

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

С.В. Родькин

БИОСТАТИСТИКА

Учебно-методическое пособие

Ростов-на-Дону
ДГТУ
2024

УДК 57.087.1
Р60

Рецензенты:

доктор биологических наук, профессор
зав. кафедрой «Биоинженерия» *Е.Ю. Кириченко*
(Донской государственный технический университет);

кандидат биологических наук, доцент *М.Ю. Бибов*
(Ростовский государственный медицинский университет)

Родькин, Станислав Владимирович.

Р60 Биостатистика : учебно-методическое пособие / С.В. Родькин ;
Донской государственный технический университет. – Ростов-на-
Дону : ДГТУ, 2024. – 281 с.

ISBN 978-5-7890-2251-1

Детально изложены теория и методы статистического анализа в программных пакетах MS Excel и JASP.

Предназначено для студентов биологических, медицинских и ветеринарных вузов, а также преподавателей и сотрудников, осуществляющих статистическую обработку медико-биологических данных.

УДК 57.087.1

Печатается по решению редакционно-издательского совета
Донского государственного технического университета

ISBN 978-5-7890-2251-1

© Родькин С.В., 2024
© Донской государственный
технический университет, 2024

Предисловие

Статистика играет важную роль в современной биологии, позволяя исследователям анализировать и интерпретировать данные, полученные в ходе биологических экспериментов. Биостатистика – это область статистики, которая специализируется на применении статистических методов к биологическим данным. Она играет ключевую роль в разработке новых лекарств и вакцин, оценке эффективности лечения, проведении клинических испытаний, исследовании генетических факторов риска заболеваний и многих других областях медицины и биологии.

Основными задачами биостатистики являются сбор и обработка данных, анализ и интерпретация результатов исследований, прогнозирование развития заболеваний, оценка рисков и выгод различных методов лечения, а также разработка статистических моделей для принятия решений в области здравоохранения и т. д. Кроме того, биостатистика тесно связана с другими областями науки, такими как эпидемиология, молекулярная биология, фармакология и генетика. Для решения этих задач биостатистика использует широкий спектр статистических методов, начиная от простых описательных статистик и заканчивая сложными моделями машинного обучения.

Целью данного учебно-методического пособия является формирование у обучающихся комплексного представления об анализе биостатистических данных, а также выработка практических навыков их обработки в программных пакетах MS Excel и JASP.

Пособие имеет четкую структуру и состоит из двух ключевых блоков – вводной и основной частей. В вводной части в доступной форме представлен ряд теоретических положений и простых практических работ, направленных на освоение рабочей среды MS Excel и JASP. Основная часть включает теорию и практику работы с биостатистическими данными – от самых простых описательных методов до сложных процедур кластерного и факторного анализа – с детализованным описанием всех шагов.

После освоения материала, представленного в этом пособии, у обучающихся будет сформировано комплексное представление о биостатистике, а также освоены навыки практической работы со статистическими биологическими данными для решения реальных задач в области биологии и медицины.

Учебно-методическое пособие подготовлено на основе результатов работы, выполненной при финансовой поддержке гранта Министерства науки и высшего образования РФ № FZNE-2024-0004.

1. ОСВОЕНИЕ ПРОГРАММНОЙ СРЕДЫ MS EXCEL И JASP (ВВОДНАЯ ЧАСТЬ)

1.1. Первоначальное знакомство с MS Excel

Программа Microsoft Excel (далее – MS Excel) служит для создания электронных и печатных документов, содержащих различные типы данных, включая текстовые, числовые и графические. Кроме того, она предлагает функционал для автоматизации математических расчетов и представления результатов в удобном для пользователя формате. Все данные, вводимые в программу и обрабатываемые в ней, представлены в виде таблиц в окне приложения.

Запуск MS Excel

Существует несколько способов запуска программы MS Excel:

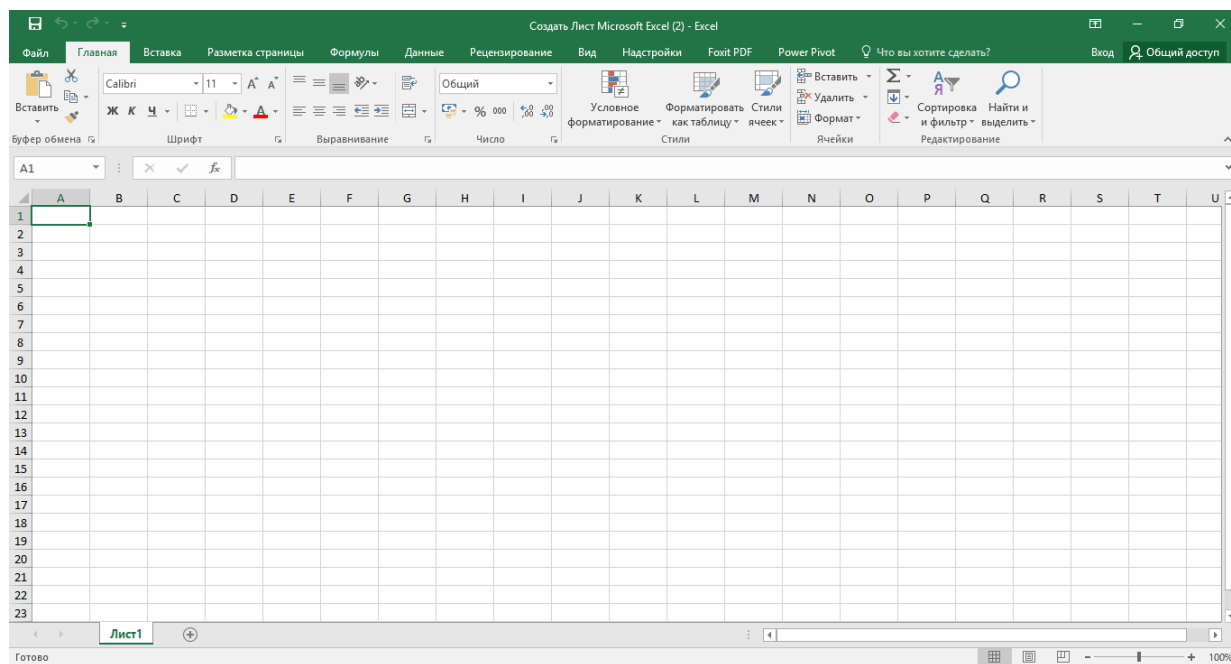
1. Двойной клик по ярлыку на рабочем столе, если он есть.
2. Нажать кнопку «Пуск» и в списке «Программы» найти кнопку «Excel 2016», кликнуть по ней.
3. Двойной клик по уже существующему файлу документа (.xlsx), чтобы открыть его.

Интерфейс MS Excel 2016

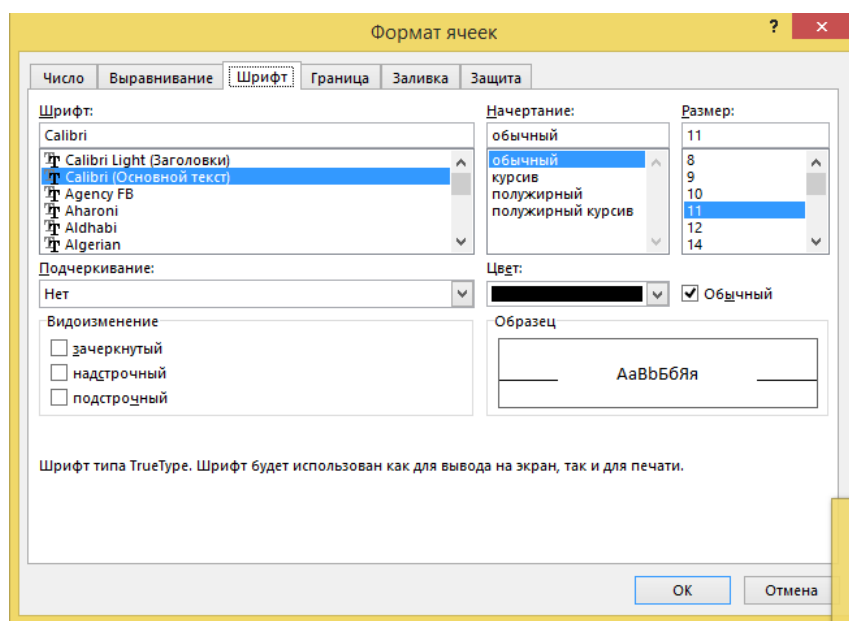
Интерфейс MS Excel 2016 представляет собой набор инструментов для работы с электронными таблицами. Он включает в себя различные элементы управления, такие как кнопки, раскрывающиеся кнопки, списки, раскрывающиеся списки, счетчики, кнопки с меню, флажки и значки.

Данные элементы позволяют пользователям выполнять различные задачи, такие как изменение стиля текста или открытие новых окон. Например, кнопки используются для выполнения определенных действий. Если нажать на кнопку «Полужирный» в группе «Шрифт» на вкладке «Главная», то можно изменить начертание текста на полужирное.

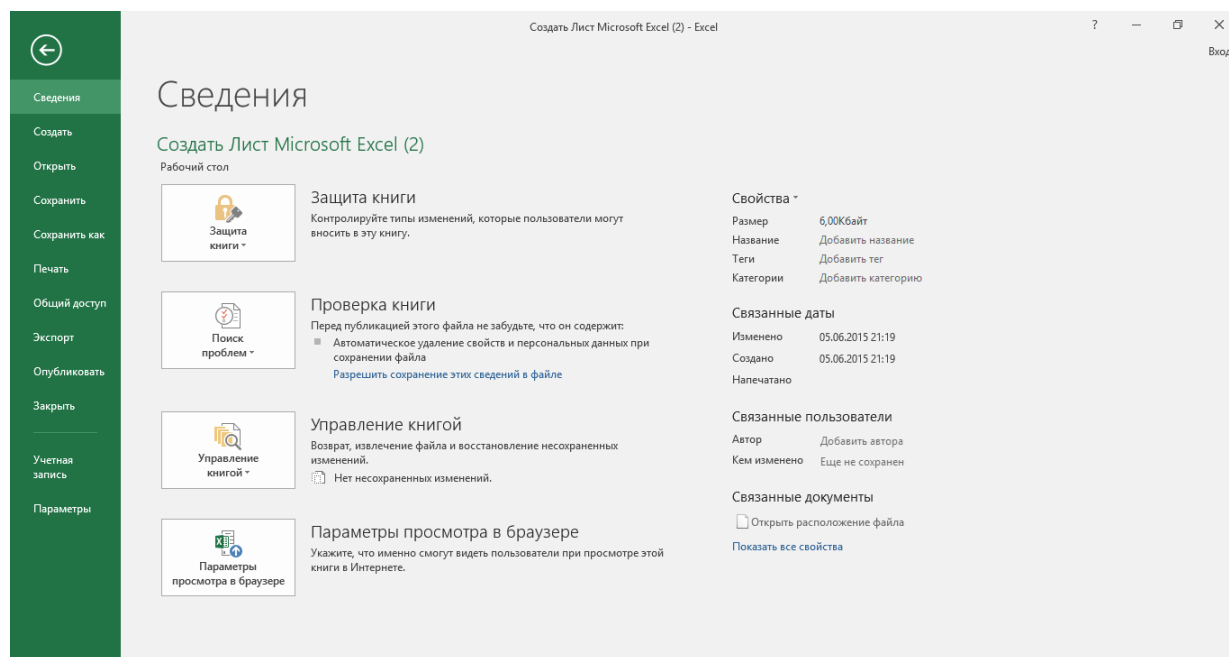
Лента в MS Excel 2016 позволяет быстро найти нужные команды. Они организованы в логические группы, которые находятся на разных вкладках. Каждая вкладка состоит из панелей, где размещены инструменты для работы с электронными таблицами:



Вкладка «Главная» содержит базовые инструменты, которые необходимы на начальном этапе работы с документом. Вкладка «Вставка» нужна для добавления различных объектов в документ. Вкладка «Разметка страницы» используется для настройки параметров страниц. Чтобы получить полный набор инструментов определенной панели, нужно нажать на стрелку в правом нижнем углу этой панели. Это откроет окно с нужными инструментами. Например, при нажатии на стрелку в правом углу панели «Шрифт» появится диалоговое окно для настройки ячеек:



В левом верхнем углу ленты расположена вкладка «Файл». При нажатии на неё открывается меню с основными командами для работы с файлами, появляется список недавно открытых документов и возможность настройки параметров приложения:



По умолчанию программа сохраняет файлы с расширением .xlsx. Однако, чтобы обеспечить совместимость с более старыми версиями электронных таблиц, при сохранении файла следует выбрать соответствующий режим.

Рабочая область в MS Excel представлена в виде сетки, состоящей из строк и столбцов. Всего в программе может быть до 1048576 строк и 16384 столбцов. Столбцы обозначаются буквами латинского алфавита, начиная от A до Z, затем от AA до AZ, BA до BZ и так далее до XFD.

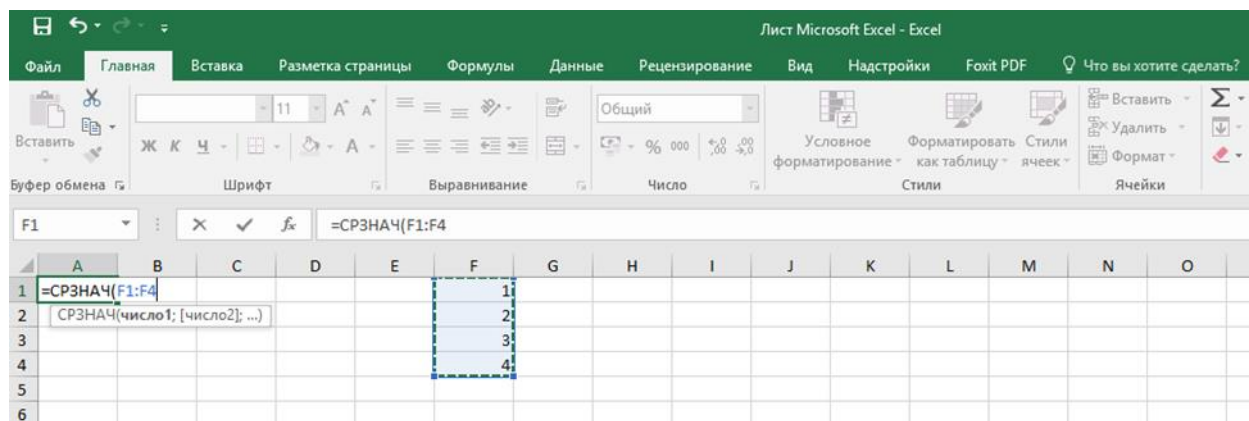
Строки обозначаются номерами. Пересечение строки и столбца образует ячейку, которая имеет свой уникальный адрес. Этот адрес используется в формулах и ссылках, например, B4 или G8. Абсолютные ссылки, такие как \$B\$4 или \$G\$8, не изменяются при копировании формулы. Выделенная ячейка считается активной. Для удобства перемещения по таблице предусмотрены горизонтальная и вертикальная строки прокрутки. Строка ввода используется для просмотра и редактирования содержимого текущей (активной) ячейки.

