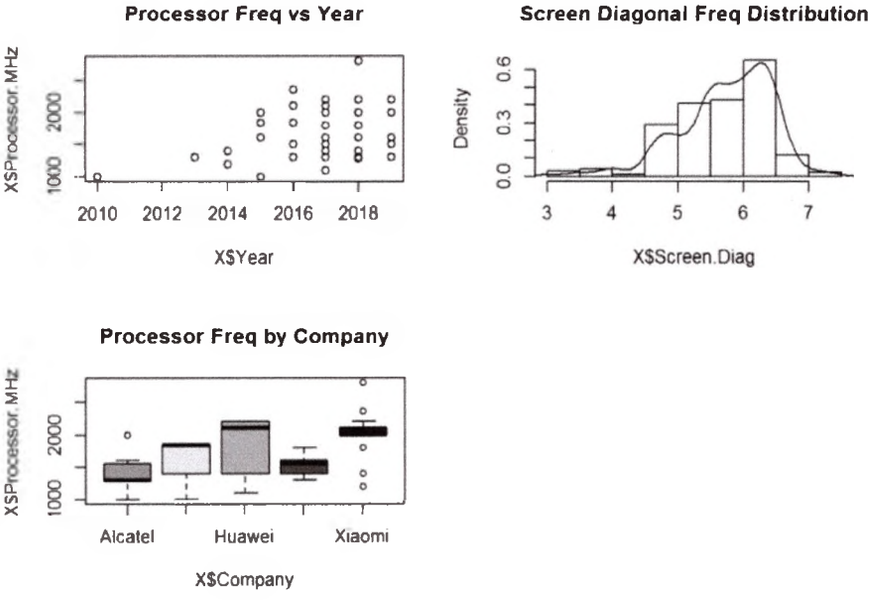

```
barplot(mean.prices2, beside=FALSE, col=1:5)
legend("topleft",legend=c("Apple","Alcatel","Huawei","Samsung","Xiaomi"), fill=1:5)
```



```
boxplot(X$Memory.Gb ~ X$Company, horizontal = FALSE, col=c("green","yellow","orange","red","blue"))
```




```
par(mfrow=c(2,2))
plot(X$Processor.MHz ~ X$Year, main="Processor Freq vs Year")
hist(X$Screen.Diag, probability=TRUE, main="Screen Diagonal Freq Distribution")
lines(density(X$Screen.Diag))
boxplot(X$Processor.MHz ~ X$Company, main="Processor Freq by Company",col=c("green","yellow","orange","red","blue"))
```



Примерная тематика НИРС по теме

1. Применение статистики в здравоохранении
2. Визуализация как средство анализа информации

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В. Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)

6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)
11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)
14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №2

Тема: Моделирование данных с заданным законом распределения (В интерактивной форме).

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.3.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

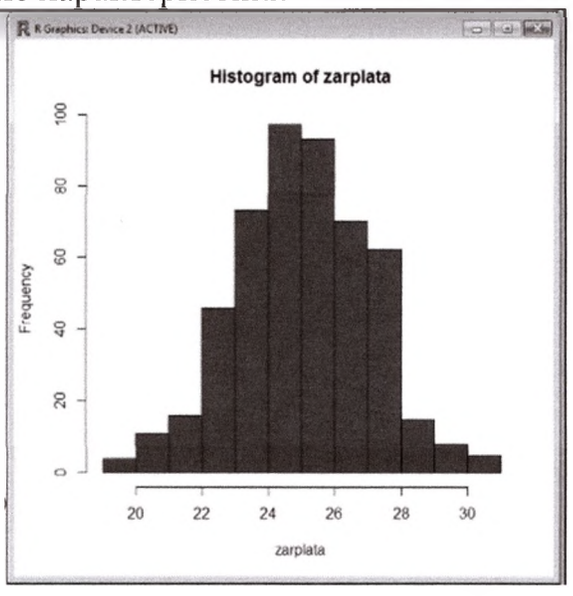
п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для

			внеаудиторной работы по теме
	ВСЕГО	90	

Аннотация (краткое содержание темы):

Смоделируем выборку из нормального распределения с заданными параметрами (значения зарплаты за месяц 500 сотрудников предприятия), построим гистограмму, найдём выборочные характеристики

```
> # Смоделируем данные о зарплате 500 работников компании
> # Распределение будем считать нормальным с параметрами :
> # математическое ожидание = 25 (тыс. руб.)
> # среднее квадратическое отклонение = 2 (тыс. руб.)
>
> zarplata = rnorm(500,25,2)
>
> # Построим гистограмму
> # col - цвет, breaks - количество столбцов,
> # freq - "флажок" для надписей на оси OY
>
> hist(zarplata,col="red",breaks=15,freq=TRUE)
>
> # Найдём среднюю зарплату по фирме (тыс. руб.)
> sred=mean(zarplata)
> sred
[1] 25.03542
>
> # Найдём выборочную дисперсию зарплаты (тыс. руб.^2)
> disp=var(zarplata)
> disp
[1] 4.114386
>
> # Найдём выборочное среднее квадратическое отклонение (тыс. руб.)
> s=sqrt(disp)
> s
```



Примерная тематика НИРС по теме

1. Возможности анализа данных медико-биологических экспериментов в различных статистических пакетах

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В. Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.

5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)
6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
10. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
11. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)
12. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
13. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
14. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №3

Тема: Первичный анализ данных в различных пакетах. Проверка статистических гипотез.

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.1.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для

			внеаудиторной работы по теме
	ВСЕГО	90	

Аннотация (краткое содержание темы):

Описательная статистика, представление данных

1. Определите **тип признака**, который вы хотите описать. Если признак - ***количественный***, переходим к **п.2**. Если ***качественный*** - к **п.5**.
2. Оцените **нормальность распределения** признака с помощью **критерия Колмогорова-Смирнова** (при числе исследуемых n>50) или **критерия Шапиро-Уилка** (при n<50). Если распределение нормальное - переходим к **п.3**, если отличается от нормального - переходим к **п.4**.
3. При описании нормально распределенного количественного признака укажите **среднее значение** (M), **стандартное отклонение** (σ) или **стандартную ошибку** (m), **95% доверительный интервал** (95% ДИ).
2. **Например:** *Систолическое артериальное давление у пациентов исследуемой группы составляло от 90 до 150 мм рт.ст., среднее значение показателя - 118±2,5 мм рт.ст. (95% ДИ 113,2 - 123,1 мм рт.ст.).*
3. При описании количественного признака, распределение которого отличается от нормального, укажите медиану, значения нижнего и верхнего квартилей (или 25% и 75% перцентилей).
4. **Например:** *Медиана частоты сердечных сокращений у пациентов основной группы составила 84 удара в минуту с интерквартильным размахом от 76,5 до 91.*
5. При описании качественного признака для каждого его значения укажите абсолютную величину, а также процентную долю в структуре всей совокупности.
6. **Например:** *В структуре исследуемой совокупности артериальная гипертония отмечалась в 24 случаях, или в 12%.*

Первичный анализ данных в R

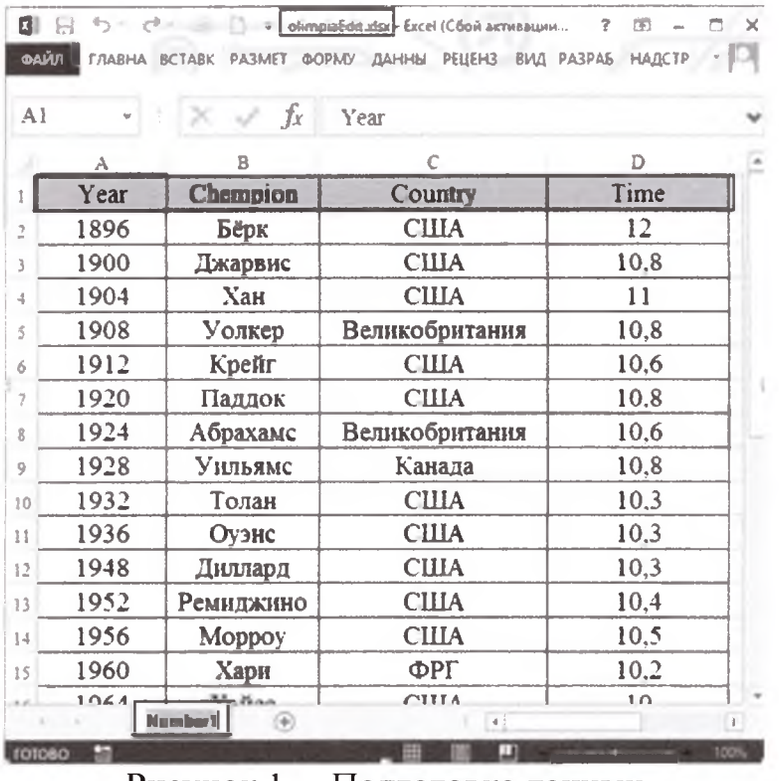
Установку последней версии R рассмотрели в лекции. Рекомендую дополнительно установить все пакеты, начинающиеся с Rcmdr.

Теперь еще раз пробежимся по первичному анализу и формированию отчета. Markdown изначально был предназначен для вывода HTML.

Рабочий файл: olimpia.xlsx (заголовки данных на кириллице!).

Сделаем копию этого файла и сохраним с именем olimpiaEdit.xlsx.

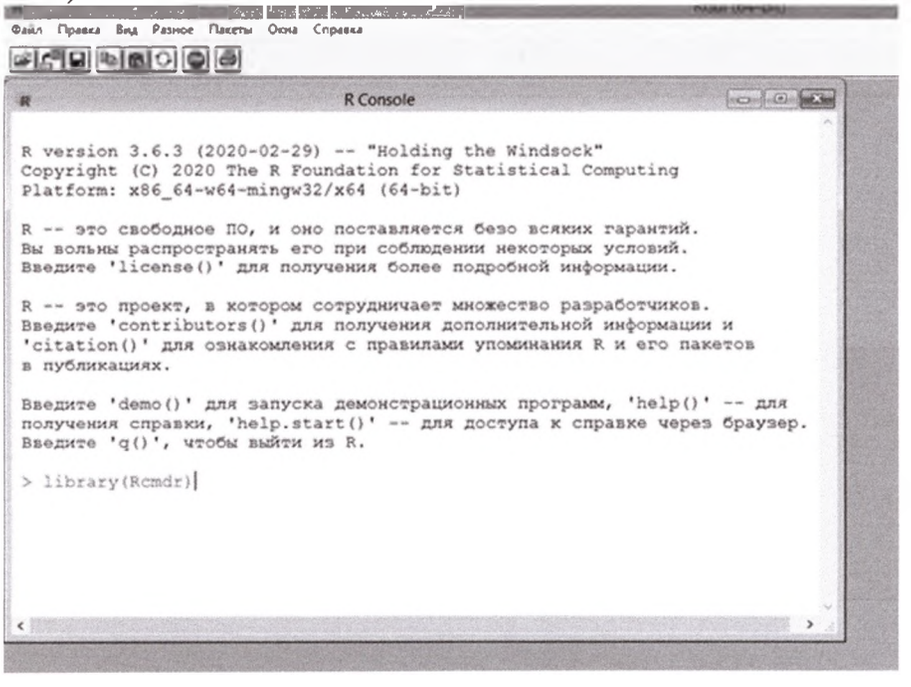
Переименуем заголовки как на рисунке 1, а также название Лист1 в Number1.



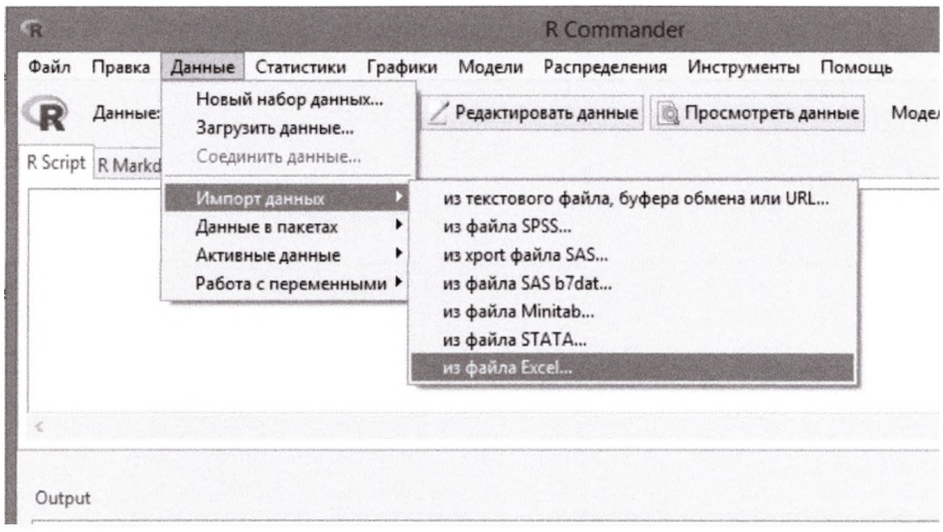
	A	B	C	D
1	Year	Champion	Country	Time
2	1896	Бёрк	США	12
3	1900	Джарвис	США	10,8
4	1904	Хан	США	11
5	1908	Уолкер	Великобритания	10,8
6	1912	Крейг	США	10,6
7	1920	Паддок	США	10,8
8	1924	Абрахамс	Великобритания	10,6
9	1928	Уильямс	Канада	10,8
10	1932	Толан	США	10,3
11	1936	Оуэнс	США	10,3
12	1948	Диллард	США	10,3
13	1952	Ремиджино	США	10,4
14	1956	Морроу	США	10,5
15	1960	Хари	ФРГ	10,2

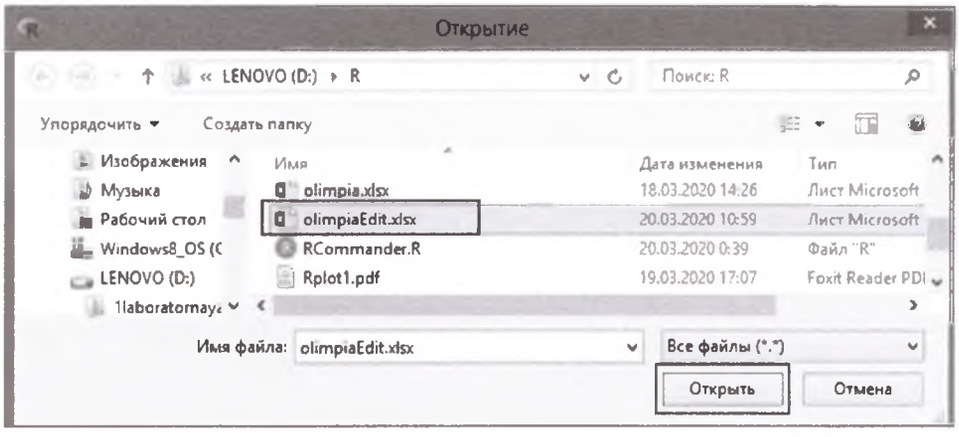
Рисунок 1. – Подготовка данных

Загружаем R – Rgui и в консоли набираем.
>library(Rcmdr)