Задания

1. Запустить программу MS Excel.
2. После открытия окна «Microsoft Excel» просмотреть команды вкладки «Главная».
3. Закрыть раскрытую книгу, используя кнопку «Закрыть».
4. Просмотреть всплывающие подсказки для всех кнопок панелей MS Excel: для вкладки «Главная» – команды панели быстрого доступа (рядом с кнопкой «Файл»), команды кнопки «Файл», панелей: «Шрифт», «Выравнивание», «Число», «Стили», «Ячейки», «Редактирование», имя ячейки, строка формул.
5. Создать таблицу по образцу и сохранить ее под названием «Био-статистика».
6. Ввести названия в соответствующие ячейки:

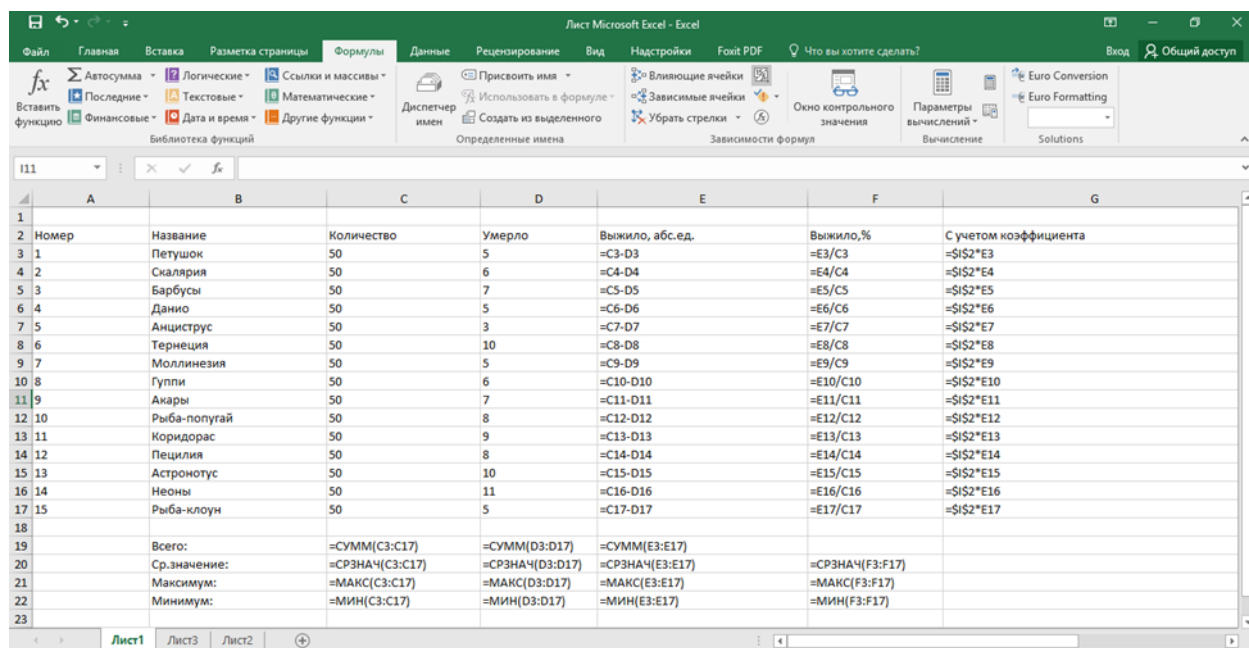
	A	B	C	D	E	F	G	H	I
1									
2	Номер	Название	Количество	Умерло	Выжило, абс.ед.	Выжило, %	С учетом коэффициента	Коэффициент	0,7
3	1	Петушок	50	5	=C3-D3	=E3/C3	=SI\$2*E3		
4	2	Скалярия	50	6					
5	3	Барбусы	50	7					
6	4	Данио	50	5					
7	5	Анциструс	50	3					
8	6	Тернеция	50	10					
9	7	Моллинезия	50	5					
10	8	Гуппи	50	6					
11	9	Акары	50	7					
12	10	Рыба-попугай	50	8					
13	11	Коридорас	50	9					
14	12	Пецилия	50	8					
15	13	Астронотус	50	10					
16	14	Неоны	50	11					
17	15	Рыба-клоун	50	5					
18									
19		Всего:	=СУММ(C3:C17)						
20		Ср.значение:	=СРЗНАЧ(C3:C17)						
21		Максимум:	=МАКС(C3:C17)			=МАКС(F3:F17)			
22		Минимум:	=МИН(C3:C17)			=МИН(F3:F17)			
23									

7. Ввести соответствующие формулы в ячейки, используя следующие варианты:
 - через Σ и вкладку «Формулы»;
 - через ячейку, используя сначала знак «=», затем ввести название и из выпадающего списка выбрать нужную формулу, кликнуть два раза левой кнопкой мышки, выбрать нужный диапазон, затем нажать «Enter»;
 - через строку функций:



8. Научиться просматривать записи формул в строке формул и результаты вычислений в ячейках таблицы. Для этого выполните следующие шаги:

- откройте вкладку «Формулы» в верхнем меню Excel;
- найдите панель зависимости формул, которая обычно находится слева от строки формул;
- нажмите на команду «Показать формулу» в этой панели;
- после этого вы увидите формулу, которая была использована для создания содержимого выбранной ячейки;
- результаты вычислений будут отображаться непосредственно в ячейках таблицы:



9. Модифицировать таблицу. Для этого вставьте столбец «Доза» между столбцами «Количество» и «Умерло». Заполните данными, как

предложено на рисунке ниже. Вставьте строку «Меченосец» между строками «Гуппи» и «Акары». Для выполнения данных операций используйте команды из контекстного меню:

Номер	Название	Количество	Доза	Умерло	Выжило, абс.ед.	Выжило, %	С учетом коэффициента	Коэффициент
1	Петушок	50	10	5	45	90%	31,5	
2	Скалярия	50	10	6	44	88%	30,8	
3	Барбусы	50	10	7	43	86%	30,1	
4	Данио	50	10	5	45	90%	31,5	
5	Анциструс	50	10	3	47	94%	32,9	
6	Тернеция	50	10	10	40	80%	28	
7	Моллинезия	50	10	5	45	90%	31,5	
8	Гуппи	50	10	6	44	88%	30,8	
9	Меченосец	50	10	5	45	90%	31,5	
10	Акары	50	10	7	43	86%	30,1	
11	Рыба-попугай	50	10	8	42	84%	29,4	
12	Коридорас	50	10	9	41	82%	28,7	
13	Пецилия	50	10	8	42	84%	29,4	
14	Астронотус	50	10	10	40	80%	28	
15	Неоны	50	10	11	39	78%	27,3	
16	Рыба-клоун	50	10	5	45	90%	31,5	
Всего:		800		110	690			
Ср. значение:		50		6,875	43,125	86%		
Максимум:		50		11	47	94%		
Минимум:		50		3	39	78%		

10. Скопировать таблицу с листа. Затем вставить таблицу на требуемом листе, расположив её в область ячеек, начинающихся, например, с A28.

11. Модифицировать продублированную таблицу на третьем листе. Для этого удалите столбцы «Количество», «Умерло», пользуясь командами «Удалить» и «Очистить». Отметьте разницу в результатах этих команд.

Контрольные вопросы

1. Что такое программа MS Excel?
2. Как скопировать формулу в MS Excel?
3. Как создать таблицу в MS Excel?
4. Что такое рабочий лист в MS Excel?
5. Как сохранить книгу в MS Excel?

1.2. Методы оформления таблиц

Для оформления таблицы необходимо:

1. Создать на первом рабочем листе таблицу по образцу и сохранить ее для дальнейшей работы:

	A	B	C	D	E	F	G	H
1	Форматирование текста							
2								
3	Левый край							
4		По центру выделения (ячейки C4:G4)						
5	Правый край							
6								
7	Центр							
8					Текст 1	Текст 2	Текст 3	Текст 4
9	1,31313E+17							

2. Скопировать на второй рабочий лист всё содержимое первого рабочего листа.

3. Отформатировать тексты таблицы по образцу:

	A	B	C	D	E	F	G	H
1	Форматирование текста							
2								
3	Левый край							
4			По центру выделения (ячейки C4:G4)					
5	Правый край							
6								
7	Центр							
8					Т е к с т 1	Текст 2	Текст 3	Текст 4
9	1313131313131000,00							

На этом примере научиться выравнивать текст всеми доступными способами. Перед выполнением этого пункта установить для всего рабочего листа стандартную ширину столбцов и высоту строк (для шрифта размером 10 стандартная высота строки составляет 12,75, а ширину строки задает сам пользователь). Использовать вкладку «Главная», панель «Выравнивание».

4. Чтобы изменить внешний вид таблицы, включить режим автоматической установки ширины столбцов. Это позволит увидеть, как таблица будет выглядеть после внесения изменений. Затем подстроить параметры таблицы, включая ширину столбцов и высоту строк, чтобы достичь желаемого вида. Использовать вкладку «Ячейки», кнопку «Формат», чтобы выровнять элементы таблицы.

5. Чтобы оформить тексты в таблице второго листа аналогично образцу на рисунке, воспользоваться режимом форматирования ячеек. Необходимо научиться применять различные варианты шрифтового оформления, такие как **жирный**, подчеркнутый, *курсив*, ***жирный курсив***, ~~перечеркнутый~~, а также верхний и нижний индексы. Для этого перейти на вкладку «Главная», нажать на кнопку «Шрифт», которая находится на панели инструментов. Откроется диалоговое окно, где можно выбрать нужные опции для форматирования текста. После выбора необходимых параметров применить их к соответствующим ячейкам таблицы.

6. Для оформления таблицы использовать подчеркивания.

7. Для раскраски таблиц использовать кнопки «Заливка» и другие опции.

8. Для выделения данных в таблице использовать различные варианты оформления, открыв вкладку «Главная», выбрав панель «Стили», нажав кнопку «Стили ячеек»; а также вкладку «Главная», панель «Ячейки», кнопку «Формат», «Формат ячейки»:

	A	B	C	D	E	F	G	H
1	Форматирование текста							
2								
3	Левый край							
4			По центру выделения (ячейки C4:G4)					
5	Правый край							
6								
7	Центр							
					Т е к с т 1	Текст 2	Текст 3	Текст 4
8								
9	131313131313131000,00							

9. Чтобы ознакомиться со всеми вариантами форматирования чисел, доступными в программе MS Excel, перейти на вкладку «Главная» и найти панель инструментов «Число». Здесь можно выбрать различные форматы чисел, такие как общие, денежные, процентные, научные и другие.

10. Отформатировать таблицу, как приведено на рисунке:

Аквариум							
Номер	Название	Количество	Умерло	Выжило, абс. ед.	Выжило, %	С учетом коэффициента	Коэффициент
1	Петушок	50	5	45	90%	31,5	
2	Скалярия	50	6	44	88%	30,8	
3	Барбусы	50	7	43	86%	30,1	
4	Данио	50	5	45	90%	31,5	
5	Анциструс	50	3	47	94%	32,9	
6	Тернеция	50	10	40	80%	28	
7	Моллинезия	50	5	45	90%	31,5	
8	Гуппи	50	6	44	88%	30,8	
9	Акары	50	7	43	86%	30,1	
10	Рыба-попугай	50	8	42	84%	29,4	
11	Коридорас	50	9	41	82%	28,7	
12	Пещерия	50	8	42	84%	29,4	
13	Астронотус	50	10	40	80%	28	
14	Неоны	50	11	39	78%	27,3	
15	Рыба-клоун	50	5	45	90%	31,5	
Всего:		750	105	645			
Ср. значение:		50	7	43	86%		
Максимум:		50	11	47	94%		
Минимум:		50	3	39	78%		

Контрольные вопросы

1. Как изменить размер ячейки в MS Excel?
2. Как объединить ячейки в MS Excel?
3. Как добавить заливку ячейки в MS Excel?
4. Как добавить текстовый маркер в ячейку в MS Excel?
5. Как добавить заголовок в MS Excel?

1.3. Визуализация данных

Графически табличные данные можно представить в виде диаграмм. Для этого необходимо:

1. Ввести таблицу на первый и второй листы рабочего документа MS Excel с какими-либо данными. Например:

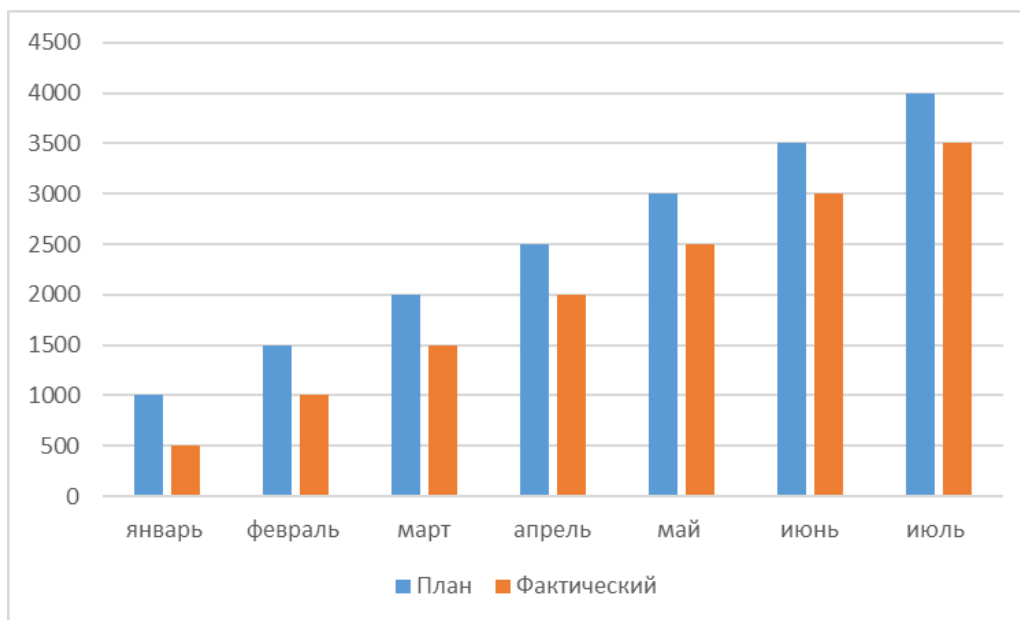
	A	B	C	D	E	F	G	H
1	Рождаемость							
2		январь	февраль	март	апрель	май	июнь	июль
3	План	1000	1500	2000	2500	3000	3500	4000
4	Фактический	500	1000	1500	2000	2500	3000	3500

2. Научиться создавать диаграммы на рабочем листе. Для этого выбрать вкладку «Вставка», открыть панель «Работа с диаграммами». Эта панель содержит вкладки «Конструктор» и «Формат»:

а) во вкладке «Конструктор» можно изменить тип диаграммы и данные, которые она отображает, а также изменить макет, стиль и расположение диаграммы;

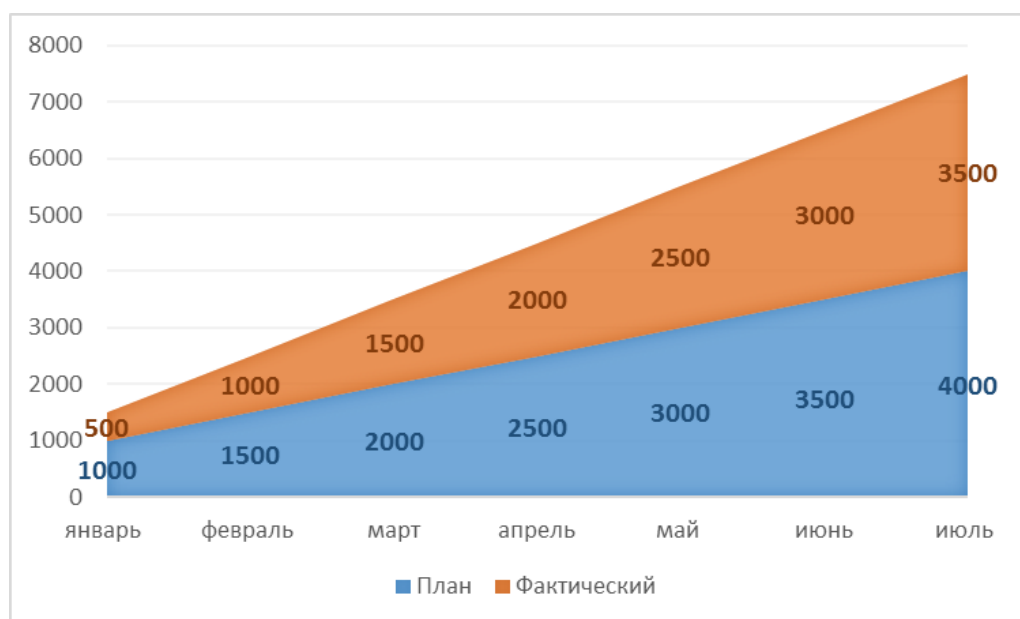
б) во вкладке «Формат» можно изменять стили фигур, упорядочивать объекты и изменять их размеры.

3. Построить на рабочем поле листа *гистограмму*, как показано на рисунке:

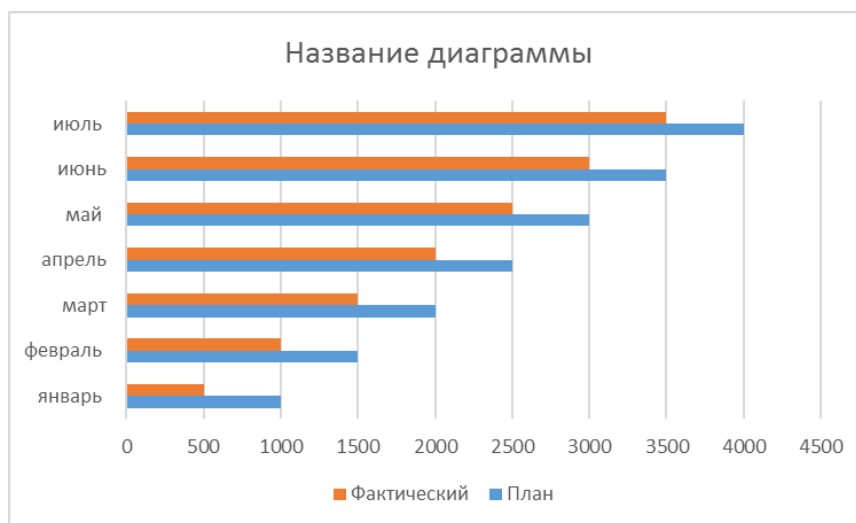


4. Научиться создавать и оформлять диаграммы на отдельных листах. Каждый такой лист должен иметь название, которое соответствует типу диаграммы, размещенной на нём.

5. Построить диаграмму *с областями*, как показано на рисунке:



6. Построить *линейчатую* диаграмму:



7. Построить диаграмму типа *график*.

8. Построить *круговую* диаграмму.

9. Построить *кольцевую* диаграмму.

10. Построить *лепестковую* диаграмму.

11. Построить *точечную* диаграмму (XY).

12. Построить *поверхностную* диаграмму.

13. Построить *пузырьковую* диаграмму.

14. Научиться располагать на одном листе несколько диаграмм.

15. Научиться редактировать объемные диаграммы:



Контрольные вопросы

1. Как создать диаграмму в MS Excel?

2. Как изменить тип диаграммы в MS Excel?

3. Как изменить цвет, шрифт или стиль диаграммы в MS Excel?