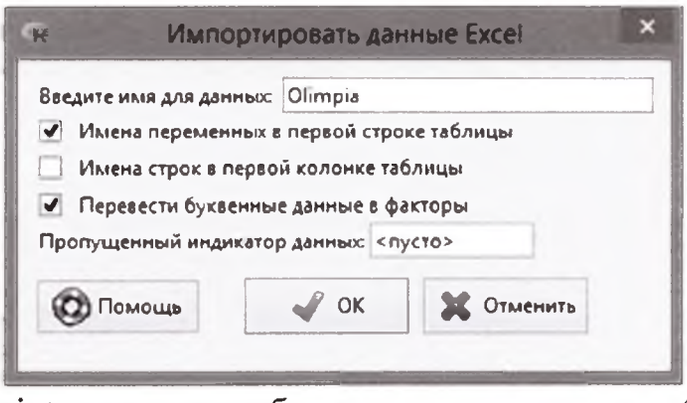


Далее, в окне R Commander импортируем данные из Excel:





Дадим имя для рабочей базы данных в R – Olimpia:

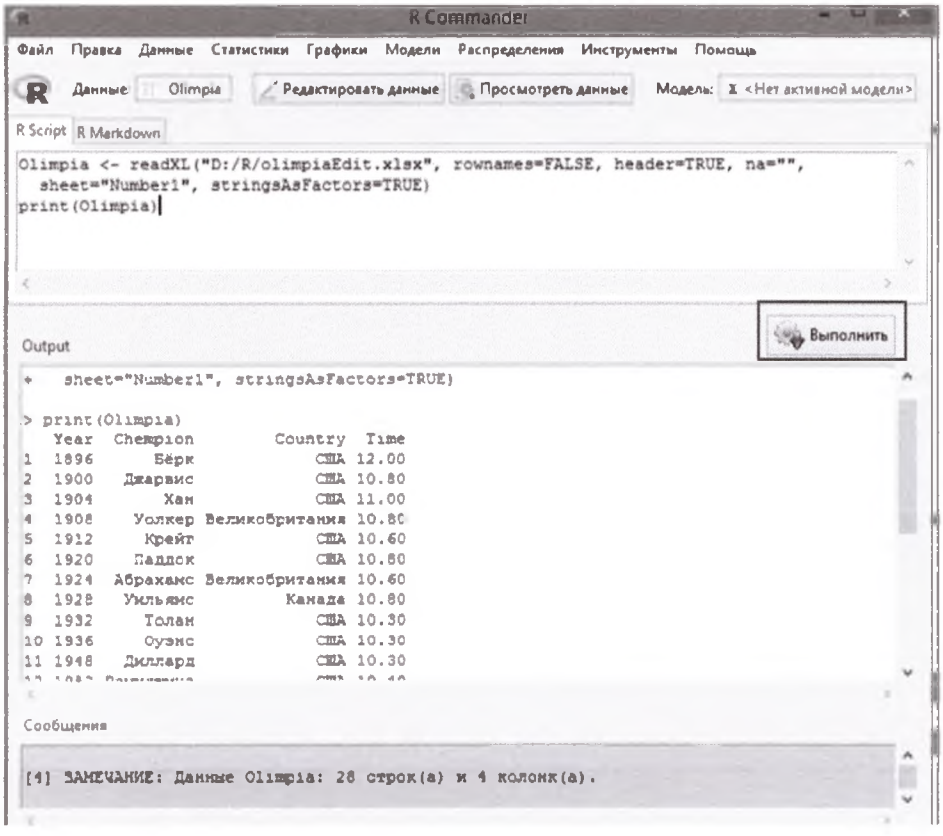


В окне R Script появится сообщение с указанием пути (в моем случае это D:/R/olimpiaEdit.xlsx), название листа – sheet="Number1":
Olimpia

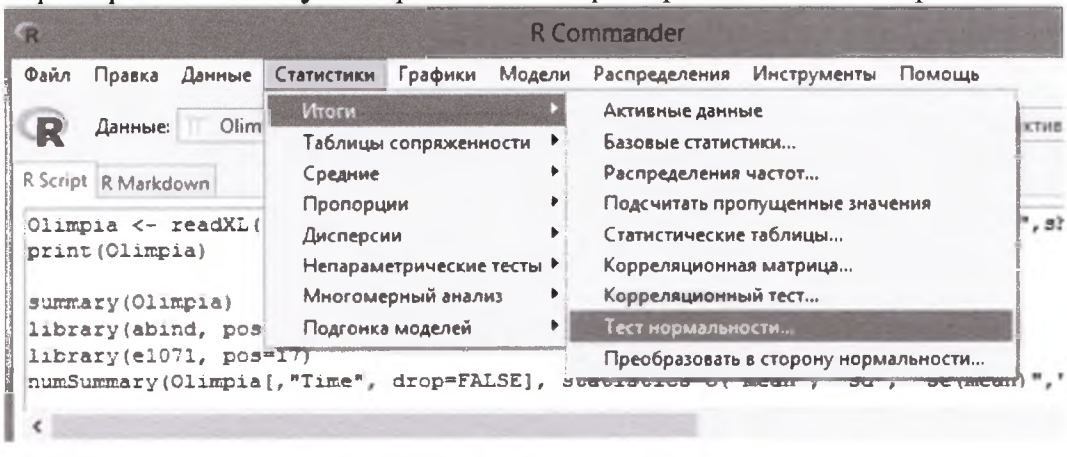
Распечатаем данные для будущего отчета. Для этого набираем команду в окне R Script:

print(Olimpia)

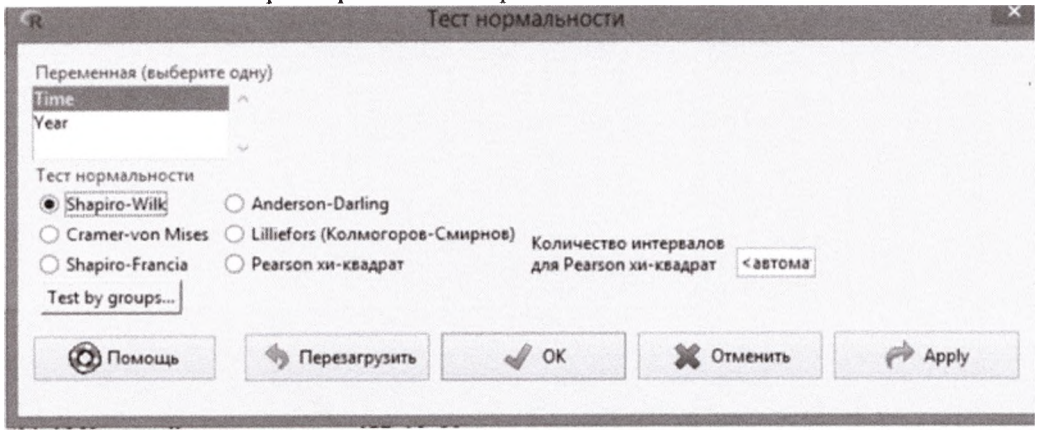
И нажимаем на Выполнить.



Проверим гипотезу о нормальности распределения выборки



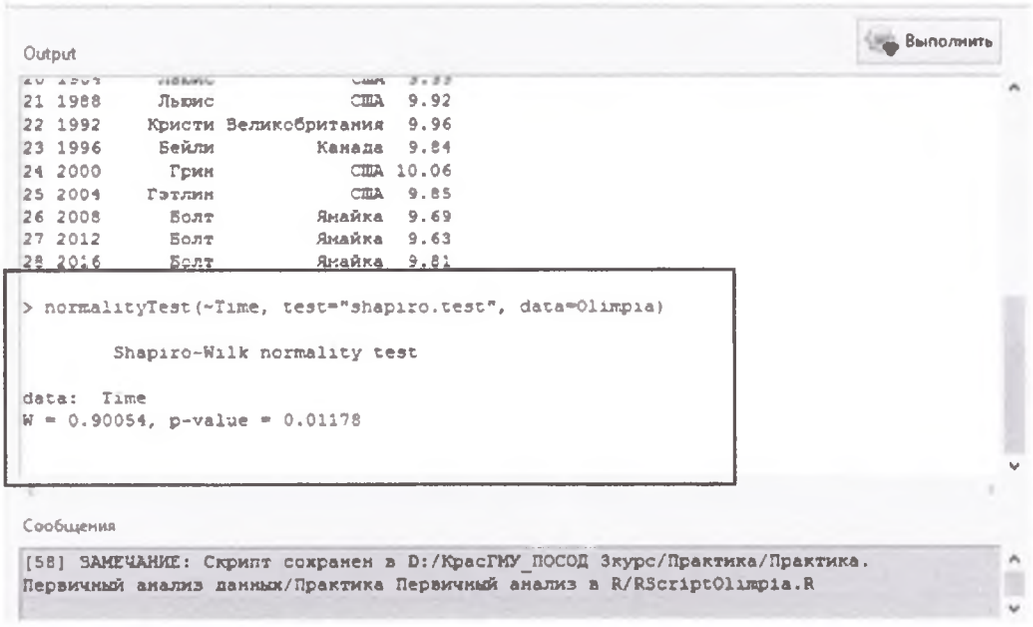
Поскольку выборка имеет малый объем (n = 28), в этой ситуации может помочь только критерий Шапиро–Уилка.



В окне R Script выражается следующей командой:

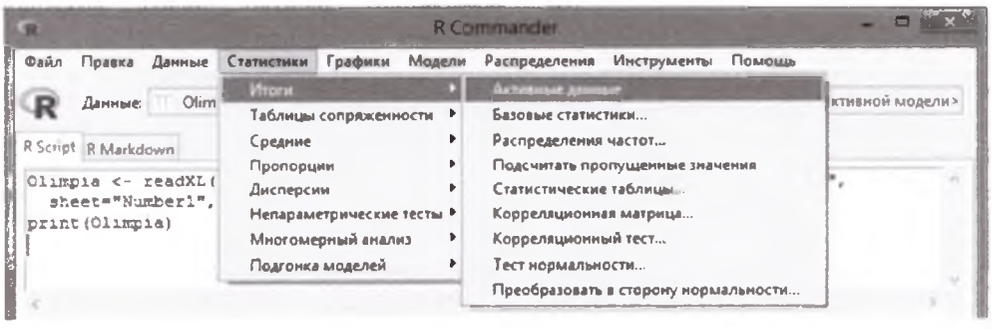
normalityTest(~Time, test="shapiro.test", data=Olimpia)

В окне Output (вывод результата) получаем

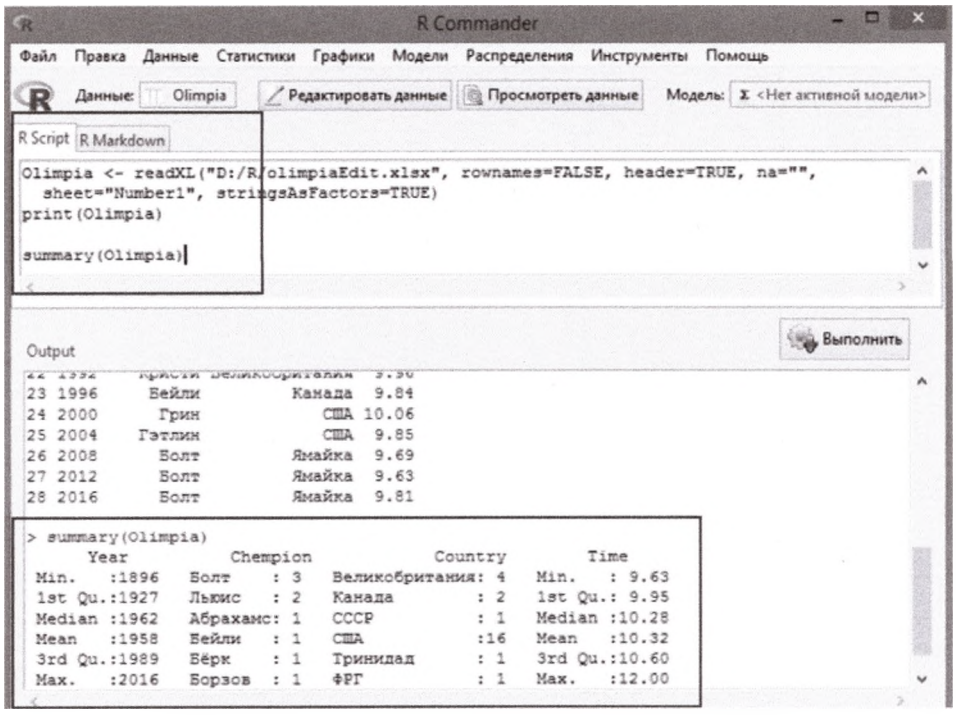


Получаем, что значение p=0.01178, что меньше 0.05, следовательно, гипотезу о нормальности распределения отклоняем.

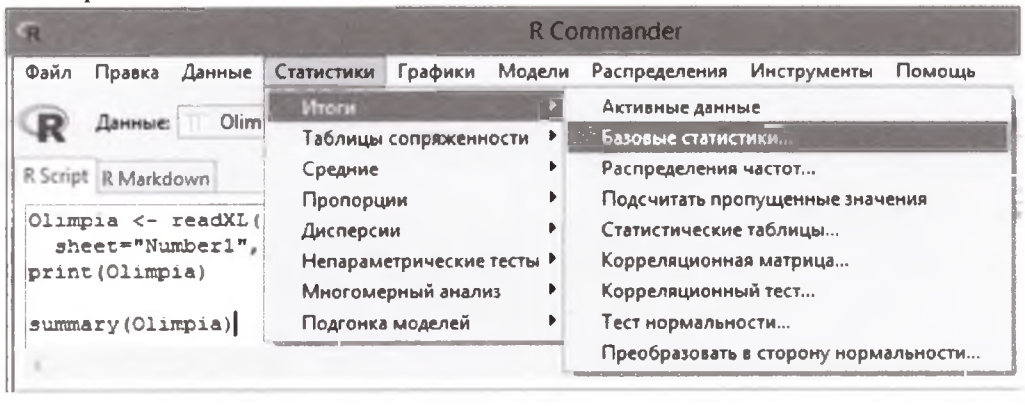
Проводим расчет базовых статистик. Сначала выберем Активные данные:

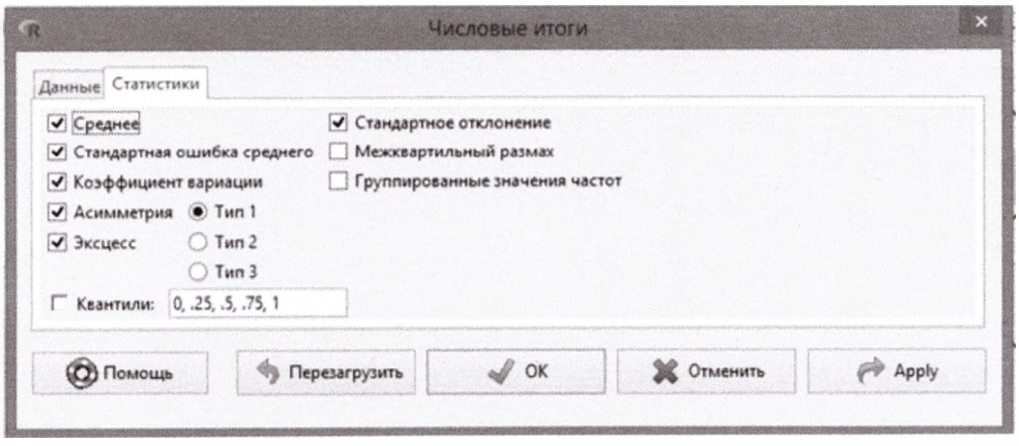


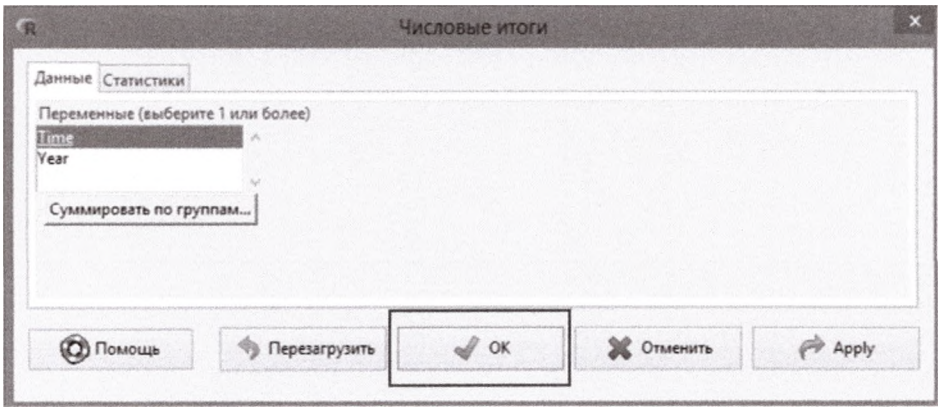
Результат формируется из команды:
summary(Olimpia)



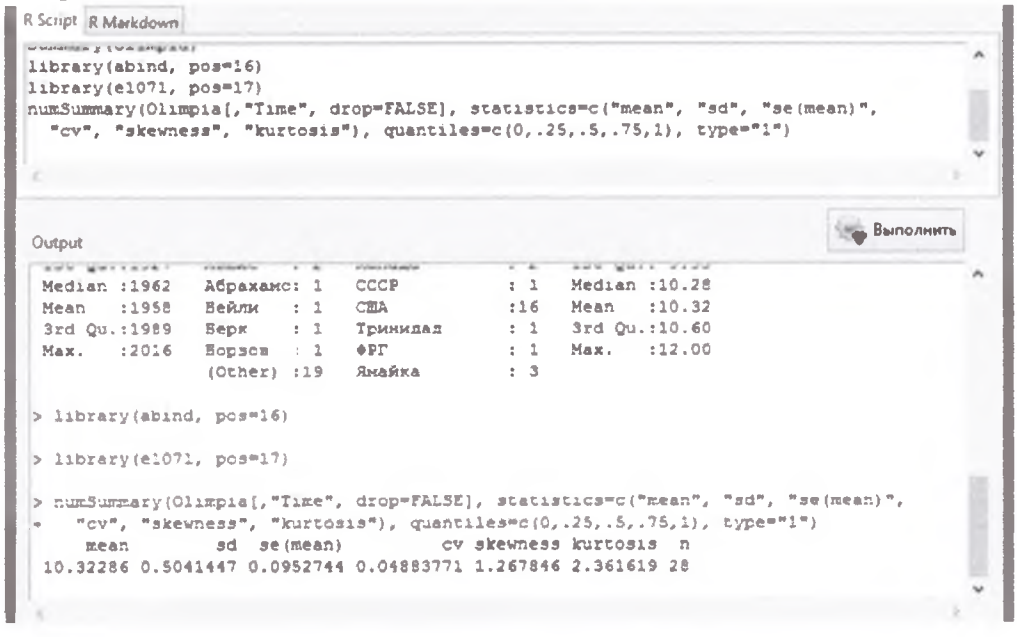
Далее, выбираем Базовые статистики:





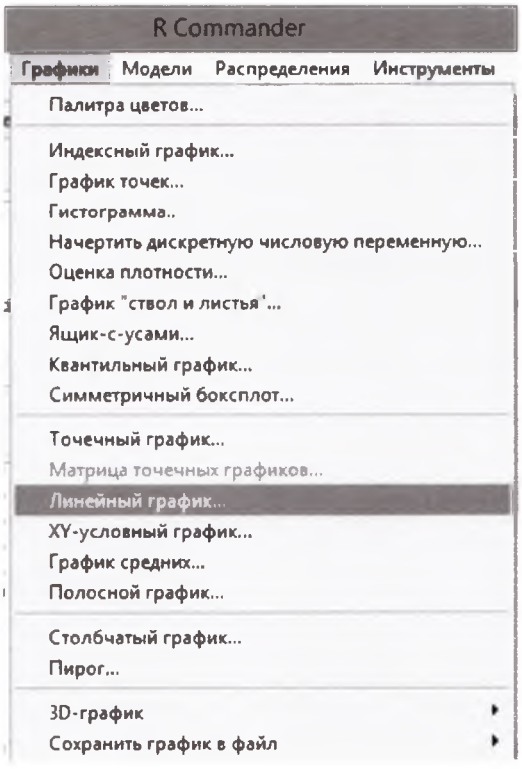


В окне Output (вывод результата) получаем:

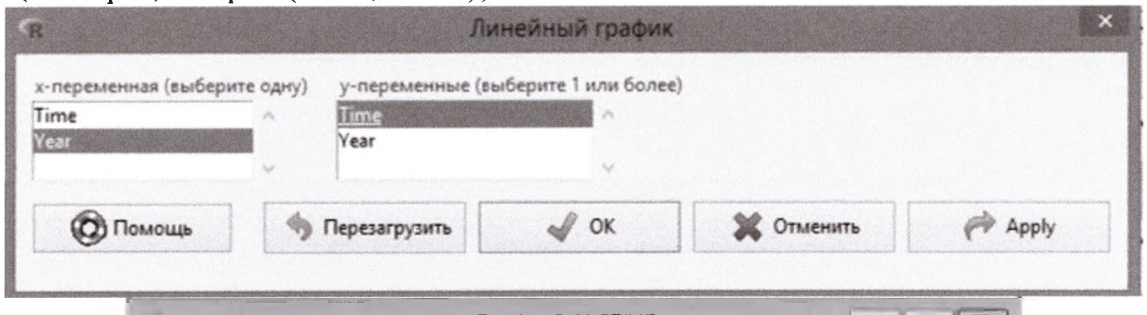


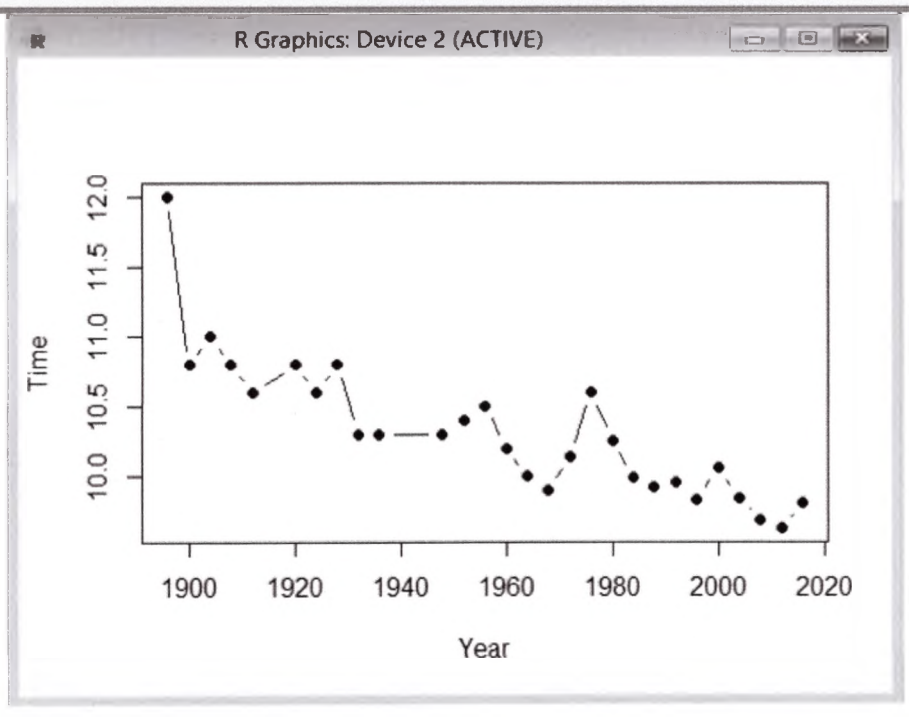
Строим графики.

1. Линейный график. По оси абсцисс выбираем года (Year), по оси ординат – время(Time).



В окне R Script построение графика выражено командой:
with(Olimpia, lineplot(Year, Time))

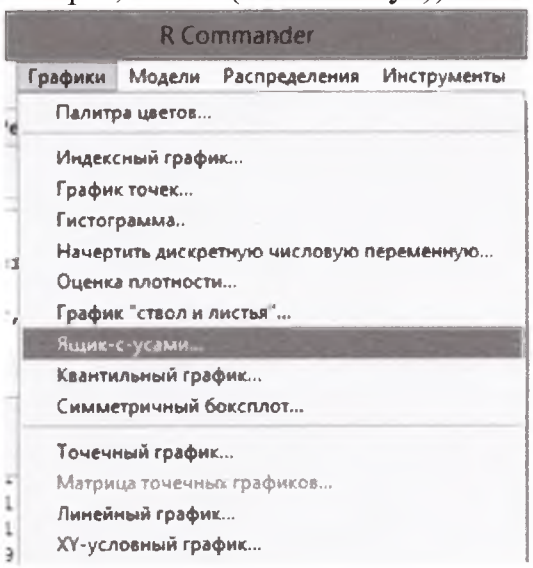


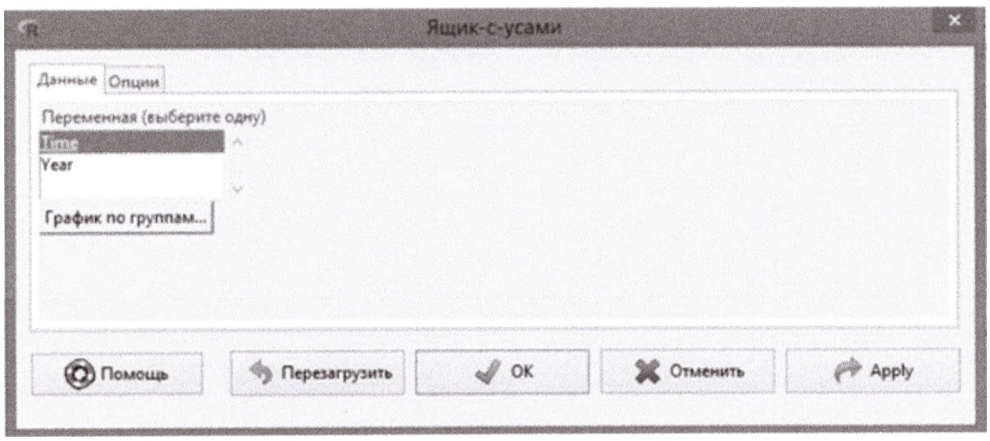


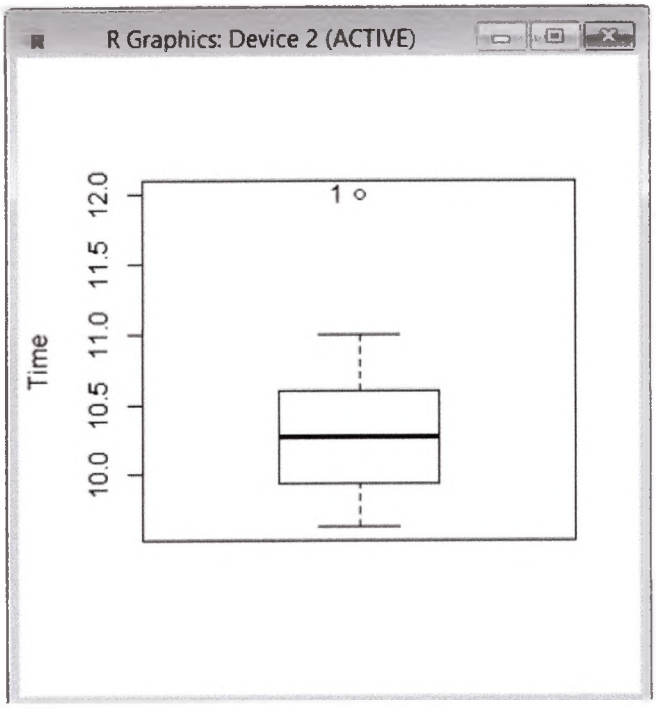
2. График Ящик-с-усами

В окне R Script построение ящика выражено командой:

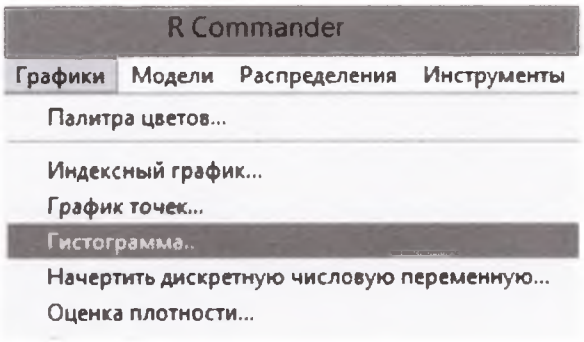
Boxplot(~ Time, data=Olimpia, id=list(method="y"))