1.4. Методы обработки данных

Рассмотрим, каким образом может выполняться обработка данных в таблицах. Для этого необходимо:

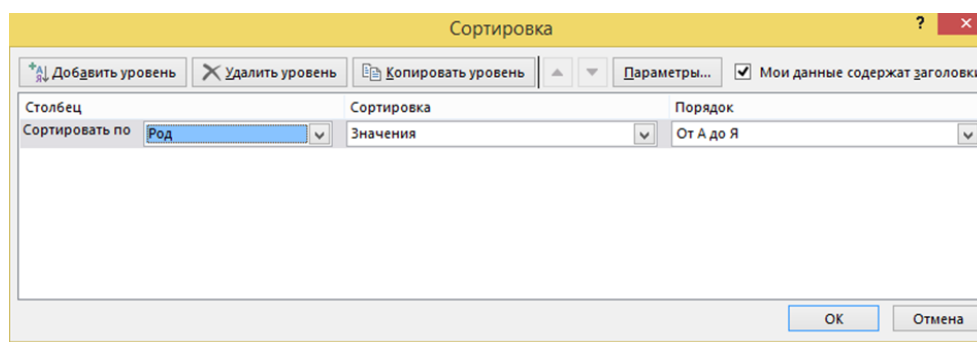
1. Ввести таблицу с какими-либо данными. Например:

	A	B	C	D	E	F	G	H	I
1	№	Название	Класс	Отряд	Семейство	Род	Вид	Обитание	Мак.длина панцыря, см
2	1		Reptilia	Testudines	Dermochelys	Dermochelys	Dermochelys coriacea L.	Море	200
3	2		Reptilia	Testudines	Cheloniidae	Caretta	Caretta caretta L.	Море	100
4	3		Reptilia	Testudines	Chelydridae	Chelydra	Chelydra serpentina	Водоемы	80
5	4		Reptilia	Testudines	Trionychidae	Trionyx	Trionyx sinensis Wieg.	Водоемы	33
6	5		Reptilia	Testudines	Emydidae	Clemmys	Clemmys caspica Gmel	Водоемы	22
7	6		Reptilia	Testudines	Testudinidae	Testudo	Testudo horsfieldi Gray.	Суша	25
8	7		Reptilia	Testudines	Emydidae	Emys	Emys orbicularis L.	Водоемы	20
9	8		Reptilia	Testudines	Testudinidae	Testudo	Testudo graeca L.	Суша	30

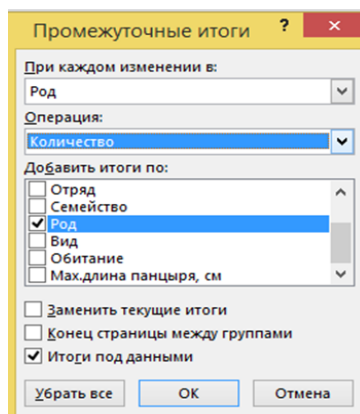
2. Пользуясь вкладкой «Данные», кнопками «Сортировка» и «Промежуточные итоги», провести анализ данных по нижеприведенной схеме, ответив на следующие вопросы:

а) Какой род в таблице повторяется дважды?

– Для этого необходимо выделить всю таблицу и нажать кнопку «Сортировка», в открывшемся окне указать параметры в соответствии с рисунком ниже и нажать кнопку «ОК»:



– Далее нажать кнопку «Промежуточные итоги» на панели «Структура», в открывшемся окне указать параметры в соответствии с рисунком ниже и нажать кнопку «ОК»:



1	2	3	A	B	C	D	E	F	G	H	I
1	№	Название	Класс	Отряд	Семейство	Род	Вид	Обитание	Мак.длина панцыря, см		
2	2		Reptilia	Testudines	Cheloniidae	Caretta	Caretta caretta L.	Море	100		
3					Caretta Количество		1				
4	3		Reptilia	Testudines	Chelydridae	Chelydra	Chelydra serpentina	Водоемы	80		
5					Chelydra Количество		1				
6	5		Reptilia	Testudines	Emydidae	Clemmys	Clemmys caspica Gmel	Водоемы	22		
7					Clemmys Количество		1				
8	1		Reptilia	Testudines	Dermochelys	Dermochelys	Dermochelys coriacea L.	Море	200		
9					Dermochelys Количество		1				
10	7		Reptilia	Testudines	Emydidae	Emys	Emys orbicularis L.	Водоемы	20		
11					Emys Количество		1				
12	6		Reptilia	Testudines	Testudinidae	Testudo	Testudo horsfieldi Gray.	Суша	25		
13	8		Reptilia	Testudines	Testudinidae	Testudo	Testudo graeca L.	Суша	30		
14					Testudo Количество		2				
15	4		Reptilia	Testudines	Trionychidae	Trionyx	Trionyx sinensis Wieg.	Водоемы	33		
16					Trionyx Количество		1				
17					Общее количество		8				

б) У какого вида черепах, обитающих в водоемах, длина панциря наименьшая?

– Для ответа на этот вопрос необходимо нажать на кнопку «Сортировка», в открывшемся окне указать параметры в соответствии с рисунком ниже и нажать кнопку «ОК»:

Сортировка

Добавить уровень Удалить уровень Копировать уровень

Столбец Сортировка Порядок

Сортировать по Обитание Значения От А до Я

Затем по Max.длина панцыря, см Значения По возрастанию

ОК Отмена

– Получим следующую таблицу:

1	№	Название	Класс	Отряд	Семейство	Род	Вид	Обитание	Мак.длина панцыря, см
2	7		Reptilia	Testudines	Emydidae	Emys	Emys orbicularis L.	Водоемы	20
3	5		Reptilia	Testudines	Emydidae	Clemmys	Clemmys caspica Gmel	Водоемы	22
4	4		Reptilia	Testudines	Trionychidae	Trionyx	Trionyx sinensis Wieg.	Водоемы	33
5	3		Reptilia	Testudines	Cheloniidae	Chelydra	Chelydra serpentina	Водоемы	80
6	2		Reptilia	Testudines	Cheloniidae	Caretta	Caretta caretta L.	Море	100
7	1		Reptilia	Testudines	Dermochelys	Dermochelys	Dermochelys coriacea L.	Море	200
8	6		Reptilia	Testudines	Testudinidae	Testudo	Testudo horsfieldi Gray	Суша	25
9	8		Reptilia	Testudines	Testudinidae	Testudo	Testudo graeca L.	Суша	30

– Выделим таблицу и нажмем кнопку «Промежуточные итоги». В «Промежуточных итогах» введем следующие параметры:

Промежуточные итоги

При каждом изменении в:

Обитание

Операция:

Минимум

Добавить итоги по:

☐ Отряд

☐ Семейство

☐ Род

☐ Вид

☐ Обитание

☒ Max.длина панцыря, см

☐ Заменить текущие итоги

☐ Конец страницы между группами

☒ Итоги под данными

Убрать все ОК Отмена

– Получим следующую таблицу:

	A	B	C	D	E	F	G	H	I
1	№	Название	Класс	Отряд	Семейство	Род	Вид	Обитание	Мак.длина панцыря, см
2	7		Reptilia	Testudines	Emydidae	Emys	Emys orbicularis L.	Водоемы	20
3	5		Reptilia	Testudines	Emydidae	Clemmys	Clemmys caspica Gmel	Водоемы	22
4	4		Reptilia	Testudines	Trionychinae	Trionyx	Trionyx sinensis Wiegman	Водоемы	33
5	3		Reptilia	Testudines	Cheloniidae	Chelydra	Chelydra serpentina	Водоемы	80
6								Водоемы Минимум	20
7	2		Reptilia	Testudines	Cheloniidae	Caretta	Caretta caretta L.	Море	100
8	1		Reptilia	Testudines	Dermochelys	Dermochelys	Dermochelys coriacea L.	Море	200
9								Море Минимум	100
10	6		Reptilia	Testudines	Testudinidae	Testudo	Testudo horsfieldi Gray	Суша	25
11	8		Reptilia	Testudines	Testudinidae	Testudo	Testudo graeca L.	Суша	30
12								Суша Минимум	25
13								Общий минимум	20

Задания

1. Ввести следующую таблицу:

	A	B	C	D	E	F
1	№	Фамилия	Вес	Рост	Бег на 100 м, сек	Бег на 3000 м, мин
2	1	Сидров	70,3	169,3	12,5	12
3	2	Иванов	73,6	264,3	11,7	11,5
4	3	Пугачев	105,9	158,9	11	14,7
5	4	Лебедь	122,6	131,8	17,5	16,5
6	5	Иванов	63,7	266,1	12	10,5
7	6	Яковлев	134,0	242,3	20,5	18
8	7	Абрамов	147,5	136,5	25,4	25
9	8	Мировной	92,7	165,5	13	14,5
10	9	Холяк	60,1	168,2	12	11
11	10	Курдов	114,9	214,4	15	17
12	11	Емельян	115,6	209,3	16,7	17,5
13	12	Ильин	113,0	293,3	15	18,4
14	13	Имак	116,2	201,0	16	19,2

2. Распределить участников в алфавитном порядке.

3. Распределить участников по росту и весу.

4. Распределить участников по скорости забегов на дистанции.

Контрольные вопросы

1. Как отсортировать данные в MS Excel по алфавиту?

2. Как отсортировать данные в MS Excel по возрастанию или убыванию?

1.5. Создание связей между таблицами и консолидация данных

Установление связей между таблицами проиллюстрируем на следующем примере:

1. Создадим три таблицы, содержащие сведения о рождаемости в разных областях по годам. Для каждого года на отдельном листе в документе создается собственная таблица:

	А	В	С
1	Область	Рождаем	Умерших
2	Белгородская область	16790	22422
3	Брянская область	14259	23111
4	Владимирская область	15569	27119
5	Воронежская область	22361	40316
6	Ивановская область	11138	20769
7	Костромская область	10491	17374

Лист1 (2008)

	А	В	С
1	Область	Рождаем	Умерших
2	Белгородская область	16845	22011
3	Брянская область	14401	21964
4	Владимирская область	15551	26407
5	Воронежская область	23559	38581
6	Ивановская область	11390	19887
7	Костромская область	10537	16762

Лист 2 (2009)

	А	В	С
1	Область	Рождаемость	Умерших
2	Белгородская область	16635	22027
3	Брянская область	13727	21775
4	Владимирская область	15542	26082
5	Воронежская область	23804	39767
6	Ивановская область	11078	19604
7	Костромская область	11131	16730

Лист 3 (2010)

2. Для установления связи между таблицами «2008», «2009» и «2010» необходимо выполнить следующие шаги:

а) скопировать диапазон ячеек А1:А7 таблицы «2008» в буфер обмена;

б) перейди в таблицу «2009» и вставить скопированный диапазон, используя команды «Главная» – «Вставить» – «Специальная вставка» – «Вставить связь». Это создаст связь между таблицами «2008» и «2009»;

в) повторить шаг 2 для таблицы «2010», чтобы установить связь между таблицами «2008» и «2010».

3. Чтобы проверить, как работают ссылки между таблицами «2008» и «2009», необходимо выполнить следующие шаги:

а) открыть обе таблицы в MS Excel;

б) выделите ячейку в таблице «2008», которая связана с таблицей «2009»;

в) посмотреть на строку формул – там должна быть ссылка на ячейку в таблице «2009»;

г) изменить содержимое ячейки А7 в таблице «2008»;

д) вернуться к таблице «2009» и проверить, изменилась ли соответствующая ячейка.

Это позволит убедиться, что связь между таблицами работает корректно и что изменение данных в одной таблице приводит к обновлению связанных ячеек в другой таблице.

4. На листе «2010» посчитаем сумму за три года, родившихся и умерших, используя формулы, приведенные на рисунке:

	А	В	С
1	= '2008'!A1	Рождаемость	Умерших
2	= '2008'!A2	16635	22027
3	= '2008'!A3	13727	21775
4	= '2008'!A4	15542	26082
5	= '2008'!A5	23804	39767
6	= '2008'!A6	11078	19604
7	= '2008'!A7	11131	16730
8			
9		Рождаемость	Умерших
10	Итого за три года	=СУММ(B2:B7)+СУММ('2009'!B2:B7)+СУММ('2008'!B2:B7)	=СУММ(C2:C7)+СУММ('2009'!C2:C7)+СУММ('2008'!C2:C7)

5. Вычислим среднее по рождаемости и смертности за три года.

6. Вычислим дисперсию по рождаемости и смертности за три года.

7. Вычислим стандартное отклонение по рождаемости и смертности за три года.

Консолидация электронных таблиц или их частей представляет собой процесс объединения нескольких однотипных таблиц в одну. Этот метод позволяет упростить работу с данными, особенно когда их много и они разрознены. Консолидация данных – это процедура получения итоговых значений для данных, которые расположены в разных частях таблицы. Диапазоны ячеек могут находиться на различных листах и даже в разных книгах. Таким образом, консолидация помогает собрать всю необходимую информацию в одном месте, что значительно облегчает ее анализ и обработку.

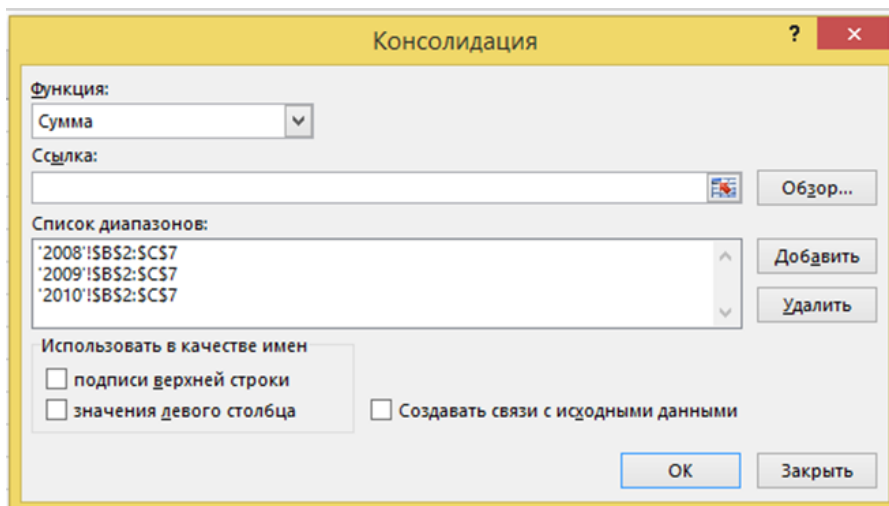
Для выполнения консолидации данных можно использовать прошлые таблицы по рождаемости и смертности. Необходимо:

1. Создать лист и назвать его «Консолидация данных».

2. Подписать столбцы, а также строки, как в прошлых таблицах по рождаемости и смертности.

3. Для вызова диалогового окна «Консолидация» необходимо перейти в меню «Данные» и выбрать пункт «Консолидация». Далее, в поле

«Ссылка» следует указать адреса консолидируемых областей. Чтобы выделить области на листах, нужно воспользоваться пунктом «Обзор», а затем добавить их в «Список диапазонов». Важно учесть, что список должен состоять из трех записей, как показано на рисунке:



4. В итоге получим следующую таблицу на отдельном рабочем листе:

B2					33480
	A	B	C		
1	Область	Рождаемость	Умерших		
2	Белгородская область	33480	66460		
3	Брянская область	28128	66850		
4	Владимирская область	46662	79608		
5	Воронежская область	69724	118664		
6	Ивановская область	33606	60260		
7	Костромская область	21622	50866		

Контрольные вопросы

1. Как создать связь между таблицами в MS Excel?
2. Как обновить связанные данные в MS Excel?
3. Что такое консолидация данных в MS Excel?
4. Как выполнить консолидацию данных в MS Excel?
5. Как выбрать диапазоны для консолидации в MS Excel?
6. Как добавить новые строки или столбцы в консолидированные данные в MS Excel?

1.6. Знакомство с программой для статистического анализа данных JASP

JASP означает «Программа удивительной статистики Джеффриса» в честь пионера байесовского вывода сэра Гарольда Джеффриса. Это бесплатный многоплатформенный статистический пакет с открытым исходным кодом, разработанный и постоянно обновляемый группой исследователей из Амстердамского университета. Они стремились разработать бесплатную программу с открытым исходным кодом, которая включала бы как стандартные, так и более продвинутое статистические методы, уделяя особое внимание обеспечению простого, интуитивно понятного пользовательского интерфейса.

В отличие от многих статистических пакетов, JASP предоставляет простой интерфейс перетаскивания, удобные меню доступа, возможность интуитивно понятного анализа с вычислениями в реальном времени и отображением всех результатов. Все таблицы и графики представлены в формате APA и могут быть скопированы напрямую и/или сохранены самостоятельно. Таблицы также можно экспортировать из JASP в формат LaTeX.

JASP можно бесплатно загрузить с сайта <https://jasp-stats.org/>, программа доступна для Windows, Mac OS X и Linux. Можно также загрузить предустановленную версию Windows, которая будет запускаться непосредственно с USB-накопителя или внешнего жесткого диска без необходимости устанавливать ее локально. Установщик WIX для Windows позволяет выбрать путь для установки JASP, однако в некоторых учреждениях это может быть заблокировано правами локального администратора.