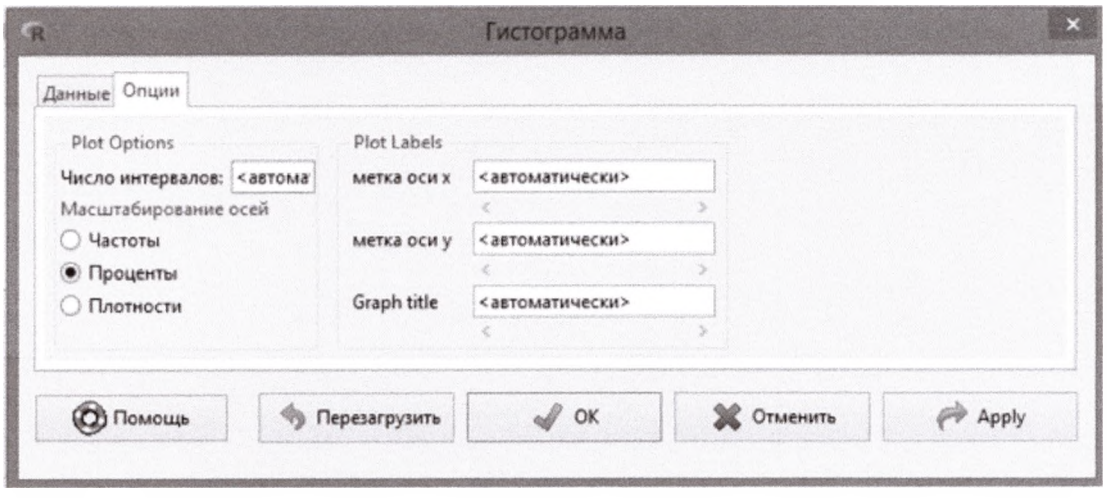


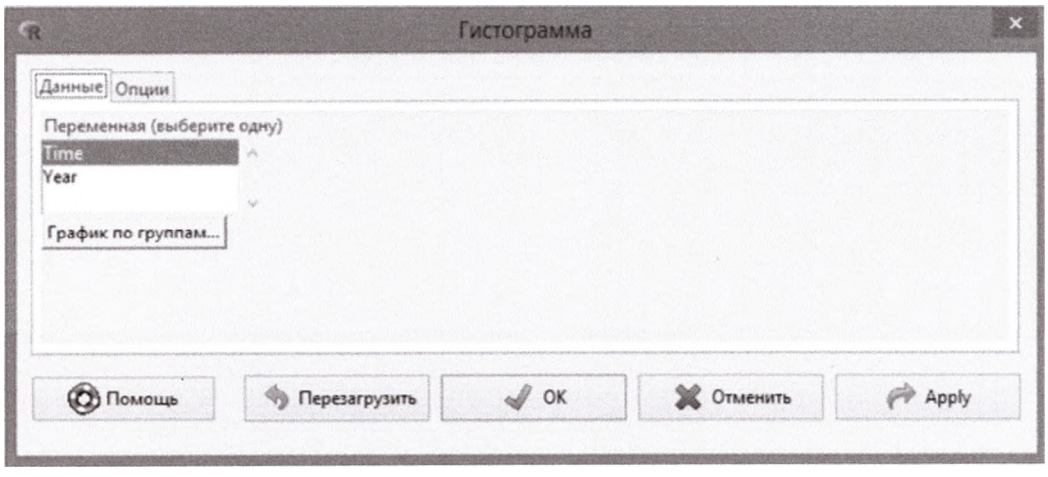




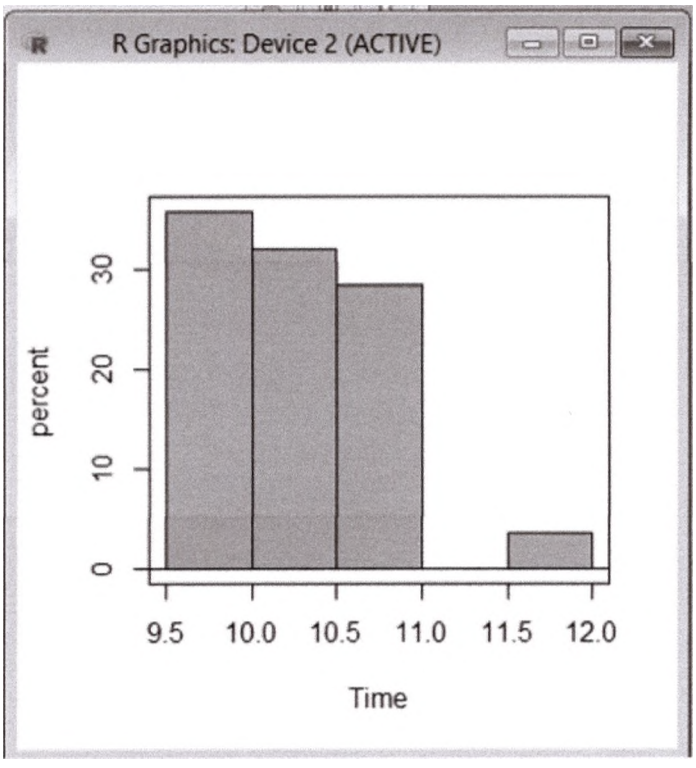
3. Гистограмма





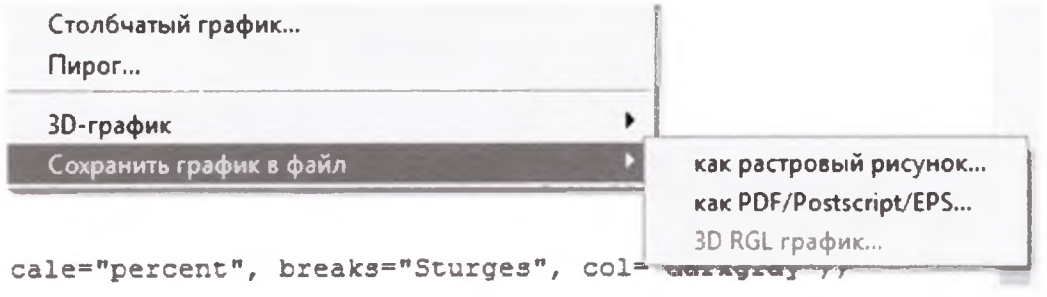


В окне R Script построение ящика выражено командой:
with(Olimpia, Hist(Time, scale="percent", breaks="Sturges", col="darkgray"))



Сохранить программный код – Сохранить скрипт как (например, RScriptOlimpia.R) В Сообщении появляется Замечание с указанием пути.

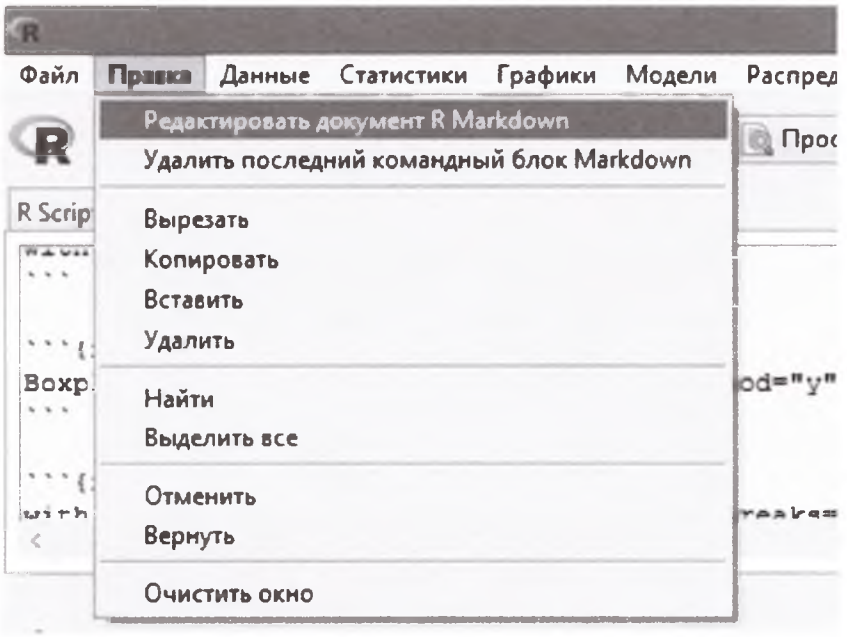
Далее, переходим в окно R Markdown

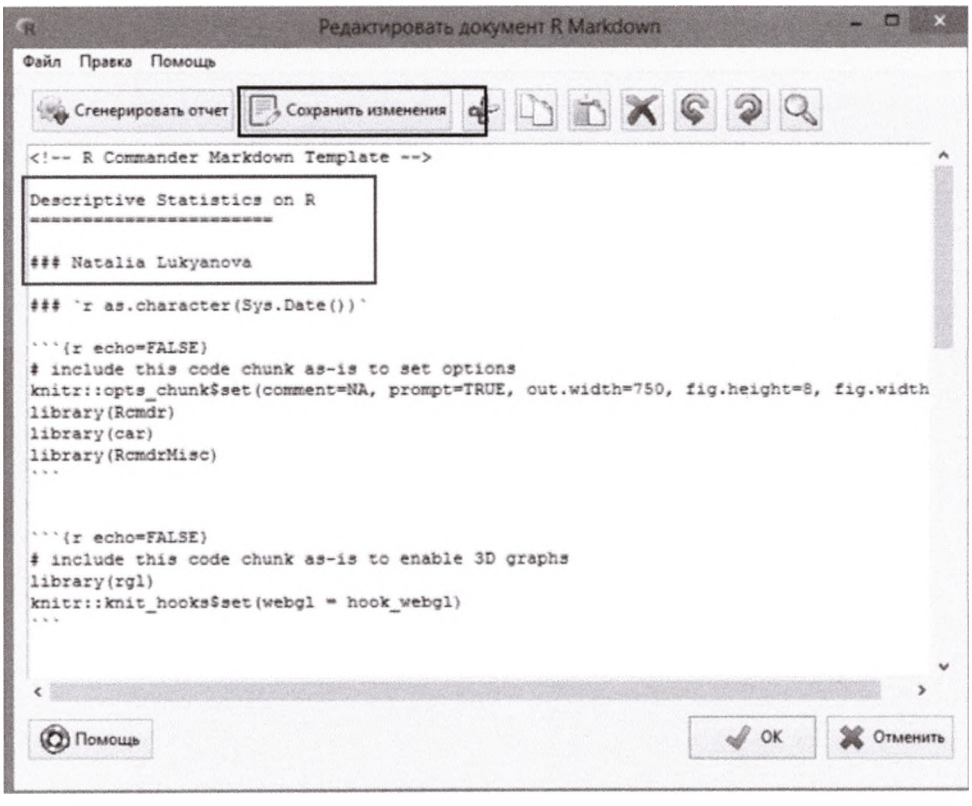


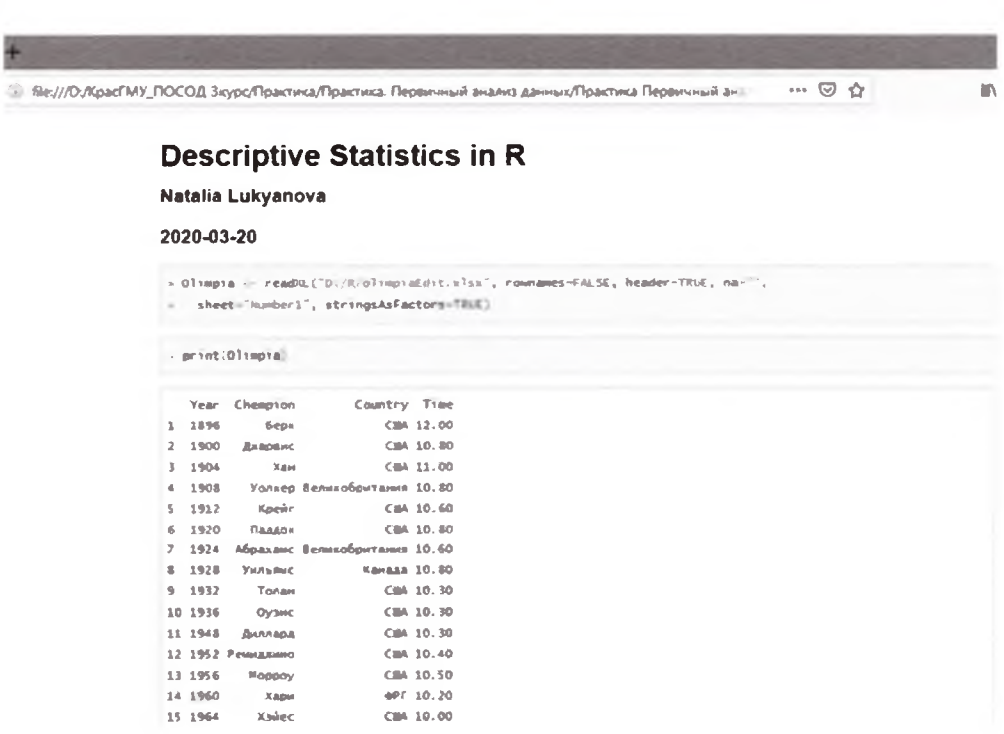
Просматриваем последовательность действий. Можно удалить какое-то действие или вставить. Добавить имя автора (на латинице). Дать название отчету (на латинице) - Descriptive Statistics in R (Описательная статистика в R).

После сохранить изменения (будет сформирован файл RMarkdownOlimpia.Rmd).

Сгенерировать отчет. В папке появится файл RMarkdownOlimpia.html. В Сообщении появляется Замечание с указанием пути.


Файл
Правка
Данные
Статистики
Графики
Модели
Редактировать документ R Markdown
Удалить последний командный блок Markdown
Вырезать
Копировать
Вставить
Удалить
Найти
Выделить все
Отменить
Вернуть
Очистить окно


Редактировать документ R Markdown
Файл
Правка
Помощь
Сгенерировать отчет
Сохранить изменения
<!-- R Commander Markdown Template -->
Descriptive Statistics on R
Natalia Lukyanova
`r as.character(Sys.Date())`
```{r echo=FALSE}
# include this code chunk as-is to set options
knitr::opts_chunk$set(comment=NA, prompt=TRUE, out.width=750, fig.height=8, fig.width
library(Rcmdr)
library(car)
library(RcmdrMisc)
```
```{r echo=FALSE}
# include this code chunk as-is to enable 3D graphs
library(rgl)
knitr::knit_hooks$set(webgl = hook_webgl)
```
Помощь
OK
Отменить



Descriptive Statistics in R

Natalia Lukyanova

2020-03-20

```
> Olimpia <- readXL("D:/R/OlimpiaEdit.xlsx", rownames=FALSE, header=TRUE, na="",
+   sheet="Number1", stringsAsFactors=TRUE)
```

```
> print(Olimpia)
```

```
   Year  Chempion        Country  Time
1  1896      Бёрк            США 12.00
2  1900   Дэвриес            США 10.80
3  1904       Хан            США 11.00
4  1908    Уолкер Великобритания 10.80
5  1912     Крейг            США 10.60
6  1920    Паддок            США 10.80
7  1924 Абрахамс Великобритания 10.60
8  1928    Уильямс         Канада 10.80
9  1932     Толан            США 10.30
10 1936     Оуэнс            США 10.30
11 1948   Диллард            США 10.30
12 1952 Ремиджино            США 10.40
13 1956    Морроу            США 10.50
14 1960      Хари            ФРГ 10.20
15 1964     Хэйес            США 10.00
```

Итак, полный R Scrupt:

```
Olimpia
print(Olimpia)
normalityTest(~Time, test="shapiro.test", data=Olimpia)
summary(Olimpia)
library(abind, pos=16)
library(e1071, pos=17)
numSummary(Olimpia[,"Time", drop=FALSE], statistics=c("mean", "sd",
"se(mean)","cv", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1),
type="1")
with(Olimpia, lineplot(Year, Time))
Boxplot( ~ Time, data=Olimpia, id=list(method="y"))
with(Olimpia, Hist(Time, scale="percent", breaks="Sturges",
col="darkgray"))
```

Примерная тематика НИРС по теме

1. Возможности анализа данных медико-биологических экспериментов в различных статистических пакетах

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.

3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В. Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)
6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)

11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)
14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №4

Тема: Методы оценки связи в различных пакетах (В интерактивной форме).

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.1, ПК-4.3.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для внеаудиторной работы по теме

	ВСЕГО	90	

Аннотация (краткое содержание темы):

Пример. Виды корреляции

Предположить зависимость одной величины от другой в самом простом случае можно определив их корреляцию. Переменные X и Y имеют положительную корреляцию, если большим значениям X соответствуют большие значения Y. Если большим X соответствуют меньшие Y, то корреляция отрицательная. Если ни одной из этих зависимостей установить не удаётся, то корреляция будет нулевой.

```
X=read.csv("D:/R/Smarts.csv",sep=",",header=TRUE)
library(rpart) par( mfrow = c(1,3))
plot(Mass.Gr ~ Battery.mAh,data=X, main="Положительная корреляция")
lin.regr1=lm(Mass.Gr ~ Battery.mAh, data=X)
lin.regr1
```

```
##
## Call:
## lm(formula = Mass.Gr ~ Battery.mAh, data = X)
##
## Coefficients:
## (Intercept) Battery.mAh
## 86.7690 0.0232
```

```
abline(lin.regr1)

plot(Main.Camera.Mpix ~ Main.Camera.Diafr,data=X, main="Отрицательная
корреляция")
lin.regr2=lm(Main.Camera.Mpix ~ Main.Camera.Diafr, data=X)
lin.regr2
```

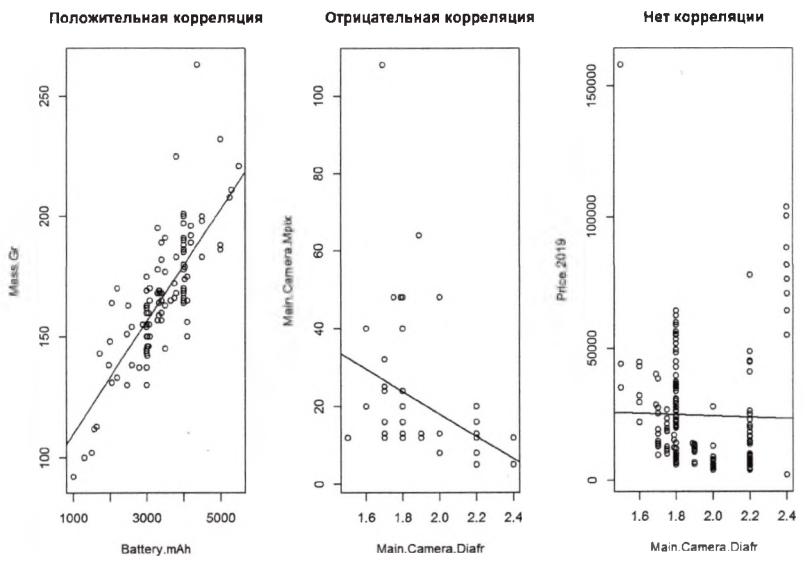
```
##
## Call:
## lm(formula = Main.Camera.Mpix ~ Main.Camera.Diafr, data = X)
##
## Coefficients:
## (Intercept) Main.Camera.Diafr
## 75.13 -28.53
```

```
abline(lin.regr2)

plot(Price.2019~Main.Camera.Diafr,data=X, main="Нет корреляции")
lin.regr3=lm(Price.2019 ~ Main.Camera.Diafr, data=X)
```

```
lin.regr3
```

```
##
## Call:
## lm(formula = Price.2019 ~ Main.Camera.Diafr, data = X)
##
## Coefficients:
## (Intercept) Main.Camera.Diafr
## 29215 -2351
```

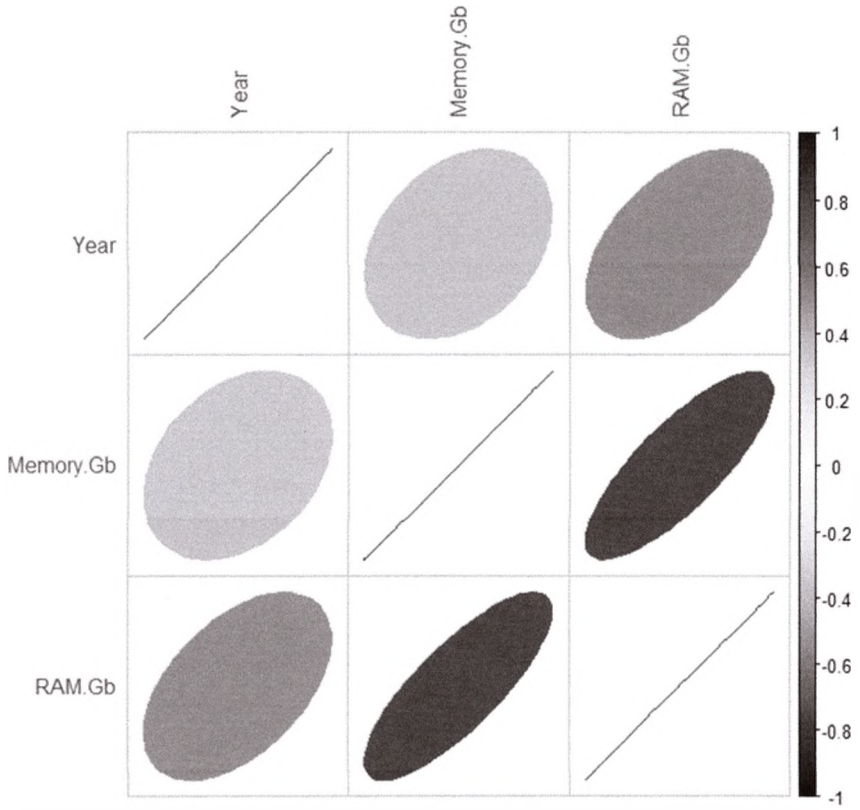
```
abline(lin.regr3)
```



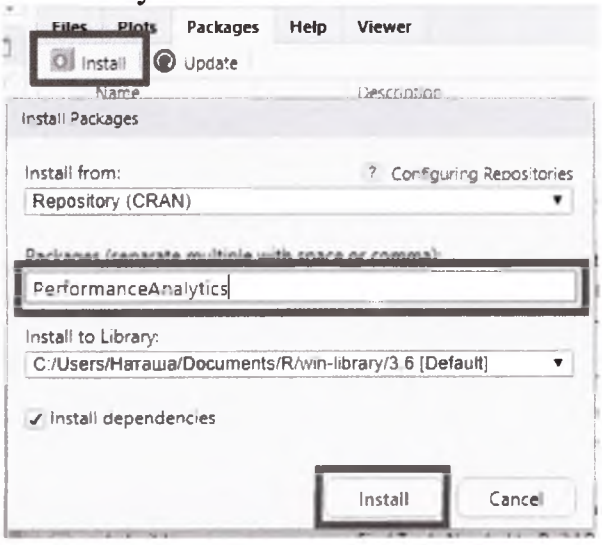
При наличии корреляции на графике одной величины против другой точки будут стремиться выстроиться в одну линию, с положительным (положительная корреляция) или отрицательным (отрицательная корреляция) наклоном. В идеальном случае все точки лягут на линию. Мера корреляции - коэффициент, принимающий значения от +1 до -1 (идеальные случаи). Коэффициент, близкий к 0 будет свидетельствовать об отсутствии зависимости.

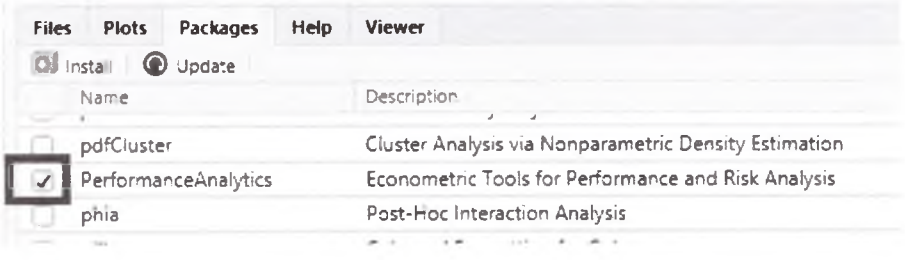
В R можно наглядно представить корреляции между парами значений с помощью библиотеки corrplot: (если библиотеки нет, то загрузить с помощью

```
>install.packages("corrplot") в окне Console)
X=read.csv("D:/R/Smarts.csv",sep=",",header=TRUE)
library(corrplot)
xc=X[c("Year","Memory.Gb","RAM.Gb")] # выбор переменных
xc=na.omit(xc) # удаление строк с NA
corr_mat=cor(xc, method = "s") # вычисление корреляций
corrplot(corr_mat,method = "ellipse") # график корреляций
```



Чем ближе график к линии, тем корреляция сильнее. Цвет определяет знак корреляции. Такая наглядная демонстрация удобна для быстрой оценки, однако можно получить и численную характеристику, с помощью библиотеки PerformanceAnalytics:



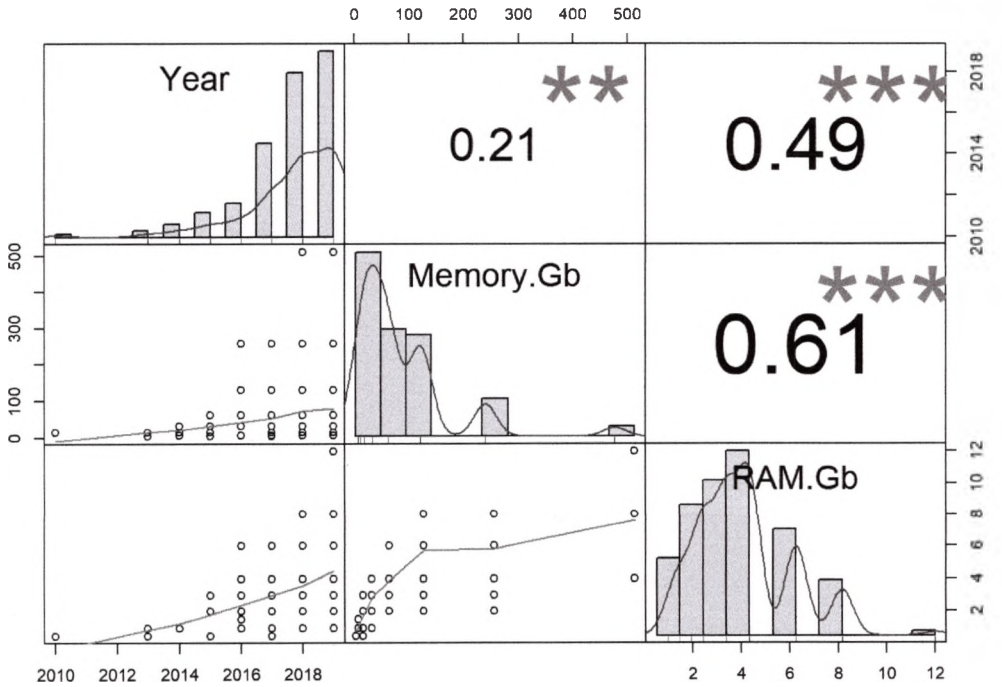


```
library(PerformanceAnalytics)
chart.Correlation(xc, histogram=TRUE,pch=19)
```

Помимо коэффициентов корреляции в верхней треугольной области добавлены обычные графики и гистограммы.

Существует примерная оценка силы корреляции:

Границы коэффициента	**Степень корреляции**
-0.3 - 0.3	**Нет корреляции**
± 0.3 - ±0.5	**Слабая корреляция**
±0.5 - ±0.7	**Умеренная**
±0.7 - ±1	**Сильная**



Примерная тематика НИРС по теме

1. Исследование корреляционной зависимости случайных величин, регрессионный анализ

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В. Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)
6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)

8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)
11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)
14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №5

Тема: Дисперсионный анализ в различных пакетах.

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.3.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для внеаудиторной работы по теме

	ВСЕГО	90	

Аннотация (краткое содержание темы):

Дисперсионный анализ — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. В отличие от t-критерия, позволяет сравнивать средние значения трёх и более групп. Разработан Р. Фишером для анализа результатов экспериментальных исследований. В литературе также встречается обозначение ANOVA (от англ. ANalysis Of VAriance)

Суть дисперсионного анализа сводится к изучению влияния одной или нескольких независимых переменных, обычно именуемых факторами, на зависимую переменную. Зависимые переменные представлены значениями абсолютных шкал (шкала отношений). Независимые переменные являются номинативными (шкала наименований), то есть отражают групповую принадлежность, и могут иметь два или более значения (типа, градации или уровня). Примерами независимой переменной X_i с двумя значениями могут служить пол (женский: X_1, мужской: X_2) или тип экспериментальной группы (контрольная: X_1, экспериментальная: X_2). Градации, соответствующие независимым выборкам объектов, называются межгрупповыми, а градации, соответствующие зависимым выборкам, — внутригрупповыми.

В зависимости от типа и количества переменных различают:

- однофакторный и многофакторный дисперсионный анализ (одна или несколько независимых переменных);
- одномерный и многомерный дисперсионный анализ (одна или несколько зависимых переменных);
- дисперсионный анализ с повторными измерениями (для зависимых выборок);
- дисперсионный анализ с постоянными факторами, случайными факторами, и смешанные модели с факторами обоих типов

Установка пакетов и рабочей директории

Установим необходимые для занятия пакеты (это действие необходимо, только если пакеты ещё не установлены на ваш компьютер) (для установки выполните следующую строку без символа # вначале).

```
# install.packages(c('ggplot2', 'reshape2', 'dplyr', 'gplots', 'compute.es', 'lsr', 'pwr'))
```

Подгружаем загруженные пакеты для их использования в текущей сессии работы R.

```
library(ggplot2)
library(reshape2)
library(dplyr)
library(gplots)
library(compute.es)
```

```
library(lsr)
library(pwr)
```
Устанавливаем рабочую директорию: через меню или командой: Session -> Set working directory -> Choose directory
```
#setwd('...')
#(вместо '...' укажите путь к папке, в которой хранится файл)
```

Однофакторный дисперсионный анализ (One-way ANOVA) с более чем 2 уровнями фактора

В R есть встроенный набор данных **ToothGrowth**, содержащий данный о влиянии витамина С на рост зубов свинок. Переменная *len* - длина зубов, переменная *dose* - доза витамина С (было протестировано три варианта дозы: 0.5, 1, 2). Глянем на структуру данных.

```
str(ToothGrowth)
```

Вывод:

```
## 'data.frame': 60 obs. of 3 variables:
## $ len : num 4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
## $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
## $ dose: num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Попробуем применить ANOVA, чтобы проверить, влияет ли дозировака витамина С на длину зубов.

```
fit2 <- aov(len ~ as.factor(dose), data=ToothGrowth)
# Обратите внимание на то, что мы одновременно изменили тип переменной dose с количественной (num, см. выше)
# на номинальную, или качественную as.factor(dose). Это необходимо для того, чтобы с помощью этой переменной
# образовать три группы.

summary(fit2)
```

Вывод:

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(dose)  2   2426    1213   67.42 9.53e-16 ***
## Residuals       57   1026      18
##                                                   ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Дисперсионный анализ показал, что дозировака витамина С влияет на длину зубов $F(2, 57) = 67.42$, p-value < 0.001. Вспомним, что p-value = 9.53e-16 означает, что если *нулевая гипотеза (H0: независимая переменная не влияет на зависимую) верна, то вероятность получить вычесленное значение критерия* F* (в данном случае $F = 67.42$) равна 9.53e-16, что сильно меньше конвенционального значения 0.05. Следовательно, мы должны отклонить нулевую гипотезу об отсутствии эффекта и пинять альтернативную, т.е. эффект дозироввки на длину зубов есть. Однако у нас три уровня фактора дозировки (три группы). Различия есть между всеми тремя? Полученные результаты не дают возможности это утверждать, т.к. различия только между двуми из них могут показать значимый эффект. Чтобы понять, между какими именно группами (уровнчми фактора) есть различия, необходимо их попарно сравнить. Для этого существуют несколько тестов. В данном случае можно воспользоваться тестом Тьюки (Tukey's HSD test).