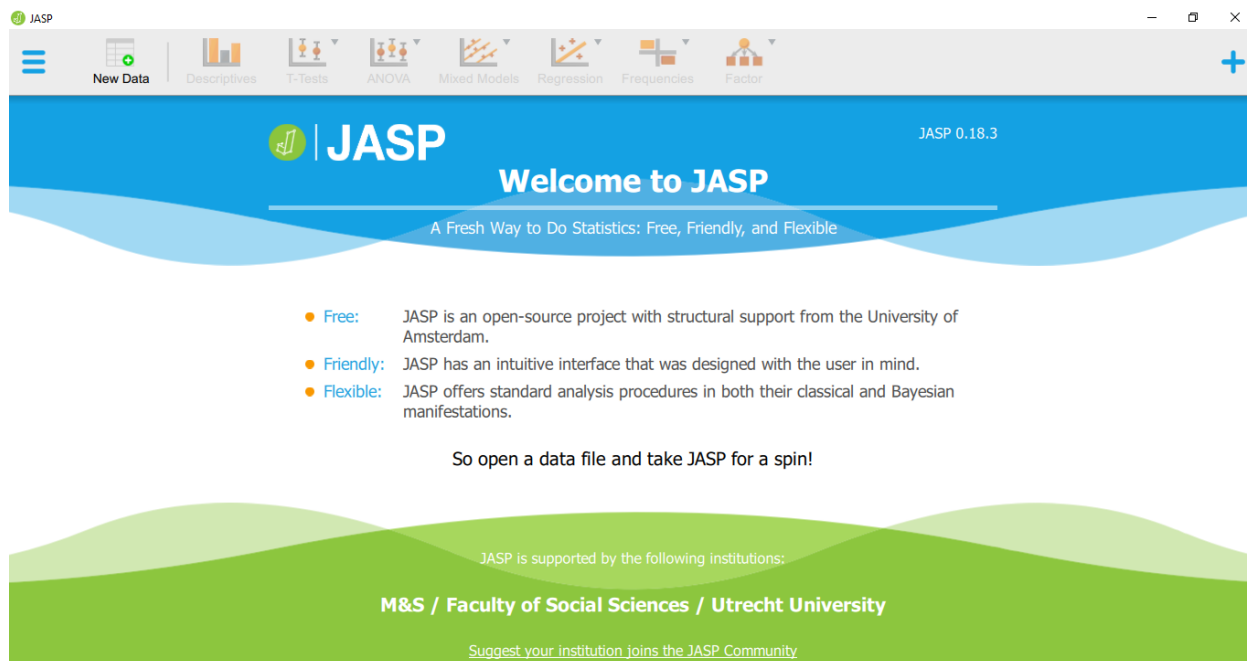
Программа также содержит библиотеку данных с первоначальной коллекцией из более чем 50 наборов данных из книги Энди Филдса «Обнаружение статистики с использованием статистики IBM SPSS» и «Введение в практику статистики» Мура, Маккейба и Крейга.

С мая 2018 г. JASP также можно запускать непосредственно в браузере через rollApp™ без необходимости устанавливать его на свой компьютер.

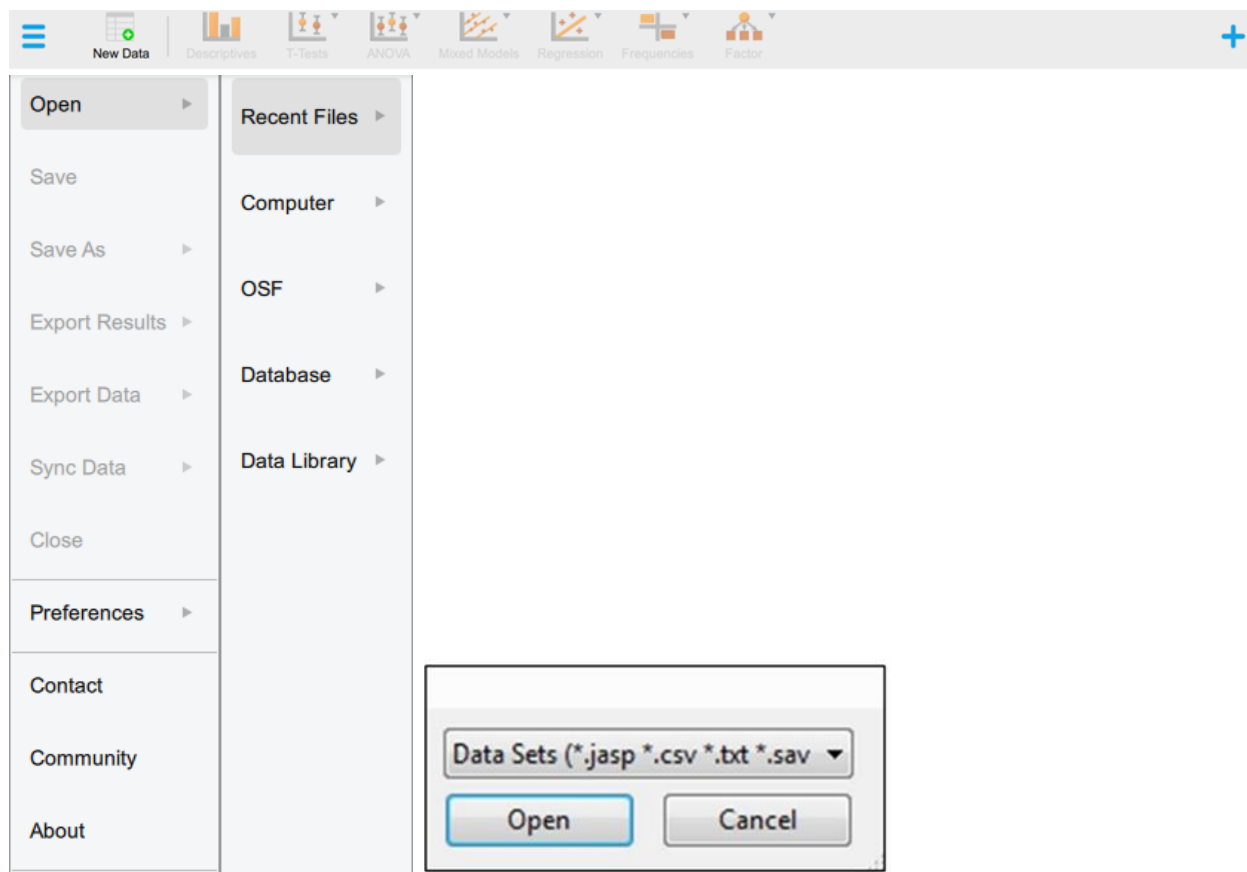
На сайте JASP (<https://www.rollapp.com/app/jasp>) регулярно появляются обновления, а также полезные видеоролики и сообщения в блогах.

Использование среды JASP

1. Откроем программу JASP:



2. В главное меню можно попасть, нажав на значок в верхнем левом углу:



JASP имеет собственный формат jasp, но может открывать множество различных форматов наборов данных, таких как:

- .csv (значения, разделенные запятыми) – можно сохранить в MS Excel;
- .txt (обычный текст) – можно сохранить в MS Excel;
- .tsv (значения, разделенные табуляцией) – можно сохранить в MS Excel;
- .sav (файл данных IBM SPSS);
- .ods (таблица открытого документа).

Можно открывать последние файлы, просматривать файлы на своем компьютере, получать доступ к Open Science Framework (OSF) или открывать широкий спектр примеров, включенных в библиотеку данных в JASP.

3. Создадим в MS Excel на первом листе таблицу, имеющую следующий вид:

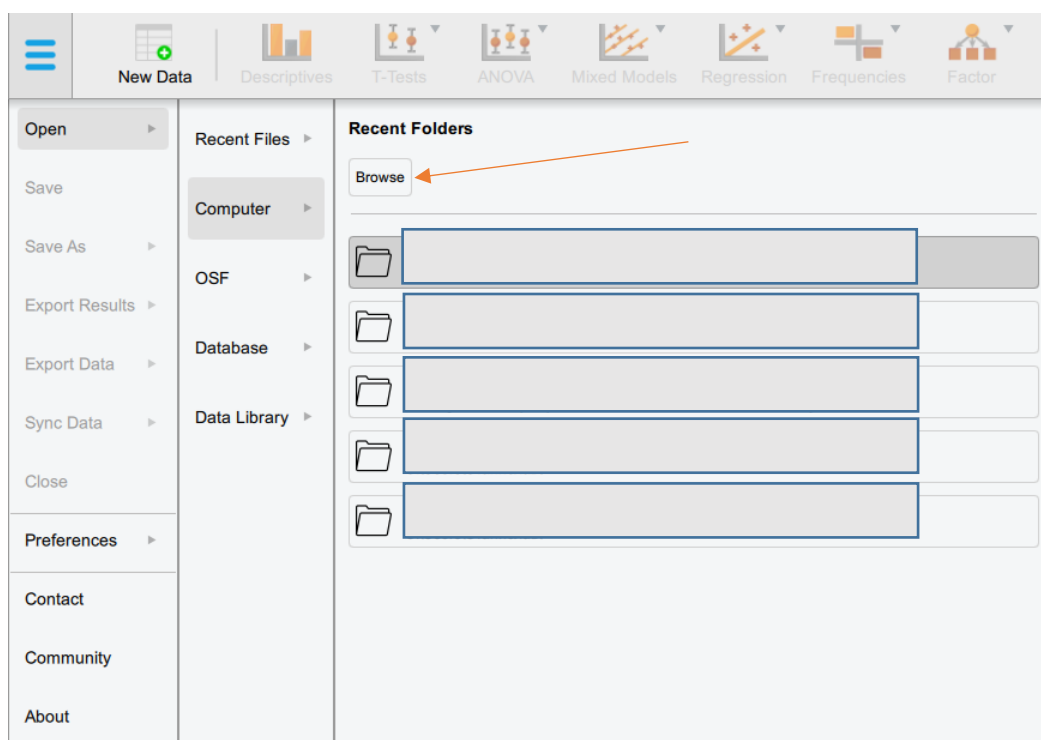
Рост, см	Вес, кг
150	50
160	60
170	70
180	80
190	90
200	100
210	110

	А	В
1	Рост, см	Вес, кг
2	150	50
3	160	60
4	170	70
5	180	80
6	190	90
7	200	100
8	210	110

4. В MS Excel через «Файл» и «Сохранить как» сохраним полученную таблицу в формате OpenDocument в своей рабочей папке.

Имя файла:	Рост и вес Microsoft Excel	▼
Тип файла:	Электронная таблица OpenDocument	▼

5. Откроем файл в формате OpenDocument программе JASP:



6. Появится следующая таблица:

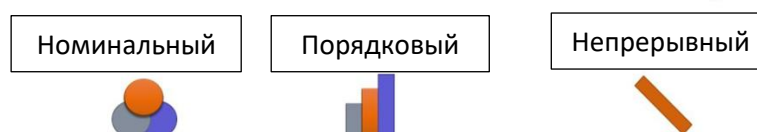
The screenshot shows the JASP 'Edit Data' window. The table has 7 rows and 2 columns: 'Рост, см' (Height, cm) and 'Вес, кг' (Weight, kg). The data is as follows:

	Рост, см	Вес, кг
1	150	50
2	160	60
3	170	70
4	180	80
5	190	90
6	200	100
7	210	110

Все файлы должны иметь метку заголовка в первой строке. После загрузки набор данных появится в окне.

Для больших наборов данных имеется значок руки, который позволяет легко прокручивать данные.

7. При импорте JASP делает наилучшее предположение о назначении данных различным типам переменных:



Если JASP неправильно определил тип данных, нужно просто щелкнуть соответствующий значок переменных данных в заголовке столбца, чтобы изменить его на правильный формат.

8. В нашем случае тип переменной был определен, поэтому исправим это, щелкнув по названию столбца переменной два раза левой кнопкой мышки:

The screenshot shows the JASP variable editor for a variable named 'Рост, см' (Height, cm). The 'Label editor' tab is active, displaying a table of values and labels. The 'Column type' dropdown is currently set to 'Ordinal' and is being changed to 'Scale'.

Filter	Value	Label
✓	150	150
✓	160	160
✓	170	170
✓	180	180
✓	190	190

Variable settings:

- Name: Рост, см
- Long name: Рост, см
- Column type: Scale (selected)
- Computed type: Not computed
- Description: ...

9. Аналогичную процедуру проделаем и со второй переменной:

The screenshot shows the JASP variable editor for a variable named 'Вес, кг' (Weight, kg). The 'Column type' dropdown is currently set to 'Ordinal' and is being changed to 'Scale'.

Variable settings:

- Name: Вес, кг
- Long name: Вес, кг
- Column type: Scale (selected)
- Computed type: Not computed
- Description: ...

10. Добавим номинальную переменную и назовем её «Участники», нажав на значок плюса справа. Не забудем сразу выбрать тип переменной «Номинальный»:

	 Рост, см	 Вес, кг	+
1	150	50	
2	160	60	
3	170	70	
4	180	80	
5	190	90	
6	200	100	
7	210	110	

Create Computed Column

Name:




 Scale
  Ordinal
  **Nominal**
 Text





11. Получим следующий вид таблицы:

	 Рост, см	 Вес, кг	 f_x Участники	+
1	150	50		
2	160	60		
3	170	70		
4	180	80		
5	190	90		
6	200	100		
7	210	110		

12. Поменяем вычислительный тип, дважды щелкнув по столбцу переменной и перейдя в раздел редактирования переменных:

The image shows two identical panels of a variable editor. Each panel contains the following fields:

- Name:** Участники
- Long name:** Участники
- Column type:** Nominal
- Description:** ...
- Computed type:** Computed with drag-and-drop (top) / Not computed (bottom)

13. Впишем имена в ячейки столбца «Участники», дважды щелкнув левой кнопкой мыши по ячейке:

	Рост, см	Вес, кг	Участники
1	150	50	Мария
2	160	60	Людмила
3	170	70	Иван
4	180	80	Степан
5	190	90	Александр
6	200	100	Петр
7	210	110	Владимир

14. Столбец «Участники» целесообразно расположить перед столбцом «Рост, см». Для этого щелкнем один раз по первой ячейке столбца «Рост, см» правой кнопкой мыши и выберем пункт из выпадающего списка «Insert column before»:

	Рост, см	Вес, кг	Участники	
1	150	50	Мария	
2	160			
3	170			
4	180			
5	190			
6	200			
7	210			
8				
9				
10				
11				
12				
13				

15. Появится новый столбец:

	Column 1	Рост, см	Вес, кг	Участники
1		150	50	Мария
2		160	60	Людмила
3		170	70	Иван
4		180	80	Степан
5		190	90	Александр
6		200	100	Петр
7		210	110	Владимир

16. Вернемся к столбцу «Участники». Наждем на название столбца левой кнопкой мышки один раз. Столбец будет выделен. Наждем правую кнопку мышки и выберем из выпадающего списка «Сору»:

	Column 1	Рост, см	Вес, кг	Участники			
1		150	50	Мария			
2		160	60	Людмила			
3		170	70	Иван			
4		180	80	Степан			
5		190	90	Александр			
6		200	100	Петр			
7		210	110	Владимир			
8							
9							
10							
11							
12							
13							
14							

Select All Ctrl+A
Cut Ctrl+X
Copy Ctrl+C
Paste Ctrl+V
Clear cells Del
Undo: Insert column 'Рост, см' Ctrl+Z
Redo: Ctrl+Shift+Z
Select column
Insert column before
Insert column after
Delete column
Select row
Insert row above
Insert row below
Delete row

17. Вернемся к первому столбцу. Наждем на название столбца левой кнопкой мышки. Затем наждем правую кнопку мышки и из выпадающего списка выберем «Paste»:

	Column 1			
1				
2				
3				
4				
5				
6				
7				
8				

Select All Ctrl+A
Cut Ctrl+X
Copy Ctrl+C
Paste Ctrl+V
Clear cells Del
Undo: Insert column 'Рост, см' Ctrl+Z
Redo: Copy column 'Участники' into column number '0' Ctrl+Shift+Z
Insert column before
Insert column after
Delete column

18. Получим таблицу следующего вида:

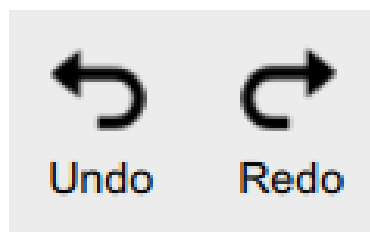
	Участники 2	Рост, см	Вес, кг	Участники
1	Мария	150	50	Мария
2	Людмила	160	60	Людмила
3	Иван	170	70	Иван
4	Степан	180	80	Степан
5	Александр	190	90	Александр
6	Петр	200	100	Петр
7	Владимир	210	110	Владимир

19. Удалим четвертый столбец, выбрав «Delete column»:

	Участники 2	Рост, см	Вес, кг	Участники
1	Мария	150	50	Мария
2	Людмила	160	60	Людмила
3	Иван	170	70	Иван
4	Степан	180	80	Степан
5	Александр	190	90	Александр
6	Петр	200	100	Петр
7	Владимир	210	110	Владимир
8				

- Select All Ctrl+A
- Cut Ctrl+X
- Copy Ctrl+C
- Paste Ctrl+V
- Clear cells Del
- Undo: Copy column 'Участники' into column number '0' Ctrl+Z
- Redo: Ctrl+Shift+Z
- Insert column before
- Insert column after
- Delete column

20. Мы также могли изначально использовать вырезку столбца, выбрав пункт «Cut». Для этого воспользуемся отменой действий и откажемся до предыдущего этапа:

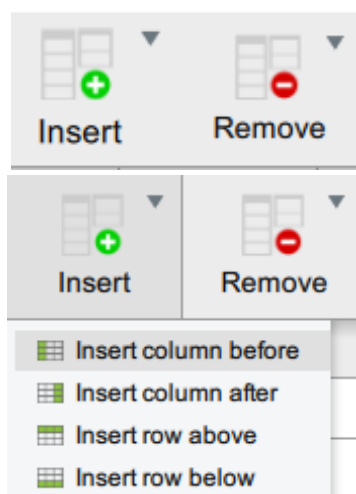


	Рост, см	Вес, кг	Участники
1	150	50	Мария
2	160	60	Людмила
3	170	70	Иван
4	180	80	Степан
5	190	90	Александр
6	200	100	Петр
7	210	110	Владимир

21. Выберем столбец «Участники», а из выпадающего списка выберем пункт «Cut»:

	Рост, см	Вес, кг	Участники		
1	150	50	Мария		
2	160	60	Людмила		
3	170	70	Иван		
4	180	80	Степан		
5	190	90	Александр		
6	200	100	Петр		
7	210	110	Владимир		
8					
9					
10					
11					
12					
13					
14					
15					

22. Выберем название столбца «Рост, см», но теперь воспользуемся верхней лентой меню, а именно пунктом «Insert», и выберем из выпадающего списка «Insert column before»:



23. Создана таблица следующего вида:

	Column 1	Рост, см	Вес, кг
1		150	50
2		160	60
3		170	70
4		180	80
5		190	90
6		200	100
7		210	110

24. Остается только вставить вырезанный столбец, который сейчас находится в буфере обмена. Получим таблицу следующего вида:

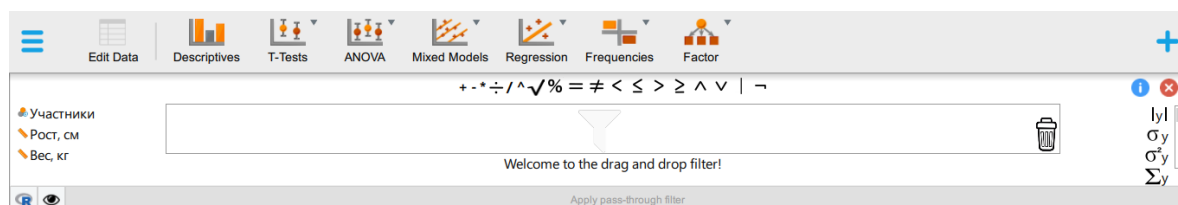
	Участники	Рост, см	Вес, кг
1	Мария	150	50
2	Людмила	160	60
3	Иван	170	70
4	Степан	180	80
5	Александр	190	90
6	Петр	200	100
7	Владимир	210	110

25. Аналогичным способом можно удалять, копировать, вставлять строки и т. д.