```
TukeyHSD(fit2)
```

Вывод:

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = len ~ as.factor(dose), data = ToothGrowth)
##
## $`as.factor(dose)`
## diff lwr upr p adj
## 1-0.5 9.130 5.901805 12.358195 0.00e+00
## 2-0.5 15.495 12.266805 18.723195 0.00e+00
## 2-1 6.365 3.136805 9.593195 4.25e-05
```

Результатом применения теста является три попарных сравнения, трёх групп между собой. В колонке *p adj* приводятся значения p-value. В данном случае нулевая гипотеза о том, что между двумя группами нет различий. Поскольку во всех трёх случаях p-value меньше конвенционального значения 0.05, то различия в длине зубок есть между всеми тремя группами свинок.

Визуализируем результат с помощью более красивого рисунка. Но за красоту надо платить, поэтому придётся немного поработать руками.
Сначала рассчитаем средние значения длины зубов и их стандартные отклонения для каждой группы. Сделаем это с помощью очень полезной функции tapply. Почитайте про неё, например, здесь

```
means <- tapply(ToothGrowth$len, as.factor(ToothGrowth$dose), mean)
sd <- tapply(ToothGrowth$len, as.factor(ToothGrowth$dose), sd)
dose <- c("0.5", "1", "2")
```

```
plot_data <- data.frame(means, sd, dose)
# объединим их в отдельный data.frame, добавив к нему групирующую переменную
```

Посмотрим на получившийся data.frame

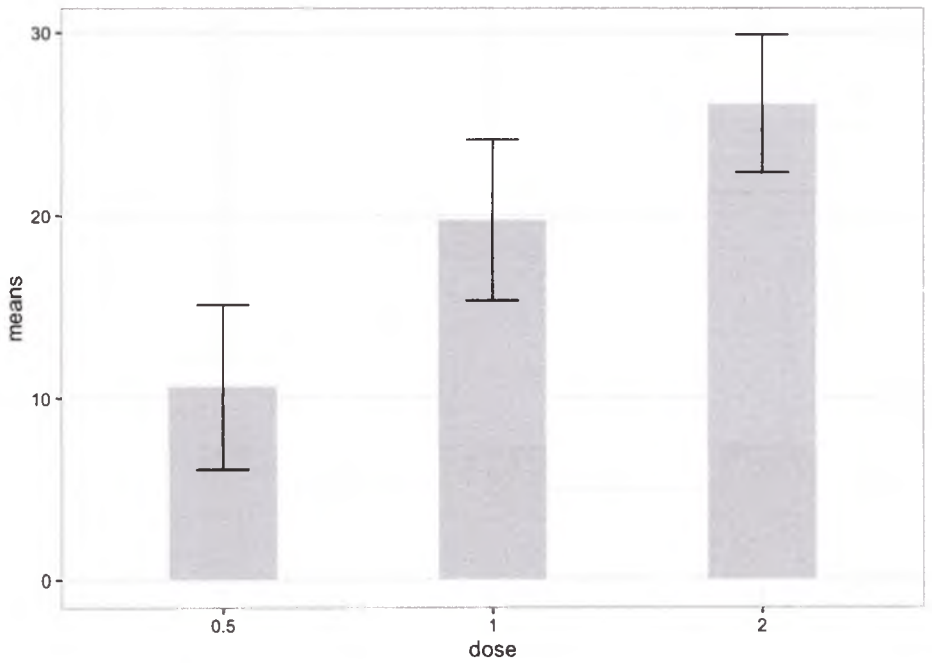
```
plot_data
```

Вывод:

```
## means sd dose
## 0.5 10.605 4.499763 0.5
## 1 19.735 4.415436 1
## 2 26.100 3.774150 2
```

Теперь на его основе построим стобликовую диаграмму с доверительными интервалами.

```
ggplot(plot_data, aes(x=dose, y=means)) +
 geom_bar(stat="identity", width=.4, fill="gold") +
 geom_errorbar(aes(ymin=means-sd, ymax=means+sd), width=.2) +
 theme_bw()
```



Двухфакторный дисперсионный анализ (Two-way ANOVA)

С помощью двухфакторного дисперсионного анализа можно проверить влияние двух факторов на зависимую переменную. Продолжим работать с данными про свинок и их зубы. Попробуем проверить, влияет ли на длину зубов и дозировака витамина С, и тип введения его в организм свинки (*supp*: апельсиновый сок (OJ) или аскорбиновая кислота (VC)).

```
fit3 <- aov(len ~ as.factor(dose) + supp, data=ToothGrowth)
# Обратите внимание на то, что мы одновременно изменили тип переменной
```

dose с количественной (num, см. выше)
# на номинальную, или качественную as.factor(dose). Это необходимо для того, чтобы с помощью этой переменной
# образовать три группы.
summary(fit3)

Вывод:

```
## Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(dose) 2 2426.4 1213.2 82.81 < 2e-16 ***
## supp 1 205.4 205.4 14.02 0.000429 ***
## Residuals 56 820.4 14.7
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Дисперсионный анализ (3x2 ANOVA) показал, что и дозировака витамина С влияет на длину зубов, $F(2, 56)=82.81$, $p < 0.001$, и тип введения его в организм свинки, $F(1, 56) = 14.02$, $p < 0.001$.

Независимые переменные могут взаимодействовать дуг с другом, и их взаимодействие также может оказывать влияние на зависимую переменную. В предыдущий анализ кроме двух основных факторов можно добавить их взаимодействие. Давайте это сделаем.

```
fit4 <- aov(len ~ as.factor(dose) + supp + as.factor(dose):supp, data=ToothGrowth)
# взаимодействие факторов добавляется через с помощью добавление в формула нового члена,
# состоящего из имен двух взаимодействующих факторов, разделённых двоеточием - as.factor(dose):supp
summary(fit4)
```

Вывод:

```
## Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(dose) 2 2426.4 1213.2 92.000 < 2e-16 ***
## supp 1 205.4 205.4 15.572 0.000231 ***
## as.factor(dose):supp 2 108.3 54.2 4.107 0.021860 *
## Residuals 54 712.1 13.2
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Можно было сделать короче: **len ~ as.factor(dose)*supp**. Эта формула добавляет сразу и две переменные по-отдельности, и их взаимодействие. Результат получится тот же самый.

```
fit5 <- aov(len ~ as.factor(dose)*supp, data=ToothGrowth)
summary(fit5)
```

Вывод:

```
## Df Sum Sq Mean Sq F value Pr(>F)
```

as.factor(dose) 2 2426.4 1213.2 92.000 < 2e-16 ***
supp 1 205.4 205.4 15.572 0.000231 ***
as.factor(dose):supp 2 108.3 54.2 4.107 0.021860 *
Residuals 54 712.1 13.2

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Дисперсионный анализ (3x2 ANOVA) показал, что есть как два основных эффекта дозы и типа введения по-отдельности, $F(2, 54) = 92.00$, $p < 0.001$, $F(1, 54) = 15.57$, $p < 0.001$, так и эффект их взаимодействия, $F(2, 54) = 4.107$, $p = 0.022$.

Посмотрим, какие именно группу отличаются друг от друга с помощью теста Тьюки.

TukeyHSD(fit5)

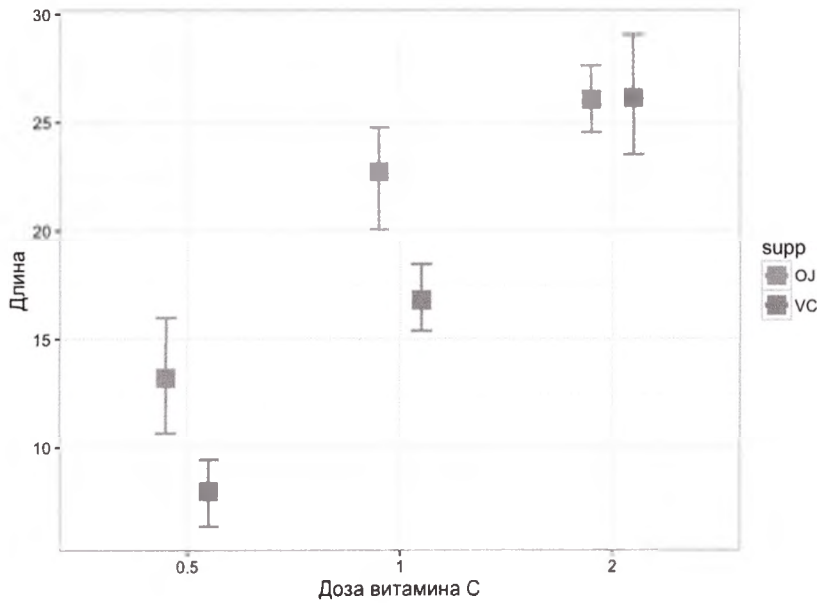
Вывод:

Tukey multiple comparisons of means
95% family-wise confidence level
##
Fit: aov(formula = len ~ as.factor(dose) * supp, data = ToothGrowth)
##
$`as.factor(dose)`
diff lwr upr p adj
1-0.5 9.130 6.362488 11.897512 0.0e+00
2-0.5 15.495 12.727488 18.262512 0.0e+00
2-1 6.365 3.597488 9.132512 2.7e-06
##
$supp
diff lwr upr p adj
VC-OJ -3.7 -5.579828 -1.820172 0.0002312
##
$`as.factor(dose):supp`
diff lwr upr p adj
1:OJ-0.5:OJ 9.47 4.671876 14.2681238 0.0000046
2:OJ-0.5:OJ 12.83 8.031876 17.6281238 0.0000000
0.5:VC-0.5:OJ -5.25 -10.048124 -0.4518762 0.0242521
1:VC-0.5:OJ 3.54 -1.258124 8.3381238 0.2640208
2:VC-0.5:OJ 12.91 8.111876 17.7081238 0.0000000
2:OJ-1:OJ 3.36 -1.438124 8.1581238 0.3187361
0.5:VC-1:OJ -14.72 -19.518124 -9.9218762 0.0000000
1:VC-1:OJ -5.93 -10.728124 -1.1318762 0.0073930
2:VC-1:OJ 3.44 -1.358124 8.2381238 0.2936430
0.5:VC-2:OJ -18.08 -22.878124 -13.2818762 0.0000000
1:VC-2:OJ -9.29 -14.088124 -4.4918762 0.0000069
2:VC-2:OJ 0.08 -4.718124 4.8781238 1.0000000

```
## 1:VC-0.5:VC 8.79 3.991876 13.5881238 0.0000210
## 2:VC-0.5:VC 18.16 13.361876 22.9581238 0.0000000
## 2:VC-1:VC 9.37 4.571876 14.1681238 0.0000058
```

Визуализируем результаты

```
pd = position_dodge(0.4)
ggplot(ToothGrowth, aes(as.factor(dose), len, color = supp)) +
 stat_summary(fun.data = mean_cl_boot, geom = 'errorbar', width = 0.2, lwd = 0.8, position = pd)+
 stat_summary(fun.data = mean_cl_boot, geom = 'line', size = 1.5, position = pd) +
 stat_summary(fun.data = mean_cl_boot, geom = 'point', size = 5, position = pd, pch=15) +
 theme_bw() +
 xlab('Доза витамина C')+
 ylab('Длина')
```



Size effect

Рассчитаем размер эффекта типа введение Витамина С на длину зубов. Воспользуемся пакетом **lsr**, в котором есть функция **etaSquared**, рассчитывающая размер эффекта eta-Squared.

```
etaSquared(fit2, anova = TRUE)
```

Вывод:

```
## eta.sq eta.sq.part SS df MS F
## as.factor(dose) 0.7028642 0.7028642 2426.434 2 1213.21717 67.41574
## Residuals 0.2971358 NA 1025.775 57 17.99605 NA
## p
## as.factor(dose) 8.881784e-16
```

Residuals NA

Размер эффекта eta-Squared = 0.703. Это эффект средней величины.

Анализ мощности

Теперь проверим, сколько необходимо людей, чтобы зафиксировать эффект такого размера. Воспользуется пакетом **pwr**, в котором есть функция **pwr.anova.test** (есть и для других дизайнов).

---

pwr.anova.test(k = 2, f= 0.703, sig.level = 0.05, power = 0.80)

Вывод:

##
Balanced one-way analysis of variance power calculation
##
k = 2
n = 9.010349
f = 0.703
sig.level = 0.05
power = 0.8
##
NOTE: n is number in each group

Анализ мощности показал, что для того, чтобы с 80% мощностью зафиксировать эффект (eta-Squared = 0.703), необходимо всего по 9 человек на группу.

---

pwr.anova.test(k = 2, f= 0.773, sig.level = 0.05, power = 0.95)

Вывод:

##
Balanced one-way analysis of variance power calculation
##
k = 2
n = 11.92686
f = 0.773
sig.level = 0.05
power = 0.95
##
NOTE: n is number in each group

Анализ мощности показал, что для того, чтобы с 95% мощностью зафиксировать эффект дозы (eta-Squared = 0.773), необходимо всего по 12 человек на группу.

Дисперсионный анализ с повторными измерениями (ANOVA with repeated measures)

Пример из книги **R в действии**. Набор данных **CO2** содержит результаты исследования холодоустойчивости северных и южных популяций злака ежовника. Сравнивали интенсивность фотосинтеза охлажденных и неохлажденных растений при разных концентрациях углекислого газа в окружающей среде. Половина растений происходила из Квебека, а половина – из штата Миссисипи.

Plant - a unique identifier for each plant.
Type - the origin of the plant / штат происхождения: Квебек или Миссисипи
Treatment - nonchilled chilled / охлаждённый неохлаждённый
conc - ambient carbon dioxide concentrations / концентрация углекислого газа в окружающей среде
uptake - carbon dioxide uptake rates / уровень потребления углекислого газа
Посмотрим на структуру данных

```
str(CO2)
```

Вывод:

```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 84 obs. of 5 variables:
## $ Plant : Ord.factor w/ 12 levels "Qn1"<"Qn2"<"Qn3"
## - attr(*, "outer")=Class 'formula' language ~Treatment * Type
## .. ..- attr(*, ".Environment")=
## - attr(*, "labels")=List of 2
## ..$ x: chr "Ambient carbon dioxide concentration"
## ..$ y: chr "CO2 uptake rate"
## - attr(*, "units")=List of 2
## ..$ x: chr "(uL/L)"
## ..$ y: chr "(umol/m^2 s)"
```

Отбираем только охлажденные растения.

```
w1b1 <- subset(CO2, Treatment=='chilled') fit6 <-
 aov(uptake ~ conc*Type + Error(Plant/conc), data = w1b1) summary(fit6)
```

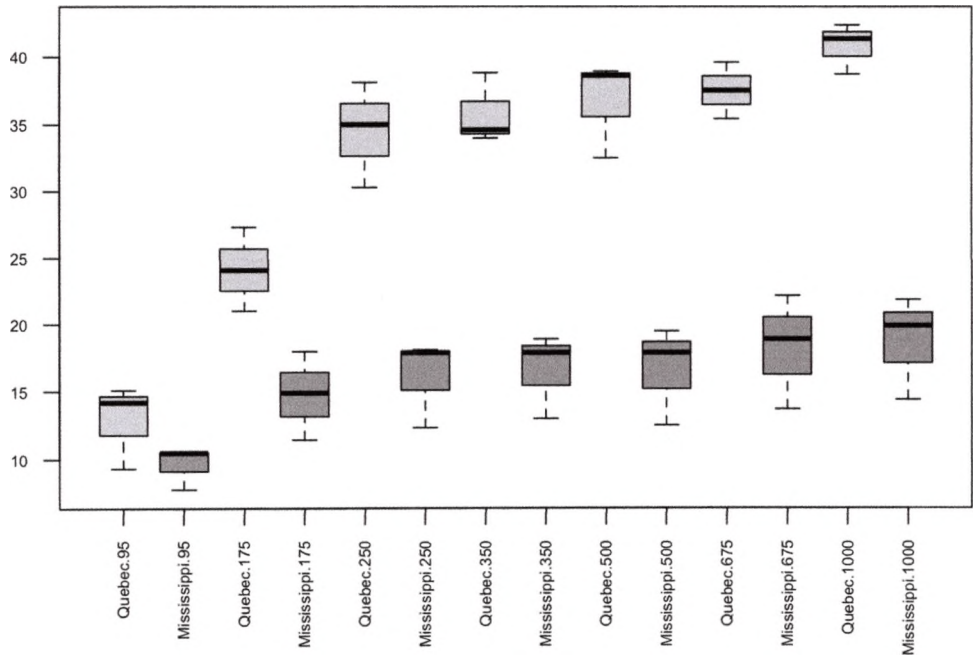
Вывод:

```
##
## Error: Plant
## Df Sum Sq Mean Sq F value Pr(>F)
## Type 1 2667.2 2667.2 60.41 0.00148 **
## Residuals 4 176.6 44.1
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Plant:conc
```

```
## Df Sum Sq Mean Sq F value Pr(>F)
## conc 1 888.6 888.6 215.46 0.000125 ***
## conc:Type 1 239.2 239.2 58.01 0.001595 **
## Residuals 4 16.5 4.1
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
## Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 30 869 28.97
```

Результаты дисперсионного анализа, представленные в таблице, свидетельствуют о том, что главные эффекты (штат и концентрация), а также взаимодействие между ними (все p-value < 0.05).
Визуализируем

```
boxplot(uptake ~ Type*conc, data=w1b1, col=(c('gold', 'green')), las = 2, cex.axis=0.6)
```



Примерная тематика НИРС по теме

1. Обработка и анализ результатов моделирования
2. Принципы математико-статистического анализа данных медико-биологических исследований

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В. Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)

6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)
11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)
14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №6

Тема: Классификационный анализ. Кластерный анализ.

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.1, ПК-4.3.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для внеаудиторной работы по теме

	ВСЕГО	90	

Аннотация (краткое содержание темы):

Основы дискриминантного анализа

Кластерный анализ позволяет разделить эмпирическую выборку на несколько классов (кластеров), однако не дает ни правил, ни четких критериев оценки качества классификации. В то же время и правила, и критерии важны прежде всего в вопросах диагностики редких, нетипичных патологических процессов, симптоматика которых весьма размыта. И особенно в процессе оказания ургентной (экстренной) медицинской помощи, когда у врача на перебор вариантов лечебно-диагностической тактики считанные минуты. Для решения подобных задач и существует **дискриминантный анализ**. И хотя дискриминантный и кластерный анализы близки по сути (направлены на решение задач классификации), но подходами к классификации принципиально различаются. **Дискриминантный анализ**, как и кластерный анализ, направлен на разделение выборки на ряд кластеров, но его конечная цель - **отнесение некоторого объекта к одному из уже построенных классов, а также проверка непротиворечивости классификации**. Термин <<дискриминация>> (от лат. discriminacio - разделение) означает не только разделение объектов на классы, но и ограничение такого разделения. Это ряд методов, с помощью которых мы можем отнести новый объект к одному из заранее построенных классов, а также проверить качество построенной классификации. Еще дискриминантный анализ называют **анализом с обучающей выборкой для распознавания образов** или **классификацией с обучением**. Кластеризация, многомерное шкалирование, эмпирическое классифицирование основывается на экспертных оценках на основании профессионального опыта врача-диагноста.

Постановка задачи дискриминантного анализа

Цель дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик (признаков, параметров) объекта **классифицировать** его, то есть отнести к одной из нескольких групп (классов) некоторым оптимальным способом.

Под оптимальным способом понимается либо **минимум математического ожидания потерь**, либо **минимум вероятности ложной классификации**.

Дискриминантный анализ является **одним из методов многомерного статистического анализа**, поскольку измеряется несколько параметров объекта, например, давление, состав крови, температура и так далее. Так, в спортивной медицине объектом исследования является спортсмен, когда по результатам измерений различных параметров, проведения диагностических тестов врач определяет степень подготовки спортсмена к участию в соревнованиях.

Математическая постановка задачи

Предположим, имеется n объектов с m характеристиками. В результате измерений каждый объект характеризуется вектором x1, ... , xm, m > 1.

Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких групп (классов) G1,..., Gk, $k \geq 2$.

Иными словами, нужно построить решающее правило, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. **Число групп заранее известно**, также известно, что объект заведомо принадлежит к определенной группе.

Пусть X - пространство значений вектора измерений.

Решающее правило называется **нерандомизированным**, если пространство X разбито на k непересекающихся областей; при попадании измерения параметров объекта в k-ю область объект относится к k-й группе. Решающее правило называется **рандомизированным**, если для каждого вектора наблюдений x задана вероятность $p_i(x)$, с которой объект принадлежит i-й группе, $p_i(x) \geq 0$, $p_1(x) + ... + p_k(x) = 1$, i = 1, ..., k.

Очевидно, при использовании **решающего правила** возникают потери, вызванные тем, что объект неправильно классифицирован - отнесен к классу i, когда в действительности он принадлежит классу j ($i \neq j$).

Если можно измерить убыток r(i, j) при неправильной классификации объекта, то вводят средние потери, к которым приводит применение данного правила, и пытаются найти правило, минимизирующее эти средние потери.

Если **значение потерь трудно оценить численно**, то при построении оптимального правила используют **критерий минимальной вероятности ложной классификации**.

В дискриминантном анализе можно задать **априорные** вероятности принадлежности объекта к определенному классу. На практике эти вероятности оцениваются из массива экспериментальных данных. Дискриминантный анализ <<работает>> при выполнении следующих предположений и ограничений:

Нормальное распределение. Предполагается, что анализируемые переменные - измеряемые характеристики объекта - представляют выборку из многомерного нормального распределения.

Однородность дисперсий и ковариаций. Предполагается, что дисперсии и ковариации наблюдаемых переменных в разных классах однородны.

Умеренные отклонения от данных предположений допустимы.

Алгоритм проверки возможности проведения дискриминантного анализа:

- Проверить, создана ли выборка в интервальных шкалах или шкалах отношений, имеют ли признаки нормальное распределение.
- Проверить, разделена ли выборка на конечное число (не менее двух) непересекающихся классов, известна ли для каждого объекта вероятность принадлежности к какому-то классу.

- Проверить отсутствие корреляции между переменными с помощью корреляционной матрицы. При наличии зависимости между средними по совокупностям дисперсиями или стандартными отклонениями (мультиколлинеарности) не существует однозначной меры относительной важности переменных.
- В каждом классе должно быть не менее двух объектов из обучающей выборки, а число дискриминантных переменных не должно превосходить объем обучающей выборки за вычетом двух объектов.

Основные вопросы дискриминации

- Принадлежит ли произвольно выбранный объект из генеральной совокупности к одному из классов, на которые разделена эмпирическая выборка, и можно ли конструировать правило классификации. Можно ли систему распознавания научить определять принадлежность объекта к тому или иному классу?
- Каково качество построенной классификации: насколько она чутка к разделению объектов на классы, насколько такая дифференцировка достоверна?
- Каковы информативные признаки из числа измеряемых у исследуемых объектов, какие из них имеют наибольшее значение для правильного и качественного дифференцирования.

Существует ряд разновидностей дискриминантного анализа, но математическая сугь у них едина, поэтому для практического применения рассмотрим основные направления дискриминантного анализа, реализованные в большинстве статистических пакетов:

- линейный дискриминантный анализ Фишера;
- канонический дискриминантный анализ (максимального правдоподобия, или вероятностный);
- методы, связанные с расстояниями;
- пошаговый дискриминантный анализ.

Линейный дискриминантный анализ Фишера

Строятся k линейных функций классификации, предназначенных для определения того, к какой группе наиболее вероятно может быть отнесен каждый объект. Соответствующие функции называются **линейными классификационными функциями** (ЛКФ) Фишера. Количество функций классификации равно количеству классов или групп. Для каждого объекта и для каждой совокупности вычисляются значения ЛФК по следующей формуле:

$$d_{mk} = a_k + b_{k1}\,x_{1k} + b_{k2}\,x_{2k} + \ldots + b_{kn}\,x_{nk}$$

или

$$d_{mk} = a_k + \Sigma\, b_{kn}\,x_{nk}, m = 1,\ldots,n;\ k = 1,\ldots,g,$$

где k - обозначает соответствующую группу;

g - количество групп;

m - номер объекта;

b_{ki} - коэффициенты, которые называют весами для i-й переменной при вычислении показателя классификации для k-й совокупности;

d_{mk} - значение ЛКФ для m-го объекта в группе k (показатель классификации);

a_k - свободный член уравнения;

x_{mj} – наблюдаемое значение j-й переменной для соответствующего m-го объекта в группе k.

Наблюдение (новый объект) приписывают к той группе, для которой классификационная функция имеет максимальное значение.

В основе метода Фишера лежит еще одно предположение, накладываемое на ковариации переменных: **признаки должны иметь статистически идентичные ковариационные матрицы**. Ковариация двух переменных - мера их совместного изменения, равноценна коэффициенту корреляции Пирсона. Однако показатель ковариации в отличие от коэффициента Пирсона может принимать произвольные значения, а не только в пределах: $-1 \leq r \leq 1$.

Канонический дискриминантный метод

Канонический дискриминантный анализ относит объект к классу k, если соответствующая апостериорная вероятность этой принадлежности максимальна.

Применяемые в этом методе **линейные дискриминантные функции** часто называют каноническими (КЛДФ). Данный анализ проводится по схеме, обратной первому виду анализа. Здесь разделение объектов ведется по минимальным значениям дискриминирующей функции. Объект относится к определенному классу только тогда, когда **Евклидово расстояние** от центра кластера до оцениваемого показателя минимально.

Примечание. Следует обратить самое серьезное внимание на **обязательную нормальность распределения в генеральной совокупности**, которая часто не выполняется для эмпирических данных. Пренебрежение этим может привести к серьезным ошибкам классификации.

Каноническая линейная дискриминантная функция имеет следующий вид:

$$D_{mk} = a_k + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n, m = 1,\ldots,n;\ k = 1,\ldots,g,$$

где

D_{mk} - значение канонической дискриминантной функции для m-го объекта в группе k;

a_k - свободный член уравнения;

g - количество групп;

x_{mi} - наблюдаемое значение i-й переменной для соответствующего m-го объекта в группе k;

b_{ki} - коэффициенты, которые оценивают с помощью дискриминантного анализа.

После того, как проведена оценка статистической значимости каждой канонической дискриминантной функции и определено, какие из них вносят наибольший вклад в дискриминацию, рассчитывают значения этих функций для каждого объекта (наблюдения). Наименьшее из значений применяют для классификации. Его сравнивают со средними значениями расстояний до центроидов каждой группы. Объект принадлежит к той группе, расстояние до которой наилучшим образом совпадает с рассчитанным значением КЛДФ. В случае применения по умолчанию **методов максимального правдаподобия** используют два набора оценок:

1. **Априорные вероятности принадлежности к классу** можно рассматривать как решающее правило, применяемое в том случае, когда нет никакой дополнительной информации об объектах. Их вычисляют либо по числу объектов каждого класса, либо считают равными друг другу.

2. **Условные вероятности принадлежности к классу** равны вероятности получить соответствующее значение дискриминантной функции при условии, что объект принадлежит к данному классу. Используется предположение о том, что значения дискриминантных функций распределены нормально.

Методы, связанные с расстояниями

Методы, связанные с расстояниями, рассматривают объекты как точки в Евклидовом пространстве. В качестве меры сходства между объектами при классификации можно использовать, например, Евклидово расстояние между объектами. **Чем меньше расстояние между объектами, тем больше сходство**.

Однако в тех случаях, когда переменные коррелированы, измерены в разных единицах и имеют различные стандартные отклонерия, трудно четко определить понятие <<расстояния>>. В этом случае полезнее применить не Евклидово расстояние, а выборочное расстояние Махаланобиса.

Расстояние Махаланобиса определяется как расстояние от наблюдаемой точки до центра тяжести в многомерном пространстве, определяемом коррелированными (неортогональными) независимыми переменными. Если независимые переменные некоррелированы, расстояние Махаланобиса совпадает с обычным Евклидовым расстоянием.

Для каждой совокупности в выборке можно определить положение точки, представляющей средние для всех переменных в многомерном пространстве,
определенном переменными рассматриваемой модели. Эти точки называются центроидами группы.

Для каждого наблюдения можно вычислить его расстояние Махаланобиса от каждого центроида группы. Мы признаем наблюдение принадлежащим к той группе, к которой он ближе, то есть **когда расстояние Махаланобиса до нее минимально**.

Апостериорные вероятности классификации

Используя для классификации расстояние Махаланобиса, можно получить вероятность того, что образец принадлежит к конкретной совокупности. Это значение будет не вполне точным так как распределение вокруг среднего для каждой совокупности будет не в точности нормальным. Поскольку принадлежность каждого образца вычисляется по априорному знанию модельных переменных, эти вероятности называются **апостериорными вероятностями**. Апостериорные вероятности это вероятности, вычисленные с использованием знания значений других переменных для образцов из частной совокупности. Имеется одно дополнительное обстоятельство, которое следует рассмотреть при классификации наблюдений. Иногда известно заранее, что в одной из групп имеется больше наблюдений, чем в другой. Поэтому априорные вероятности того, что образец принадлежит такой группе, выше. Можно установить различные априорные вероятности, которые будут затем использоваться для уточнения результатов классификации наблюдений (и для вычисления апостериорных вероятностей).

Пошаговый дискриминантный анализ

Пошаговый дискриминантный анализ вводит переменные последовательно, исходя их способности различать (дискриминировать) группы.

При пошаговом анализе <<с включением>> на каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

При пошаговом анализе <<с исключением>> движутся в обратном направлении. В этом случае все переменные сначала будут включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в различение. Тогда в качестве результата успешного анализа можно сохранить только <<важные>> переменные в модели, то есть те переменные, чей вклад в дискриминацию больше остальных.

Пошаговый дискриминантный анализ основан на использовании уровня значимости F-статистики. Он достаточно прост в реализации при компьютерной обработке данных и помогает наглядно оценивать качество полученной классификации, являясь дополнительным методом к двум вышеупомянутым.