26. Рассмотрим функцию фильтрации данных. Нажмем на значок фильтра в левом верхнем углу под значком перехода в главное меню:

	Участники	Рост, см	Вес, кг
1	Мария	150	50
2	Людмила	160	60
3	Иван	170	70
4	Степан	180	80
5	Александр	190	90
6	Петр	200	100
7	Владимир	210	110

27. Откроется раздел фильтрации данных:



28. Для примера отберем всех участников, чей рост ≥ 180 см. Применим данное условие. Мы видим, что 1, 2, 3-я строки, данные которых не подчиняются условию, стали неактивными. Полное удаление условия производится двойным щелчком левой кнопкой мышки по значку корзины.

Удаление неподходящих переменных и вычислительных операций из формулы происходит путем щелчка правой кнопкой мышки либо перетаскиванием в корзину:

Участники
 Рост, см
 Вес, кг

Рост, см ≥ 180

	Участники	Рост, см	Вес, кг
1	Мария	150	50
2	Людмила	160	60
3	Иван	170	70
4	Степан	180	80
5	Александр	190	90
6	Петр	200	100
7	Владимир	210	110

29. Усложним задачу. Отберем участников по следующему критерию: рост ≥ 180 см и вес ≥ 90 кг:

	Участники	Рост, см	Вес, кг		+
1	Мария	150	50		
2	Людмила	160	60		
3	Иван	170	70		
4	Степан	180	80		
5	Александр	190	90		
6	Петр	200	100		
7	Владимир	210	110		

30. Снова усложним нашу задачу и загрузим в JASP таблицу с большим количеством данных:

Номер	Фамилия	Рост, см	Вес, кг	Пол	Возраст, годы	Занятие спортом
1	Иванов	165	90	М	22	нет
2	Петвров	175	75	М	25	да
3	Сидоров	182	85	М	21	да
4	Губкин	190	88	М	28	нет
5	Самойлов	190	95	М	31	да
6	Тумина	155	75	Ж	30	нет
7	Минакина	162	56	Ж	24	да
8	Голубок	172	74	М	23	да
9	Гнедая	165	59	Ж	24	да
10	Гормаш	170	110	М	28	нет

11	Кумиля	178	65	Ж	25	нет
12	Боб	199	105	М	31	да
13	Брусника	168	61	Ж	24	нет
14	Морозова	198	95	М	21	да
15	Мороз	188	84	М	27	нет
16	Добрынина	165	60	Ж	19	нет
17	Дрозд	177	64	М	24	нет
18	Дроздова	172	63	Ж	27	да
19	Мулякин	187	82	М	31	нет
20	Лось	201	110	М	29	нет
21	Лососева	177	67	Ж	22	да
22	Стебельков	158	54	М	24	нет
23	Носатов	174	86	М	24	да
24	Пуля	158	52	Ж	32	нет
25	Степанов	178	74	М	24	да
26	Степанова	157	49	Ж	28	да
27	Гвоздь	176	92	М	29	нет
28	Грозный	220	140	М	29	да
29	Ищейкин	198	120	М	22	да
30	Жмайлов	178	68	М	21	нет
31	Жмайлова	172	70	Ж	25	нет
32	Чугунков	155	56	Ж	27	нет
33	Стальная	170	60	Ж	19	да
34	Могучий	221	150	М	30	да
35	Добрый	182	82	М	27	нет
36	Сульянов	185	84	М	29	нет
37	Симикина	166	90	Ж	22	нет
38	Князева	158	98	Ж	25	нет
39	Грязева	172	62	Ж	26	да
40	Сарматова	166	59	Ж	32	да

 Номер	 Фамилия	 Рост, см	 Вес, кг	 Пол	 Возраст, годы	 Занятие спортом
1	Иванов	165	90	М	22	нет
2	Петров	175	75	М	25	да
3	Сидоров	182	85	М	21	да
4	Губкин	190	88	М	28	нет
5	Самойлов	190	95	М	31	да
6	Тумина	155	75	Ж	30	нет
7	Минакина	162	56	Ж	24	да
8	Голубок	172	74	М	23	да
9	Гнедая	165	59	Ж	24	да
10	Гормаш	170	110	М	28	нет
11	Кумиля	178	65	Ж	25	нет

Как видно, с увеличением размера и набора данных визуально становится сложно ориентироваться. Продолжим освоение фильтрации данных.

31. Предположим, мы хотим исключить из анализа мужчин. Для этого щелкнем на название столбца «Пол», перейдем в режим редактирования переменных и снимем галочку с «М» в столбце «Filter»:

Name: Long name:

Column type: Description:

Computed type:

Label editor Missing Values

Filter	Value	Label
☒	Ж	Ж
☐	М	М

32. В результате получим таблицу следующего вида:

	Номер	Фамилия	Рост, см	Вес, кг	Пол	Возраст, годы	Занятие спортом
1	1	Иванов	165	90	М	22	нет
2	2	Петров	175	75	М	25	да
3	3	Сидоров	182	85	М	21	да
4	4	Губкин	190	88	М	28	нет
5	5	Самойлов	190	95	М	31	да
6	6	Тумина	155	75	Ж	30	нет
7	7	Минакина	162	56	Ж	24	да
8	8	Голубок	172	74	М	23	да
9	9	Гнедая	165	59	Ж	24	да
10	10	Гормаш	170	110	М	28	нет
11	11	Кумиля	178	65	Ж	25	нет
12	12	Боб	199	105	М	31	да
13	13	Брусника	168	61	Ж	24	нет
14	14	Морозова	198	95	М	21	да
15	15	Мороз	188	84	М	27	нет
16	16	Добрынина	165	60	Ж	19	нет
17	17	Дрозд	177	64	М	24	нет
18	18	Дроздова	172	63	Ж	27	да

Для анализа остались только женщины. Теперь отменим это действие и повторим аналогичное для мужчин. Затем вернемся к исходному варианту.

33. Как видно, в нашей выборке возраст варьируется. Для примера определим границы возраста от 22 до 28 лет. Перейдем в режим фильтрации данных и зададим следующие условия:

Номер

Фамилия

Рост, см

Вес, кг

Возраст, годы ≤ 28

Возраст, годы ≥ 22

34. В результате получим таблицу следующего вида:

	Номер	Фамилия	Рост, см	Вес, кг	Пол	Возраст, годы	Занятие спортом
1	1	Иванов	165	90	М	22	нет
2	2	Петров	175	75	М	25	да
3	3	Сидоров	182	85	М	21	да
4	4	Губкин	190	88	М	28	нет
5	5	Самойлов	190	95	М	31	да
6	6	Тумина	155	75	Ж	30	нет
7	7	Минакина	162	56	Ж	24	да
8	8	Голубок	172	74	М	23	да
9	9	Гнедая	165	59	Ж	24	да
10	10	Гормаш	170	110	М	28	нет
11	11	Кумиля	178	65	Ж	25	нет
12	12	Боб	199	105	М	31	да

После этого вернем таблицу в исходное состояние, сняв фильтрацию данных.

35. Рассчитаем индекс массы тела (ИМТ) для всех участников по формуле:

$$\text{ИМТ} = \frac{\text{вес (кг)}}{\text{рост}^2 \text{ (м)}}.$$

Для этого создадим столбец «ИМТ», определим тип вычисления «Computed with drag-and-drop», далее введем формулу, представленную на рисунке:

Name: ИМТLong name: ИМТ

Column type: ScaleDescription: ..

Computed type: Computed with drag-and-drop

Computed column definitionMissing Values

Номер

Фамилия

Рост, см

Вес, кг

Вес, кг / (Рост, см / 100 ^ 2)

Compute column

Для ввода формулы перетаскиваем переменную «Вес» в окно для вычисления, затем добавляем значок деления /, далее необходимо добавить значок возведения в степень ^, затем перенести значок дроби, далее перенести переменную «Рост» в числитель, в знаменателе указать 100 и возвести в квадрат, затем выполнить вычисления.

36. В столбце «ИМТ» появятся вычисления:

№	ИМТ	Фамилия	Рост, см	Вес, кг	Пол	Возраст, годы	Занятие спортом	ИМТ
5		Самойлов	190	95	М	31	да	26.31578947
6		Тумина	155	75	Ж	30	нет	31.21748179
7		Минакина	162	56	Ж	24	да	21.33821064
8		Голубок	172	74	М	23	да	25.01352082
9		Гнедая	165	59	Ж	24	да	21.67125803
10		Гормаш	170	110	М	28	нет	38.06228374
11		Кумила	178	65	Ж	25	нет	20.51508648
12		Боб	199	105	М	31	да	26.51448196
13		Брусника	168	61	Ж	24	нет	21.61281179
14		Морозова	198	95	М	21	да	24.2322212
15		Мороз	188	84	М	27	нет	23.76641014

37. Создадим также столбец «Рост – Вес» и произведем расчет по следующей формуле:

Name: Рост-Вес Long name: Рост-Вес

Column type: Scale Description: ...

Computed type: Computed with drag-and-drop

Computed column definition Missing Values

Рост, см - Вес, кг

Computed columns code applied

38. Получим таблицу следующего вида:

№	ИМТ	Фамилия	Рост, см	Вес, кг	Пол	Возраст, годы	Занятие спортом	ИМТ	Рост-Вес
1		Иванов	165	90	М	22	нет	33.05785124	75
2		Петров	175	75	М	25	да	24.48979592	100
3		Сидоров	182	85	М	21	да	25.66115203	97
4		Губин	190	88	М	28	нет	24.3767313	102
5		Самойлов	190	95	М	31	да	26.31578947	95
6		Тумина	155	75	Ж	30	нет	31.21748179	80
7		Минакина	162	56	Ж	24	да	21.33821064	106
8		Голубок	172	74	М	23	да	25.01352082	98
9		Гнедая	165	59	Ж	24	да	21.67125803	106
10		Гормаш	170	110	М	28	нет	38.06228374	60
11		Кумила	178	65	Ж	25	нет	20.51508648	113

39. Рассчитаем индекс массы тела по формуле Брока: рост (см) минус 100. Добавим соответствующий столбец и произведем вычисления:

Name: Рост-100 Long name: Рост-100

Column type: Scale Description: ...

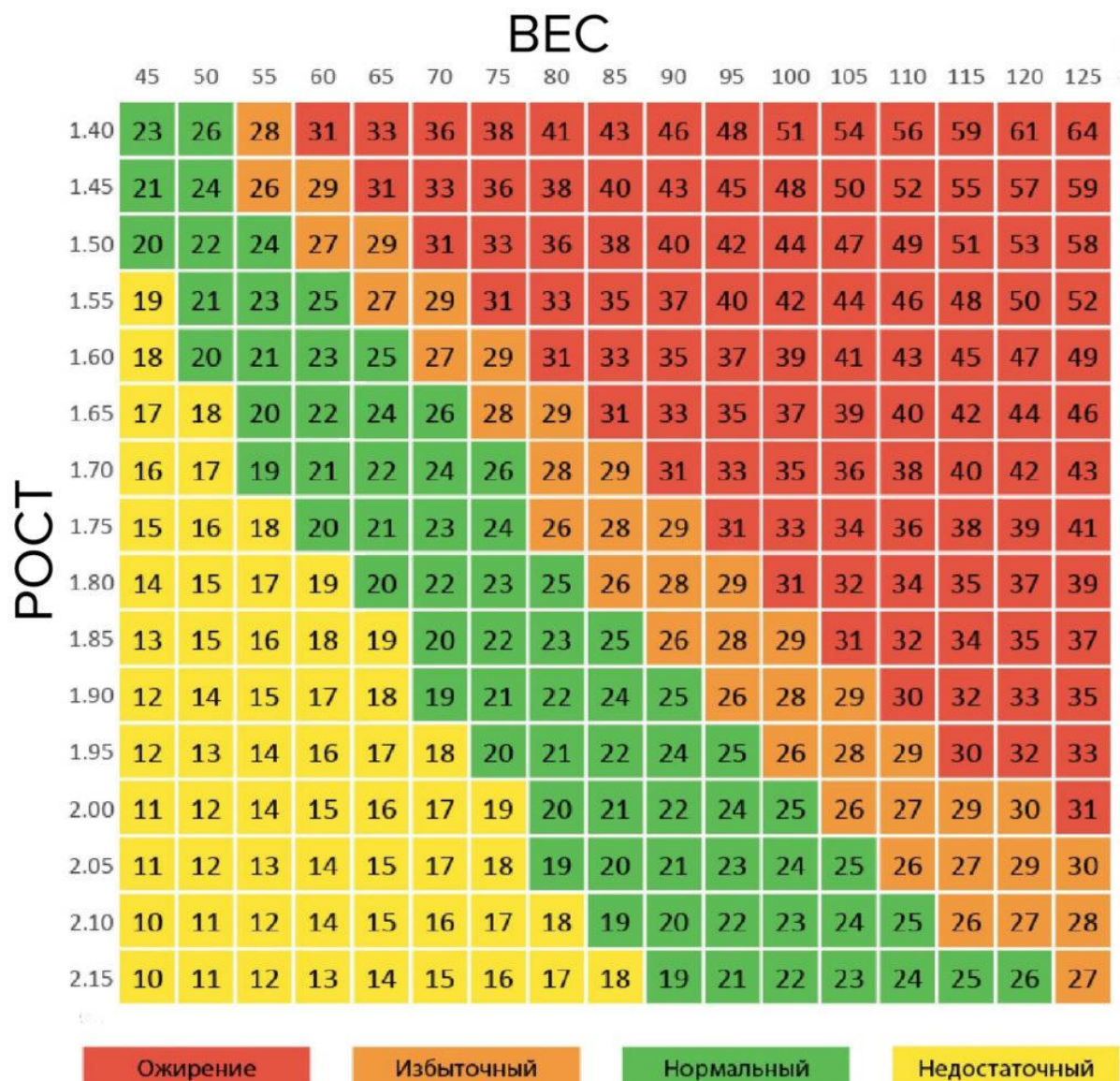
Computed type: Computed with drag-and-drop

Computed column definition Missing Values

Рост, см - 100

Computed columns code applied

40. Проанализируем полученные данные. Найдем тех, кто страдает избыточным весом в соответствии с ИМТ по таблице ИМТ:



41. Например, найдем женщин и мужчин, которые страдают ожирением. Для этого перейдем в фильтрацию данных и зададим следующие условия:

- Номер
- Фамилия
- ▮ Рост, см
- ▮ Вес, кг

▮ ИМТ > 30

42. В результате получим таблицу следующего вида:

	Номер	Фамилия	Рост, см	Вес, кг	Пол	Возраст, годы	Занятие спортом	ИМТ	Рост-Вес	Рост-100
1	1	Иванов	165	90	М	22	нет	33.05785124	75	65
2	2	Петеров	175	75	М	25	да	24.48979592	100	75
3	3	Сидоров	182	85	М	21	да	25.66115203	97	82
4	4	Губин	190	88	М	28	нет	24.3767313	102	90
5	5	Самойлов	190	95	М	31	да	26.31578947	95	90
6	6	Тумина	155	75	Ж	30	нет	31.21748179	80	55
7	7	Минаева	162	56	Ж	24	да	21.33821064	106	62
8	8	Голубок	172	74	М	23	да	25.01352082	98	72
9	9	Гнедая	165	59	Ж	24	да	21.67125803	106	65
10	10	Гормаш	170	110	М	28	нет	38.06228374	60	70

43. Найдем также тех, кто страдает дистрофией, недостаточным весом, избыточным весом, используя таблицу:

Индекс массы тела	
Результат вычисления	Значение
<16	Дистрофия
16 – 18,5	Недостаточный вес
18,5 – 25	Норма
25 – 30	Избыточный вес
30 – 35	Ожирение 1-й степени
35 – 40	Ожирение 2-й степени
>40	Ожирение 3-й степени

Полученные данные внесем в таблицу MS Excel.