Итог классификации

Общим результатом, на который следует обратить внимание при оценке качества текущей функции классификации, является **матрица классификации.**

Матрица классификации содержит число образцов, корректно классифицированных (на диагонали матрицы) и тех, которые попали не в свои совокупности (группы).

Примерная тематика НИРС по теме

1. Классификационный анализ в медико-биологических исследованиях

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В. Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)

5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)
6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)
11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)
14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №7

Тема: Анализ выживаемости. Анализ временных рядов.

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-9.1.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для внеаудиторной работы по теме

	ВСЕГО	90	

Аннотация (краткое содержание темы):

Анализ времени до наступления события в медицине чаще всего называется анализом выживаемости, так как в медицинских исследованиях принято обязательно оценивать вероятность выживания во времени, а событием (конечной точкой), по которому она определяется, является смерть.

На самом деле этим событием может быть не только смерть, но и любое другое. Функция же времени тоже может быть функцией любой другой количественной непрерывной переменной.

Ожидаемое событие при указанном анализе может не наступить. Такие случаи, когда событие ещё не наступило или о нём неизвестно, называются цензурированными. На цензурированных событиях, при конечной точке «смерть» это будет событие «жив», и основан весь анализ времени до наступления события, поэтому он получил название анализа выживаемости, а кривая её отражающая – кривой выживаемости. Другое название представляемого анализа – метод Каплана-Мейера, а кривой – кривая Каплана-Мейера.

По оси Y в кривой Каплана-Мейера откладывается процент или доля выживших (для других переменных это будет процент или доля пациентов с отсутствием события или отсутствием проецируемой точки по оси X), по оси X – время наблюдения (или для других переменных – числовое количественное непрерывное значение).

Условия для применения метода Каплана-Мейера:

1. Момент начала наблюдения должен быть четко определен (например, момент появления первых симптомов, установления диагноза и т.д.).
2. Момент исхода должен быть четко определен (например, антенатальная гибель плода, момент диагностики клинического узкого таза и т.д.).
3. Для нецензурированных случаев необходимо знать дату исхода или период времени от начала наблюдения до развития исхода.
4. Для цензурированных случаев необходимо знать дату последнего контакта или период времени от начала наблюдения до последнего контакта.
5. Цензурированные наблюдения не должны отличаться по выживаемости от нецензурированных.
6. Методы оценки выживаемости и определения исхода должны быть одинаковыми для объектов исследования, включенных как на ранних, так и на более поздних сроках исследования.
7. Условия исследования не должны меняться с течением времени и не должны влиять на выживаемость (например, лечение должно быть одинаковым на протяжении всего времени исследования для всех случаев).

8. Доли цензурированных случаев не должны статистически значимо различаться в исследуемых группах до момента окончания периода наблюдения.
9. В анализируемой выборке должно быть более 30 исследуемых объектов.

Важным в анализе является представление не только кривой, но и оценок частоты наступления события (медианы и/или средней и 95% доверительного интервала) с указанием статистической значимости различий оценок наступления событий в анализируемых выборках (расхождения кривых на момент оценки) по критериям Бреслау (другое название - генерализованный критерий Вилкоксона), учитывающий ранние различия в вероятностях события, и Лог Ранка (другое название – Мантела-Кокса), учитывающий поздние различия в вероятностях события.

R — Анализ выживания

Анализ выживания имеет дело с предсказанием времени, когда определенное событие произойдет. Это также известно как анализ времени отказа или анализ времени до смерти. Например, прогнозирование количества дней, в течение которых человек с раком выживет, или прогнозирование времени, когда механическая система выйдет из строя.

Пакет R под названием « **выживание»** используется для анализа выживаемости. Этот пакет содержит функцию **Surv (),** которая принимает входные данные в виде формулы R и создает объект выживания среди выбранных переменных для анализа. Затем мы используем функцию **Survfit (),** чтобы создать график для анализа.

Установить пакет

install.packages("survival")

Синтаксис

Основной синтаксис для создания анализа выживаемости в R —

Surv(time,event)

survfit(formula)

Ниже приведено описание используемых параметров:

время — время наблюдения до наступления события.

Событие указывает на состояние возникновения ожидаемого события.

формула — это отношение между переменными предиктора.

Пример

Мы рассмотрим набор данных с именем «pbc», присутствующий в пакетах выживания, установленных выше. Он описывает данные о выживаемости людей с первичным билиарным циррозом печени. Среди множества столбцов, представленных в наборе данных, мы в первую очередь занимаемся полями «время» и «статус». Время представляет собой количество дней между регистрацией пациента и более ранним событием между пациентом, получающим трансплантацию печени, или смертью пациента.

```
# Load the library.
library("survival")

# Print first few rows.
print(head(pbc))
```

Когда мы выполняем приведенный выше код, он дает следующий результат и диаграмму —

```
  id time status trt      age sex ascites hepato spiders edema bili chol
1  1  400      2   1 58.76523   f       1      1       1   1.0 14.5  261
2  2 4500      0   1 56.44627   f       0      1       1   0.0  1.1  302
3  3 1012      2   1 70.07255   m       0      0       0   0.5  1.4  176
4  4 1925      2   1 54.74059   f       0      1       1   0.5  1.8  244
5  5 1504      1   2 38.10541   f       0      1       1   0.0  3.4  279
6  6 2503      2   2 66.25873   f       0      1       0   0.0  0.8  248
  albumin copper alk.phos    ast trig platelet protime stage
1    2.60    156   1718.0 137.95  172      190    12.2     4
2    4.14     54   7394.8 113.52   88      221    10.6     3
3    3.48    210    516.0  96.10   55      151    12.0     4
4    2.54     64   6121.8  60.63   92      183    10.3     4
5    3.53    143    671.0 113.15   72      136    10.9     3
6    3.98     50    944.0  93.00   63       NA    11.0     3
```

Из приведенных выше данных мы рассматриваем время и статус для нашего анализа.

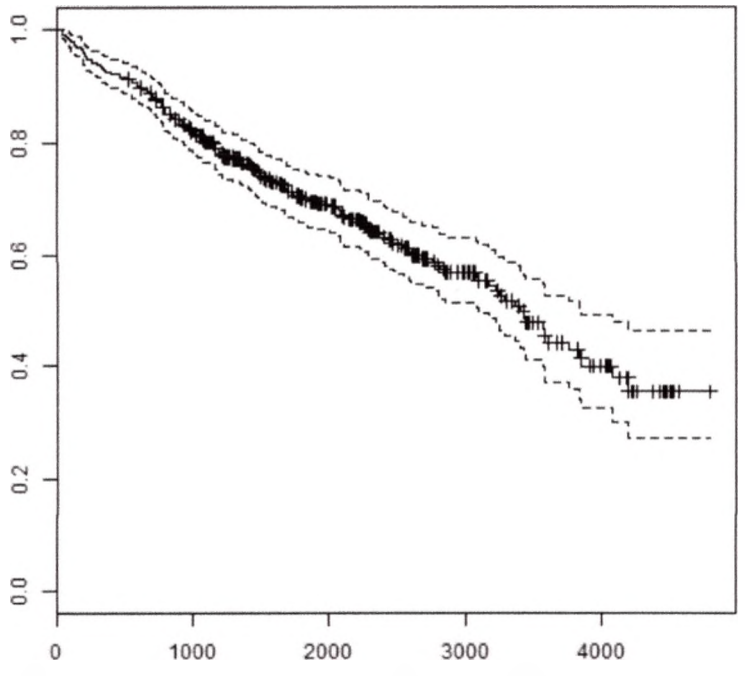
Применение функций Surv () и Survfit ()

Теперь мы переходим к применению функции **Surv ()** к указанному выше набору данных и создаем график, который покажет тренд.

```
# Load the library.
library("survival")

# Create the survival object.
survfit(Surv(pbc$time,pbc$status == 2)~1)

# Give the chart file a name.
png(file = "survival.png")

# Plot the graph.
plot(survfit(Surv(pbc$time,pbc$status == 2)~1))

# Save the file.
dev.off()
```

Когда мы выполняем приведенный выше код, он дает следующий результат и диаграмму —

```
Call: survfit(formula = Surv(pbc$time, pbc$status == 2) ~ 1)

      n  events  median 0.95LCL 0.95UCL
    418     161    3395    3090    3853
```



Тенденция на приведенном выше графике помогает нам прогнозировать вероятность выживания в конце определенного количества дней.

Примерная тематика НИРС по теме

1. Анализ времени до наступления события.

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В.

Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)
6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)
11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)

14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)

Практическое занятие №8

Тема: Систематизация пройденного материала, зачет (В интерактивной форме).

Разновидность занятия: комбинированное.

Методы обучения: объяснительно-иллюстративный, репродуктивный, метод проблемного изложения, частично-поисковый, исследовательский.

Значение темы (актуальность изучаемой проблемы): Научные исследования в сфере медицины и оздоровительных технологий приводят к накоплению большого количества данных о воздействии реабилитационных, терапевтических и болезнетворных факторов на организм человека, которые требуют количественной оценки и интерпретации. Обработка экспериментальных данных в настоящее время может осуществляться на компьютере в статистических пакетах.

Формируемые компетенции: ПК-4.1, ПК-4.3, ПК-9.1.

Место проведения и оснащение практического занятия: Компьютерный класс №6 (4-60/1) – видеопроектор, доска магнитно-маркерная, комплект учебной мебели на посадочные места, локальный сетевой сервер, персональные компьютеры, экран.

Структура содержания темы (хронокарта практического занятия)

п/п	Этапы практического занятия	Продолжительность (мин.)	Содержание этапа и оснащенность
1	Организация занятия	5.00	Проверка посещаемости и внешнего вида обучающихся
2	Формулировка темы и целей	10.00	Озвучивание преподавателем темы и ее актуальности, целей занятия
3	Контроль исходного уровня знаний и умений	10.00	Тестирование, индивидуальный устный или письменный опрос, фронтальный опрос
4	Раскрытие учебно-целевых вопросов по теме занятия	10.00	Изложение основных положений темы
5	Самостоятельная работа обучающихся (текущий контроль)	40.00	Выполнение практического задания
6	Итоговый контроль знаний (письменно или устно)	10.00	Тесты по теме, ситуационные задачи
7	Задание на дом (на следующее занятие)	5.00	Учебно-методические разработки следующего занятия и методические разработки для

			внеаудиторной работы по теме
	ВСЕГО	90	

Аннотация (краткое содержание темы):
В контрольной работе требуется продемонстрировать умение обращаться с данными и грамотно подавать информацию, построить и скомпоновать три графика - линейную регрессию, boxplot и кластерную картинку.

Задание по однофакторной линейной регрессии + пакет ggplot

- Постройте на графике облако точек;
- Найдите коэффициенты линейной регрессии;
- Совместите на графике линию регрессии с облаком точек;
- Оцените визуально характер зависимости признаков;
- Добавить уравнение регрессии и коэффициент детерминации в область графика (использовать пакет ggplot)

Выберите пару (X, Y) из предложенного списка файлов согласно распределению вариантов или возьмите данные, смоделированные Вами для двух признаков.

Задание по теме классификационный анализ (метод главных компонент)

Для проведения анализа возьмем данные Австралийского института спорта (Australian Institute of Sport).

Данные для исследования - мужчины-спортсмены, занимающиеся следующими видами спорта: BBall; Row; WPolo, Swin

Примерная тематика НИРС по теме

На этом занятии НИРС не предусмотрен.

Основная литература

1. Балдин, К. В. Теория вероятностей и математическая статистика : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев. - 2-е изд. - М. : Дашков и К, 2014. - 473 с. - Текст : электронный.

Дополнительная литература

1. Боровиков, В. П. Популярное введение в современный анализ данных в системе STATISTICA : учеб. пособие для вузов / В. П. Боровиков. - М. : Горячая линия-Телеком, 2018. - 288 с. : ил. - Текст : электронный.
2. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. - Москва : ГЭОТАР-Медиа, 2016. - Текст : электронный.
3. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рукосуев ; ред. К. В. Балдин. - 4-е изд., стер. - Москва : ФЛИНТА, 2016. - 489 с. - Текст : электронный.
4. Наркевич, А. Н. Статистические методы исследования в медицине и биологии : учеб. пособие / А. Н. Наркевич, К. А. Виноградов, К. В.

Шадрин ; Красноярский медицинский университет. - Красноярск : КрасГМУ, 2018. - 109 с. - Текст : электронный.
5. Обмачевская, С. Н. Медицинская информатика. Курс лекций : учебное пособие для вузов / С. Н. Обмачевская. - 4-е изд., стер. - Санкт-Петербург : Лань, 2022. - 184 с. - Текст : электронный.
6. Информатика и медицинская статистика : учебное пособие / ред. Г. Н. Царик. - Москва : ГЭОТАР-Медиа, 2017. - 304 с. - Текст : электронный.
7. Малугин, В. А. Математическая статистика : учебное пособие для вузов / В. А. Малугин. - Москва : Юрайт, 2020. - 218 с. - Текст : электронный.
8. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 1. - 471 с. - Текст : электронный.
9. Медик, В. А. Математическая статистика в медицине : учебное пособие для вузов : в 2 т. / В. А. Медик, М. С. Токмачев. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. - Т. 2. - 347 с. - Текст : электронный.

Электронные ресурсы

1. Электронный учебник по статистике (http://statsoft.ru/home/textbook/default.htm)
2. АНАЛИЗ И ОБРАБОТКА ДАННЫХ: ТЕОРИЯ, МЕТОДОЛОГИЯ, ПРАКТИКА (http://www.statproject.ru/)
3. Открытая лекция для студентов медицинских вузов (https://www.youtube.com/watch?v=x5QqBjerFdg&t=4868s)
4. Статистический анализ клинических испытаний (https://www.youtube.com/watch?v=aBIN1Sq-UYY)
5. Лекция 1. Анализ данных на R в примерах и задачах (https://www.youtube.com/watch?v=8mwJ3mEjdIg&list=PLlb7e2G7aSpSSa_PlFEwnd6-3gzAa08_m)
6. Официальный сайт проекта The R-Project for statistical computing (http://www.r-project.org/)
7. Официальный сайт федеральной службы государственной статистики (Росстат) (http://www.gks.ru/)
8. Основы анализа данных (R) (https://www.youtube.com/channel/UCLk-Oih8VlqF-StidijTUnw/featured)
9. Классификация, регрессия и другие алгоритмы Data Mining с использованием R (https://ranalytics.github.io/data-mining/index.html)
10. Визуализация и анализ географических данных на языке R. Глава 6 Продвинутая графика (https://tsamsonov.github.io/r-geo-course/advgraphics.html)
11. Законы распределения вероятностей, реализованные в R (https://r-analytics.blogspot.com/2012/12/r.html#.WbWaWshJaUk)
12. Классические методы статистики: t-критерий Стьюдента в R (https://r-analytics.blogspot.com/2012/03/t.html)
13. Классические методы статистики: критерий Уилкоксона в R (https://r-analytics.blogspot.com/2012/05/blog-post_20.html)

14. Однофакторный дисперсионный анализ: введение (https://r-analytics.blogspot.com/2013/01/blog-post.html)
15. Двухфакторный дисперсионный анализ (https://r-analytics.blogspot.com/2013/04/blog-post.html)
16. «Анализ данных на Python» в двух частях (https://habr.com/ru/company/JetBrains-education/blog/438058/)