44. Теперь сравним данные, полученные по таблице ИМТ, с вычисленными по формуле Брока: Индекс Брока = Рост – 100.

Данная формула показывает идеальный вес.

Более точный вариант данного индекса, учитывающий гендерные различия между мужчиной и женщиной, имеет следующий вид:

Индекс Брока для женщин:

Идеальный вес = (Рост – 100) · 0,85.

Индекс Брока для мужчин:

Идеальный вес = (Рост – 100) · 0,9.

45. Произведем расчеты индекса Брока для мужчин и женщин. Для этого создадим два столбца «Индекс Брока для женщин» и «Индекс Брока для мужчин», произведем соответствующие расчеты:

Name: Long name:

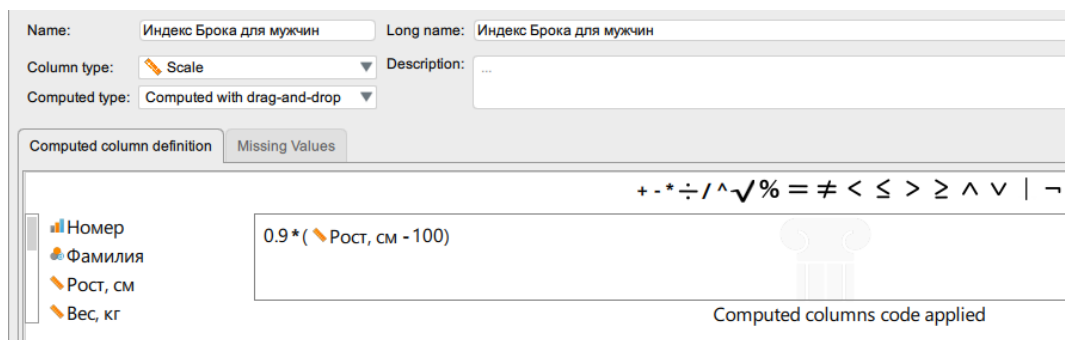
Column type: Description:

Computed type:

Computed column definition

☐ Номер
☐ Фамилия
☐ Рост, см
☐ Вес, кг

0.85 * (- 100)



46. Проанализируем полученные результаты и сравним данные идеального веса по Броку для мужчин и женщин с ИМТ. Используем фильтрацию данных.

47. Например, посмотрим, как идеальный вес по Броку соотноситься с реальным весом, используя следующие условия фильтрации:



48. Теперь по аналогии сравним идеальный вес для мужчин и женщин, используя разные условия фильтрации, основываясь на таблице ИМТ. Результаты внесем в таблицу MS Excel.

49. Ознакомимся с меню JASP-анализа:



Доступ к основным параметрам анализа можно получить с главной панели инструментов. В настоящее время JASP предлагает следующие частотные (параметрические и непараметрические стандартные статистики) и альтернативные байесовские тесты:

Descriptives Descriptive stats	Regression - Correlation - Linear regression - Logistic regression
T-Tests - Independent - Paired - One sample	Frequencies - Binomial test - Multinomial test - Contingency tables - Log-linear regression*
ANOVA - Independent - Repeated measures - ANCOVA - MANOVA *	Factor - Principal Component Analysis (PCA)* - Exploratory Factor Analysis (EFA)* - Confirmatory Factor Analysis (CFA)*
Mixed Models* - Linear Mixed Models - Generalised linear mixed models	

50. Если нажать на значок «+» в правом верхнем углу строки меню, также можно получить доступ к дополнительным параметрам, позволяющим добавлять дополнительные модули. После установки флажка они будут добавлены на главную ленту анализа. К ним относятся:

Audit	Network analysis
BAIN	Reliability analysis
Distributions	SEM
Equivalence tests	Summary statistics
JAGS	Visual modelling
Machine learning	Learning Bayes
Meta-analysis (included in this guide)	R (beta)

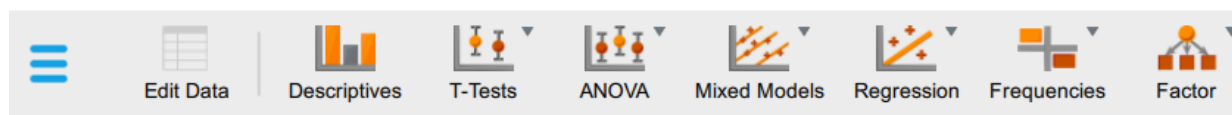
51. После того, как мы выбрали необходимый тип анализа, все возможные статистические параметры появляются в левом окне и выводятся в правом окне. JASP предоставляет возможность переименовывать и «складывать» выходные результаты, тем самым организуя несколько анализов:



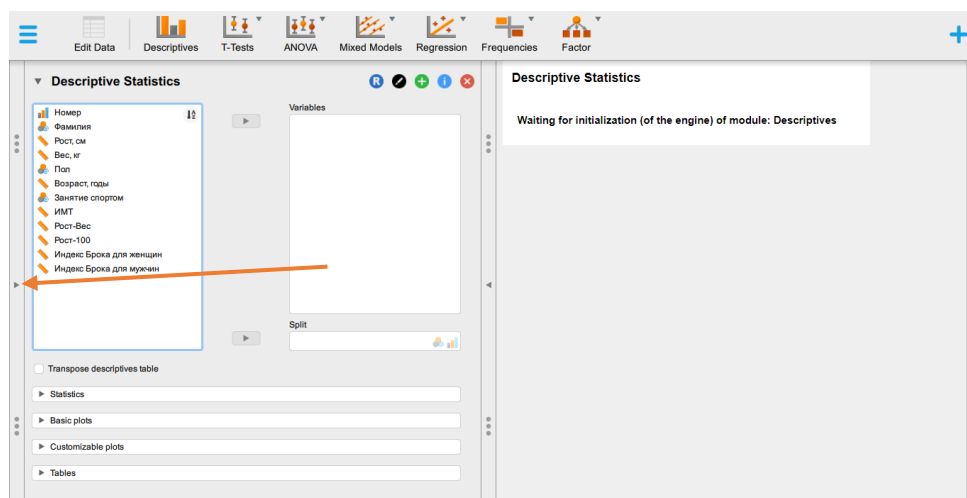
52. Отдельные анализы можно переименовать с помощью значка пера или удалить с помощью крестика  – крайнего значка справа. Нажав на опцию «Analysis» в этом списке, мы перейдем в соответствующую часть окна вывода результатов. Их также можно переупорядочить, перетаскивая каждый из анализов.

Значок  (второй слева) создает копию выбранного анализа. Значок информации  (третий слева) предоставляет подробную информацию о каждой из используемых статистических процедур и включает опцию поиска.

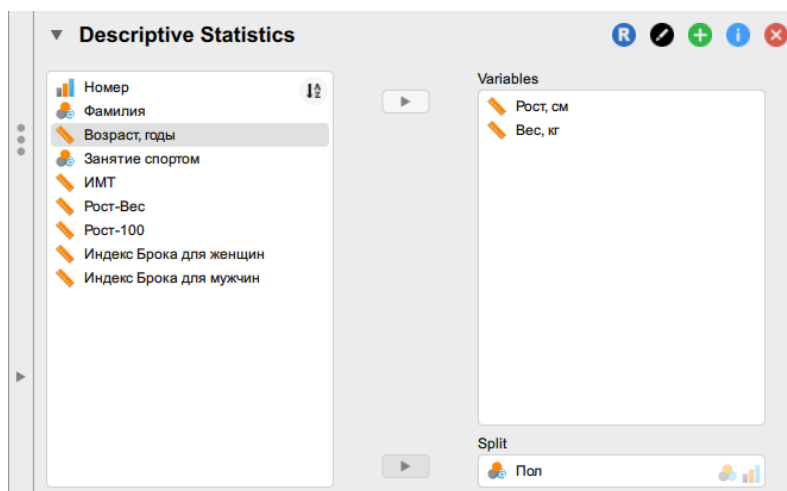
53. JASP имеет оптимизированный интерфейс для переключения между представлениями электронной таблицы, типов анализа и результатов. Например, мы хотим провести описательный анализ данных. Для этого нажмем на «Descriptives»:



54. Мы окажемся в следующем окне. Чтобы вернуться к таблице с данными, просто нажмем стрелочку сбоку. Аналогично, чтобы вернуться в раздел описательной статистики, нужно нажать стрелочку на рисунке:



55. Посчитаем среднее значение роста и веса у мужчин и женщин соответственно. Перетащим переменные, как показано на рисунке:



56. Получим соответствующие результаты вычислений:

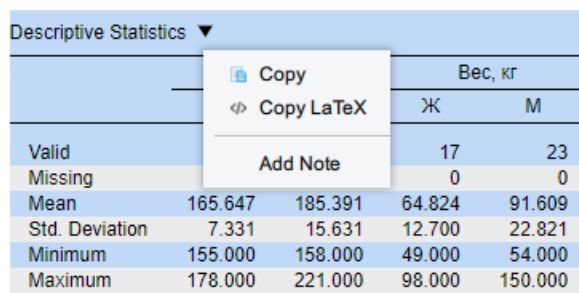
Results ▾

Descriptive Statistics ▾

Descriptive Statistics ▾

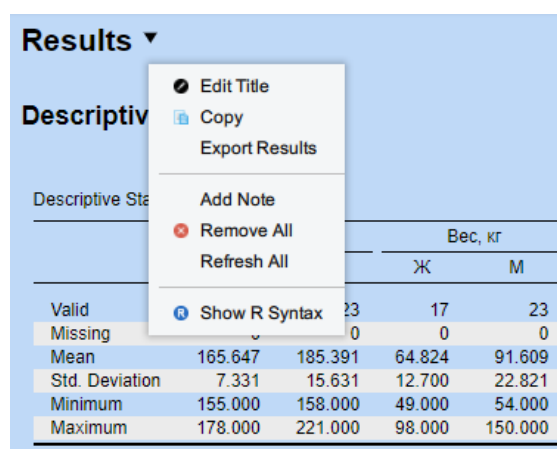
	Рост, см		Вес, кг	
	Ж	М	Ж	М
Valid	17	23	17	23
Missing	0	0	0	0
Mean	165.647	185.391	64.824	91.609
Std. Deviation	7.331	15.631	12.700	22.821
Minimum	155.000	158.000	49.000	54.000
Maximum	178.000	221.000	98.000	150.000

57. Можно скопировать таблицу либо график, нажав на стрелочку и выбрав подходящий вариант из выпадающего списка:



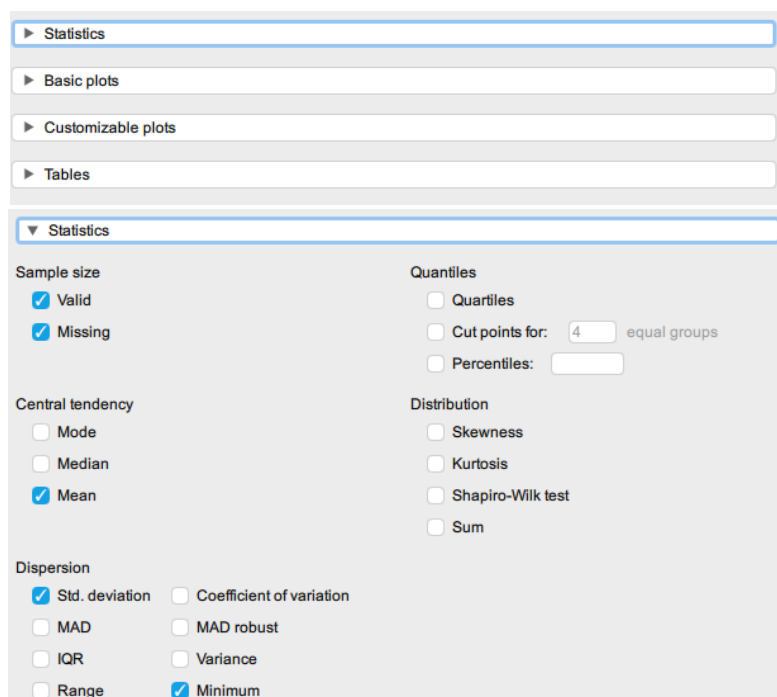
Descriptive Statistics				
	Вес, кг			
			Ж	М
Valid			17	23
Missing			0	0
Mean	165.647	185.391	64.824	91.609
Std. Deviation	7.331	15.631	12.700	22.821
Minimum	155.000	158.000	49.000	54.000
Maximum	178.000	221.000	98.000	150.000

58. Можно также целиком скопировать, удалить и т. д. результаты, нажав на стрелочку рядом с «Results»:



Results				
Descriptive Statistics				
Descriptive Statistics				
	Вес, кг			
			Ж	М
Valid			23	17
Missing			0	0
Mean	165.647	185.391	64.824	91.609
Std. Deviation	7.331	15.631	12.700	22.821
Minimum	155.000	158.000	49.000	54.000
Maximum	178.000	221.000	98.000	150.000

59. Внизу под полями переменных есть пункты настройки анализа. Если нажать на них, откроются меню настроек:



Statistics

Basic plots

Customizable plots

Tables

Statistics

Sample size

☒ Valid

☒ Missing

Central tendency

☐ Mode

☐ Median

☒ Mean

Dispersion

☒ Std. deviation

☐ Coefficient of variation

☐ MAD

☐ MAD robust

☐ IQR

☐ Variance

☐ Range

☒ Minimum

Quantiles

☐ Quartiles

☐ Cut points for: 4 equal groups

☐ Percentiles:

Distribution

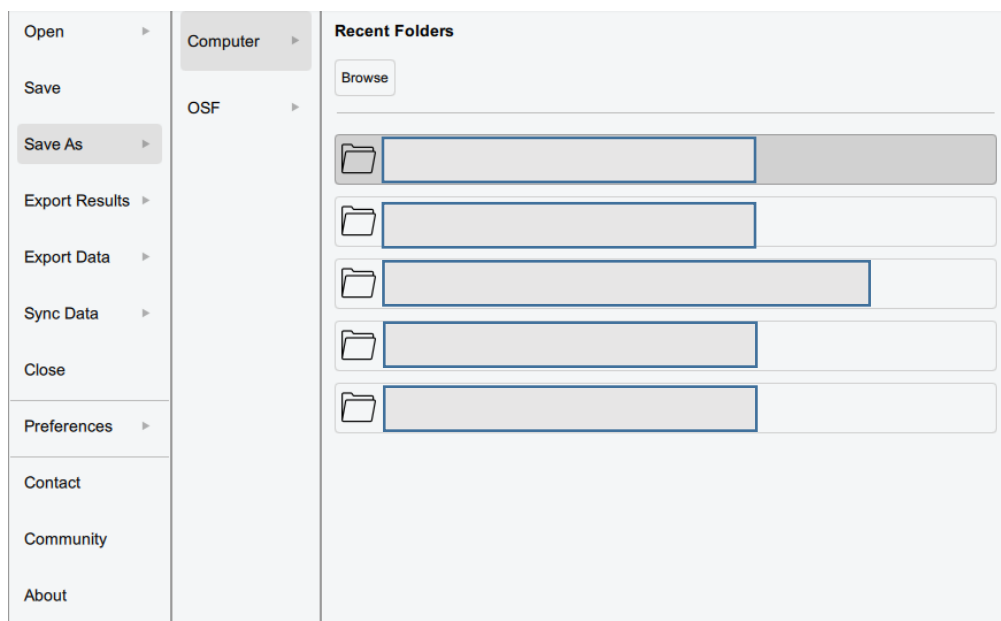
☐ Skewness

☐ Kurtosis

☐ Shapiro-Wilk test

☐ Sum

60. Теперь, когда мы ознакомились с основными функциями JASP, необходимо сохранить полученный файл. Для этого перейдем в главное меню. Через функцию «Save As» сохраним свои результаты в формате jasp, а через «Export Results» можно сохранить свои результаты в формате Portable Document Format (PDF) либо Hyper Text Markup Language (HTML). Кроме того, можно экспортировать свои данные через «Export Data» в форматах CSV (Comma-Separated Values) и т. д.:



Контрольные вопросы

1. Что такое JASP?
2. Кто разрабатывает JASP?
3. Каковы цели разработки JASP?
4. Какие статистические тесты доступны в JASP?
5. Может ли JASP обрабатывать большие объёмы данных?
6. Может ли JASP выполнять многомерный анализ данных?
7. Поддерживает ли JASP графический интерфейс?
8. Как импортировать данные в JASP?
9. Как изменить формат данных в JASP?
10. Как фильтровать данные в JASP?
11. Как группировать данные в JASP?
12. Как создать новые переменные на основе существующих данных в JASP?
13. Как преобразовать категориальные переменные в числовые в JASP?

14. Как проверить нормальность распределения данных в JASP?
15. Как проверить наличие выбросов в данных в JASP?
16. Как визуализировать корреляцию между двумя переменными в JASP?
17. Как визуализировать распределение одной переменной в JASP?

2. БИОСТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В MS EXCEL И JASP (ОСНОВНАЯ ЧАСТЬ)

2.1. Полигон частот

Краткие сведения из теории

Биомедицинская информация – это данные о биологических объектах и явлениях, которые являются предметом медицинских исследований. Она также включает в себя представления и суждения об этих объектах и явлениях.

Биомедицинская статистика – это метод анализа данных, полученных в результате экспериментов и клинических наблюдений из повседневной медицинской практики. Она также служит средством, с помощью которого исследователь может сообщить о своих результатах другим специалистам.

Статистическое исследование начинается с работы с данными, которые могут быть качественными, количественными или порядковыми. Важно понимать, какого типа данные будут обрабатываться для корректного использования статистических методов.

Количественные данные – это признаки, которые можно выразить в числовой форме, такие, например, как возраст, вес или количество детей в семье. Они могут быть непрерывными или дискретными. *Непрерывные данные* могут принимать любое значение на непрерывной шкале, например, температура или артериальное давление. *Дискретные данные* принимают конечное число значений, например, количество смертей в течение года в исследуемой группе.

Качественные данные – это признаки, которые нельзя выразить количественно, такие, например, как диагноз, место проживания или пол.

Порядковые данные – это показатели, измеряемые в шкале порядка, такие, например, как стадии болезни или оценки. Порядок состояний имеет значение, но смысл от изменения порядка не меняется.

Важно понимать, что количественные данные могут быть непрерывными или дискретными, а качественные данные не могут быть выражены в числовой форме.

Генеральная совокупность и выборка

В ходе статистического анализа исследователь обычно стремится сделать выводы обо всей совокупности объектов, например, как препарат воздействует на каждого человека с конкретной болезнью. Это означает, что исследователь хочет получить представление о свойствах всех изучаемых объектов по определенному признаку, например, «артериальное давление» или «люди в возрасте от 30 до 45 лет». Весь массив исследуемых объектов образует *генеральную совокупность*. Однако из-за различных факторов исследователь не может провести эксперимент над всеми элементами генеральной совокупности, поэтому он останавливается на достаточном для исследования количестве элементов, которые, по возможности, характеризуют всю генеральную совокупность. Это количество исследуемых объектов называется *выборкой*. Если выборка характеризует всю генеральную совокупность, то она называется *репрезентативной*.

Репрезентативность — очень важное свойство выборки. Если выборка не является репрезентативной, то исследователь может сделать ошибочные выводы обо всех объектах исследования. К сожалению, в медицинских и медико-биологических исследованиях часто бывает так, что выборки имеют очень небольшой объем, порядка 10–20 элементов.

Свойства генеральной совокупности:

1. Размер: генеральная совокупность может быть очень большой или даже бесконечной.
2. Характеристики: генеральная совокупность может иметь различные характеристики, такие как возраст, пол, доход и т. д.
3. Разнообразие: генеральная совокупность может включать в себя различные типы объектов, которые могут иметь различные характеристики.
4. Недоступность для изучения: из-за большого размера или других факторов генеральная совокупность может быть недоступна для полного изучения.

Свойства выборки:

1. Размер: размер выборки должен быть достаточно большим, чтобы обеспечить точность результатов, но не настолько большим, чтобы сделать исследование слишком дорогим или трудоемким.

2. Репрезентативность: выборка должна отражать характеристики генеральной совокупности.

3. Воспроизводимость: результаты, полученные на основе выборки, должны быть воспроизводимыми.

4. Эффективность: выборка должна быть способна дать точные результаты при минимальных затратах времени и ресурсов.

5. Валидность: выборка должна быть способна дать точные результаты, которые соответствуют целям исследования.

6. Достоверность: результаты, полученные на основе выборки, должны быть достоверными.

7. Однородность: объекты в выборке должны быть похожи друг на друга.

8. Объективность: результаты, полученные на основе выборки, должны быть независимыми от субъективных мнений или предубеждений.

9. Устойчивость: результаты, полученные на основе выборки, должны быть стабильными и не изменяться со временем.

Рандомизация

Рандомизация в статистике – это процесс выбора элементов из генеральной совокупности случайным образом. Это означает, что каждый элемент генеральной совокупности имеет равные шансы быть выбранным для включения в выборку. Рандомизация используется для того, чтобы обеспечить объективность и непредвзятость выборки, а также для уменьшения возможности ошибок и искажений в результатах исследования.

Рандомизация может быть осуществлена различными способами, включая использование таблиц случайных чисел, генераторов случайных чисел или других методов, которые позволяют выбрать элементы случайным образом.

Важно отметить, что рандомизация не гарантирует, что выборка будет репрезентативной, т. е. будет точно отражать характеристики генеральной совокупности. Однако она является важным шагом в обеспечении объективности и надежности результатов исследования.

Рандомизация также может быть использована для контроля за возможными факторами, которые могут влиять на результаты исследования. Например, в медицинском исследовании, где сравниваются две

группы пациентов, одна из которых получает лечение, а другая – плацебо, рандомизация может быть использована для того, чтобы убедиться, что обе группы пациентов имеют схожие характеристики и не отличаются по другим факторам, которые могут влиять на результаты исследования.

В целом рандомизация является важным инструментом в статистике, который помогает обеспечить объективность и надежность результатов исследования.

Полигон частот

Распределение значений признака – это графическое представление частоты встречаемости различных значений признака в выборке. Оно позволяет визуализировать, как часто встречаются различные значения признака и как они распределены в выборке.

Полигон частот – это графическое представление распределения значений признака, где каждая вершина полигона соответствует частоте встречаемости определённого значения. В отличие от гистограммы, где данные отображаются в виде прямоугольников, полигон частот строится на основе точек, соединённых отрезками. Эти точки располагаются над значениями признака на уровне их частот, создавая непрерывную ломаную линию.

Полигон частот позволяет наглядно представить распределение значений признака и определить, какие значения встречаются чаще, а какие – реже. Он также может помочь выявить наличие выбросов или аномальных значений в выборке.

Построение полигона частот начинается с определения интервалов, в которые будут сгруппированы значения признака. Затем для каждого интервала определяется частота встречаемости значений, попадающих в этот интервал. После этого строится полигон, вершины которого соответствуют частотам встречаемости, а основания – серединам интервалов.

Полигон частот является полезным инструментом для анализа данных и может помочь в выявлении закономерностей и трендов в распределении значений признака. Он также может использоваться для сравнения распределений значений признака в разных выборках или для проверки гипотез о распределении значений признака.

Пример

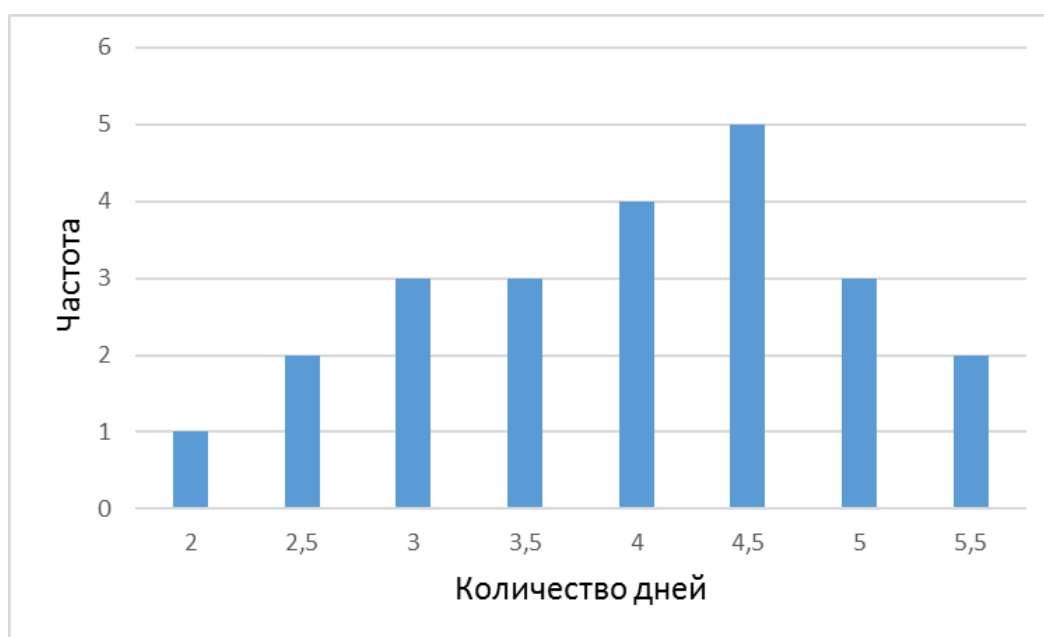
В ходе исследования группы людей на предмет влияния правильности метода лечения на сроки госпитализации (где переменной является число дней госпитализации) были получены следующие значения:

Количество дней госпитализации	2	2,5	2,5	3	3	3	3,5	3,5	3,5	4	4	4	4	4,5	4,5	4,5	4,5	4,5	5	5	5	5,5	5,5
--------------------------------------	---	-----	-----	---	---	---	-----	-----	-----	---	---	---	---	-----	-----	-----	-----	-----	---	---	---	-----	-----

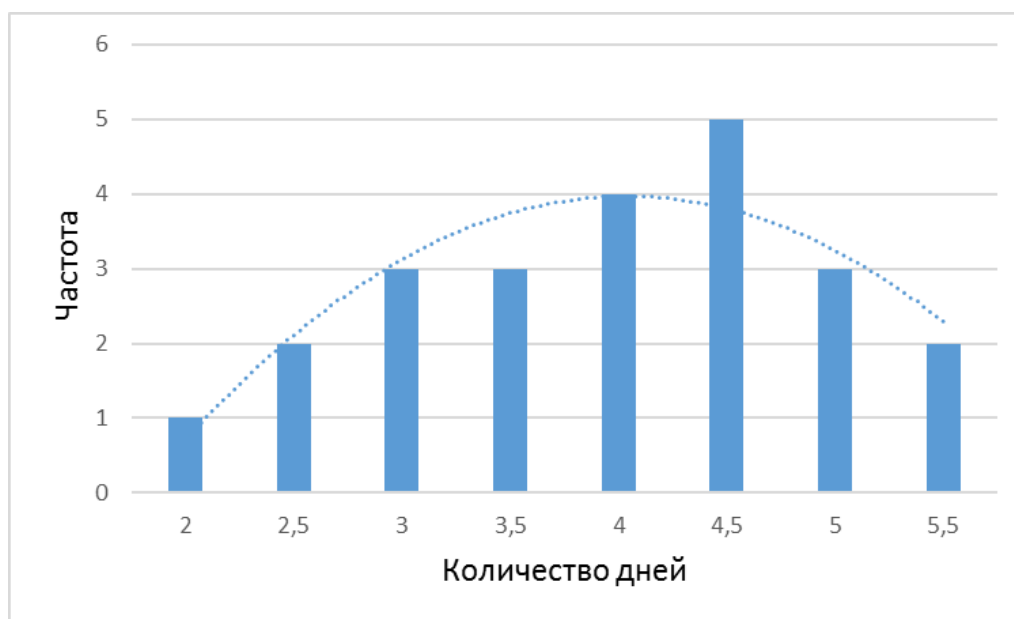
Запишем их в виде таблицы частот встречаемости данных значений:

Количество дней госпитализации	Частота встречаемости
2	1
2,5	2
3	3
3,5	3
4	4
4,5	5
5	3
5,5	2

Для построения графика распределения на горизонтальной оси X отметим значения «Количество дней госпитализации», а на вертикальной оси Y – сколько раз то или иное значение появилось в ходе исследования. Получим столбчатую диаграмму:



Строим также огибающую, полимодальную линию тренда в MS Excel:



Столбчатую диаграмму называют полигоном частот, а огибающую линию – графиком распределения частот.

Нередко вместо частоты встречаемости на графике изображают относительную частоту встречаемости.

Относительная частота встречаемости – это отношение частоты встречаемости определенного значения признака к общему числу наблюдений в выборке. Она показывает, как часто встречается данное значение признака относительно всех значений в выборке. Относительная частота встречаемости выражается в процентах и может быть использована для сравнения частоты встречаемости различных значений признака:

$$f = M/N,$$

где M – количество элементов выборки с заданным конкретным значением; N – общее количество элементов выборки.

Из вышеприведенного примера рассчитаем относительную частоту встречаемости дней госпитализации со значением 4,5:

$$f = 5/(1+2+3+3+4+5+3+2) = 5/23 = 0,2174.$$

Относительная частота встречаемости по количеству дней госпитализации со значением 4,5 дня равна 0,22, или, если это значение выразить в процентах, – 22 %.

Подсчитав все относительные частоты, получим следующую таблицу:

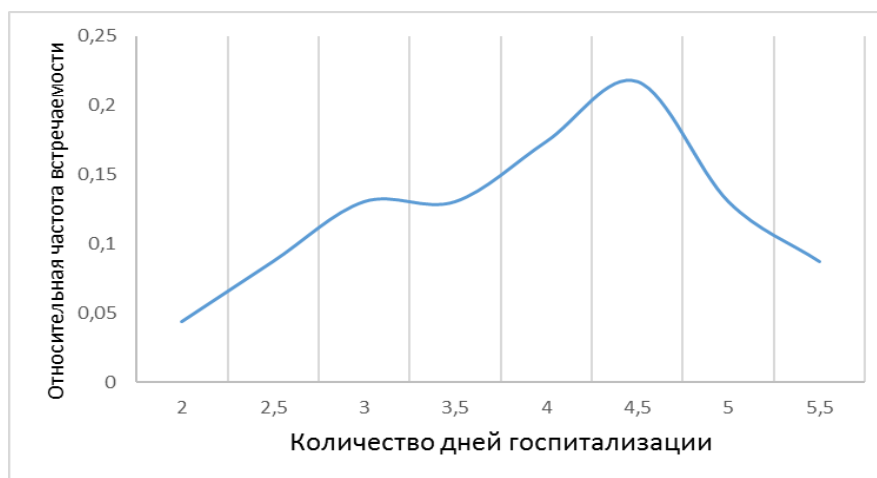
Количество дней госпитализации	Частота встречаемости	Относительная частота встречаемости	Относительная частота встречаемости, %
2	1	0,0435	4,35
2,5	2	0,0870	8,70
3	3	0,1304	13,04
3,5	3	0,1304	13,04
4	4	0,1739	17,39
4,5	5	0,2174	21,74
5	3	0,1304	13,04
5,5	2	0,0870	8,70
Сумма:	23	1	100

Для построения гистограммы случайная переменная – количество дней госпитализации – является дискретной, так как исследователь измеряет время пребывания в дискретных значениях, т. е. день делится не непрерывно – 2,5 дня, 2,75 дня, 3,8965 дня и т. п., а дискретно – 2 дня, 2,5 дня, 3 дня и т. д. Поэтому для отображения распределения частот лучше выбрать график типа гистограммы или полигона частот.

Если же представить случайную переменную как непрерывную, то можно изобразить график распределения в виде столбчатой диаграммы:

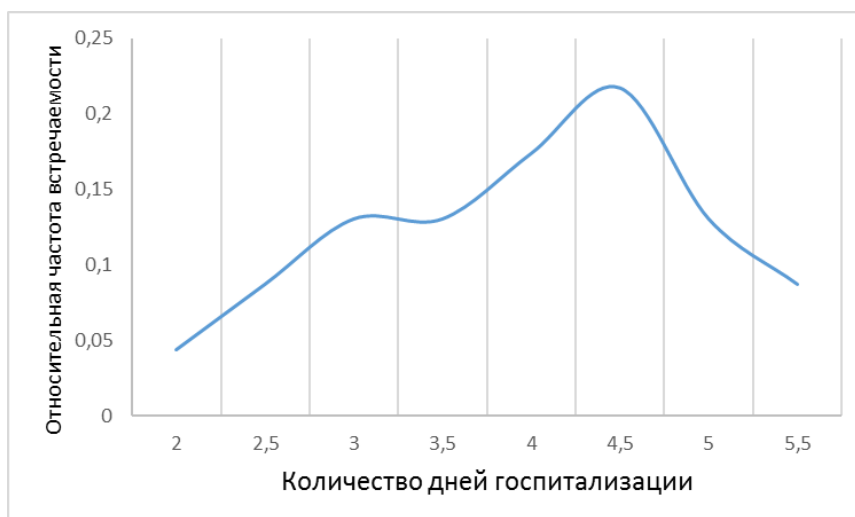


А кроме того, если случайную переменную рассматривать как непрерывную величину, можно построить график ее распределения в виде кривой:

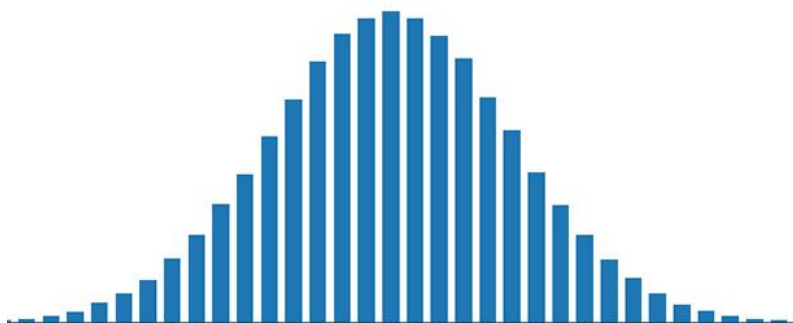


Использование относительных частот встречаемости позволяет выразить количество членов выборки с различными значениями исследуемого признака в процентном соотношении, что дает возможность предположить о частотном проявлении этого значения в генеральной совокупности при условии репрезентативности выборки. Важно отметить, что сумма относительных частот равна 1, а их процентное соотношение, соответственно, равно 100 %. Кроме того, площадь под кривой распределения всегда равна 1, если при этом используется выражение частоты встречаемости признака в виде относительной частоты встречаемости.

Площадь ограниченной области под кривой распределения равна доле и вероятности появления признака с заданными значениями, т. е., исходя из рисунка, доля членов выборки со значениями в интервале от 4,0 до 5,0 равна площади на графике в рамках этого интервала:



Если предположить, что рассматриваемая случайная величина, т. е. признак исследуемого объекта, непрерывна, то при увеличении количества измерений и уменьшении размера интервалов (иными словами, карманов), получим следующий график, который соответствует идеальному случаю нормального распределения:



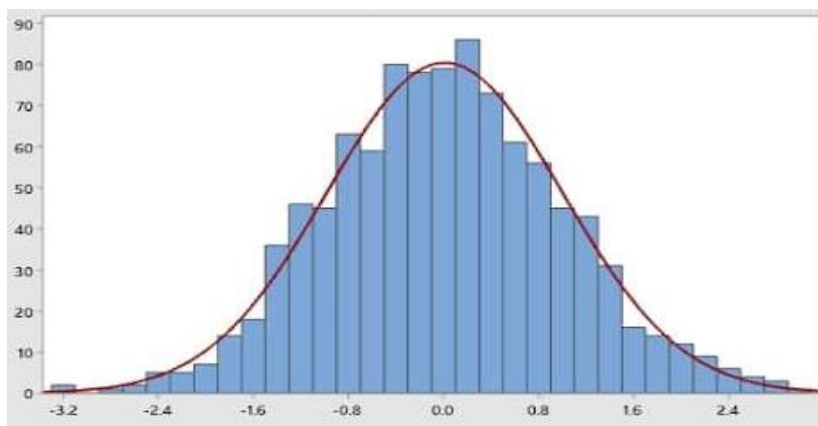
Виды распределений

Чаще всего в биологических исследованиях встречаются следующие виды распределения:

- 1) нормальное;
- 2) ассиметричное;
- 3) равномерное;
- 4) полимодальное.

Нормальное распределение (колоколообразное, гауссово) – это тип распределения вероятностей, который описывает случайную величину, имеющую непрерывное распределение. Оно имеет форму колоколообразной кривой, которая симметрична относительно своей средней точки.

Кривая нормального распределения имеет два параметра: среднее значение (μ) и стандартное отклонение (σ). Среднее значение определяет положение кривой относительно оси X, а стандартное отклонение определяет ее ширину:



Кривая нормального распределения имеет несколько важных свойств:

1. Площадь под кривой распределения равна 1. Это означает, что вероятность того, что случайная величина примет любое значение, равна 1.

2. Вероятность того, что случайная величина примет значение в определенном интервале, равна площади под кривой распределения в этом интервале.

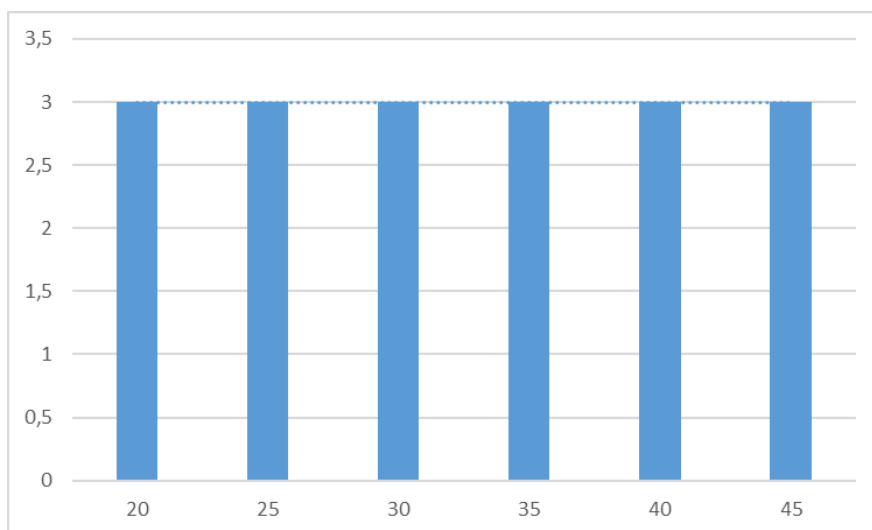
3. Вероятность того, что случайная величина примет значение, близкое к среднему, больше, чем вероятность того, что она примет значение, далекое от среднего. Это связано с тем, что кривая распределения имеет форму колокола.

4. Вероятность того, что случайная величина примет значение, равноудаленное от среднего, одинакова с обеих сторон. Это правило напрямую вытекает из симметричности кривой распределения.

Нормальное распределение широко используется в статистике и теории вероятностей для описания различных явлений, таких как ошибки измерений, результаты экспериментов и т. д. Оно также является основой для многих статистических методов, таких как метод наименьших квадратов и метод максимального правдоподобия.

Равномерное распределение – это тип распределения вероятностей, при котором все возможные значения случайной величины имеют одинаковую вероятность.

Кривая равномерного распределения имеет форму прямой линии, которая проходит через начало координат. Она имеет два параметра: нижнюю границу и верхнюю границу. Нижняя граница определяет минимальное значение случайной величины, а верхняя граница – максимальное значение:



Кривая равномерного распределения имеет несколько важных свойств:

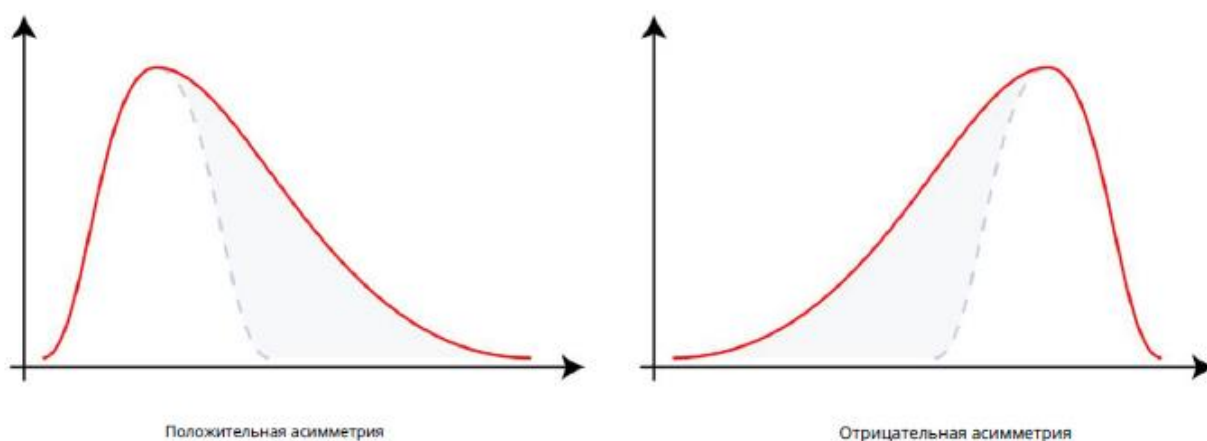
1. Вероятность того, что случайная величина примет значение в определенном интервале, равна площади под кривой распределения в этом интервале.

2. Вероятность того, что случайная величина примет значение, близкое к нижней границе, равна вероятности того, что она примет значение, близкое к верхней границе.

3. Вероятность того, что случайная величина примет значение, равномерно распределена по всему интервалу, и не зависит от положения внутри него.

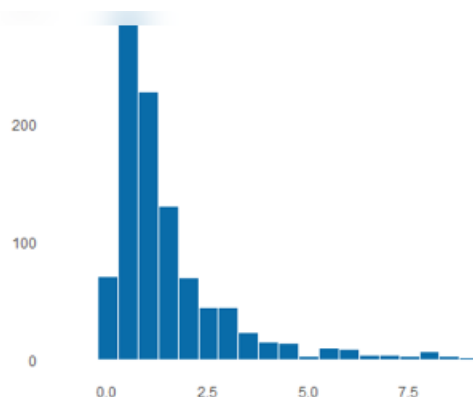
Ассиметричное распределение – это тип распределения вероятностей, при котором кривая распределения не симметрична относительно своей средней точки. Это означает, что вероятность того, что случайная величина примет значение, близкое к средней, не равна вероятности того, что она примет значение, далекое от средней.

Ассиметричное распределение может быть как левосторонним (когда кривая распределения смещена влево от средней точки – положительная асимметрия), так и правосторонним (когда кривая распределения смещена вправо от средней точки – отрицательная асимметрия):

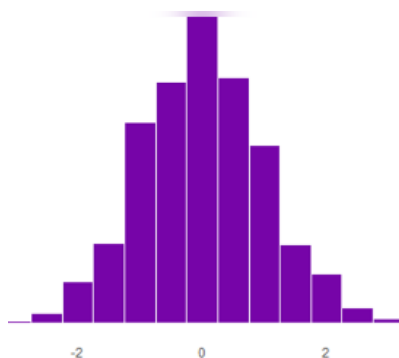


Если функцию $f(x)$ логнормального распределения преобразовать в ее логарифм $\log(f(x))$, то в этом случае полученная функция будет иметь нормальное распределение и характеризоваться теми же параметрами.

Используя графическое представление, такой случай можно продемонстрировать следующим образом:



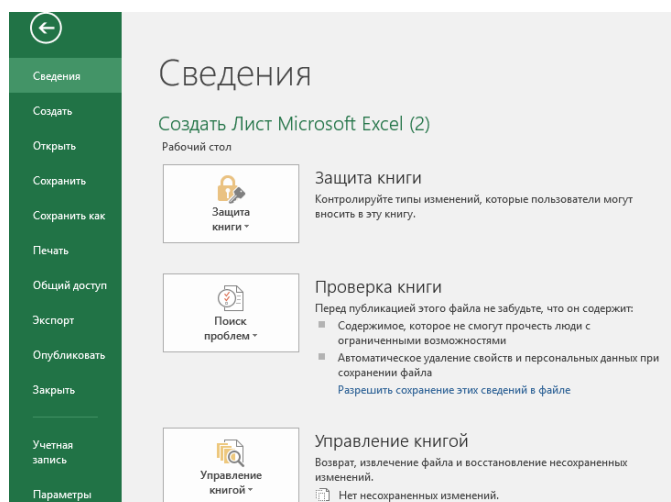
Если рассчитать десятичный логарифм от x и построить распределение получившихся значений, то мы получим следующий график:



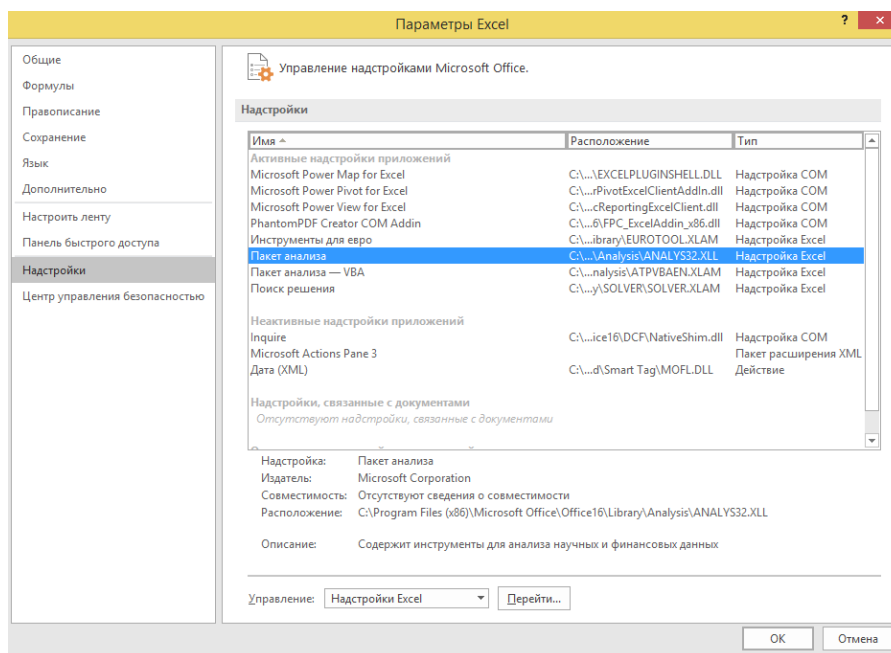
Алгоритм анализа в MS Excel

Для проведения статистического анализа данных в программе MS Excel используется модуль «Анализ данных». Он находится во вкладке «Данные». Если этот модуль отсутствует, его необходимо подключить.

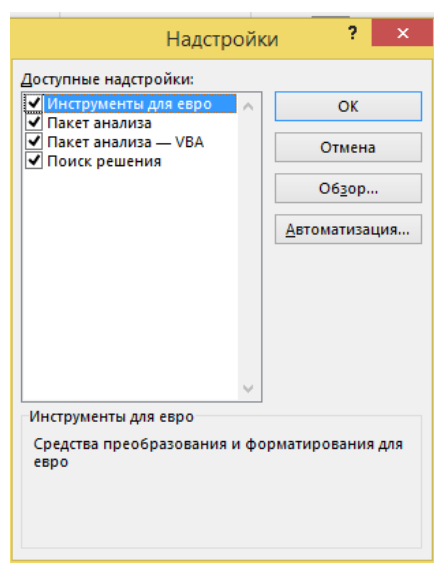
Для этого нужно вызвать меню «Надстройки» через опции «Файл» – «Параметры» – «Надстройки»:



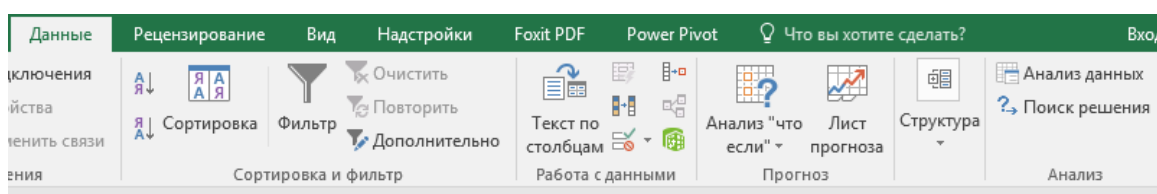
Затем в меню «Надстройки» выбрать «Пакет анализа» и нажать кнопку «Перейти»:



В открывшемся окне поставить галочку напротив опции «Пакет анализа» и нажать «ОК»:

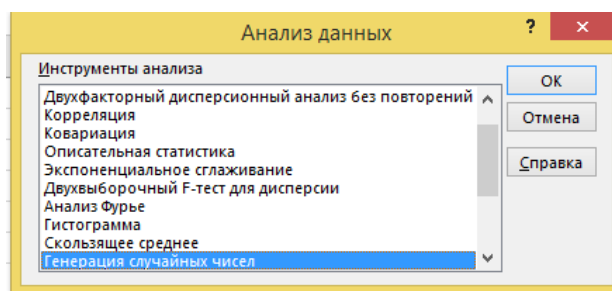


В итоге модуль «Анализ данных» появится во вкладке «Данные»:

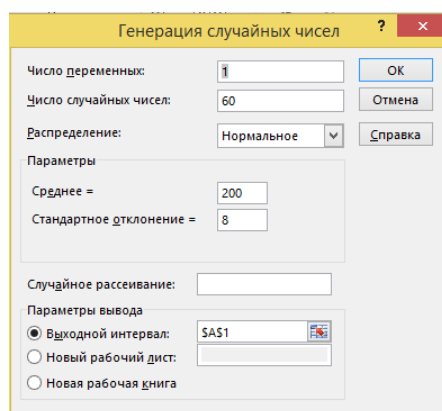


Рассмотрим работу модуля «Анализ данных» на примере генерации случайных чисел и построения графика их распределения.

1. В меню MS Excel откроем модуль «Анализ данных», выберем в нем «Генерация случайных чисел» и нажмем «ОК»:



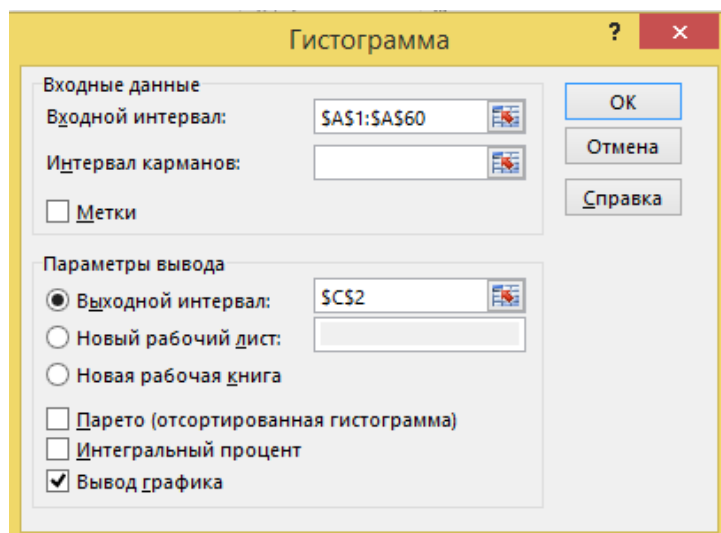
2. В появившемся окне выполним настройку с заданными для наглядности параметрами, как показано на рисунке:



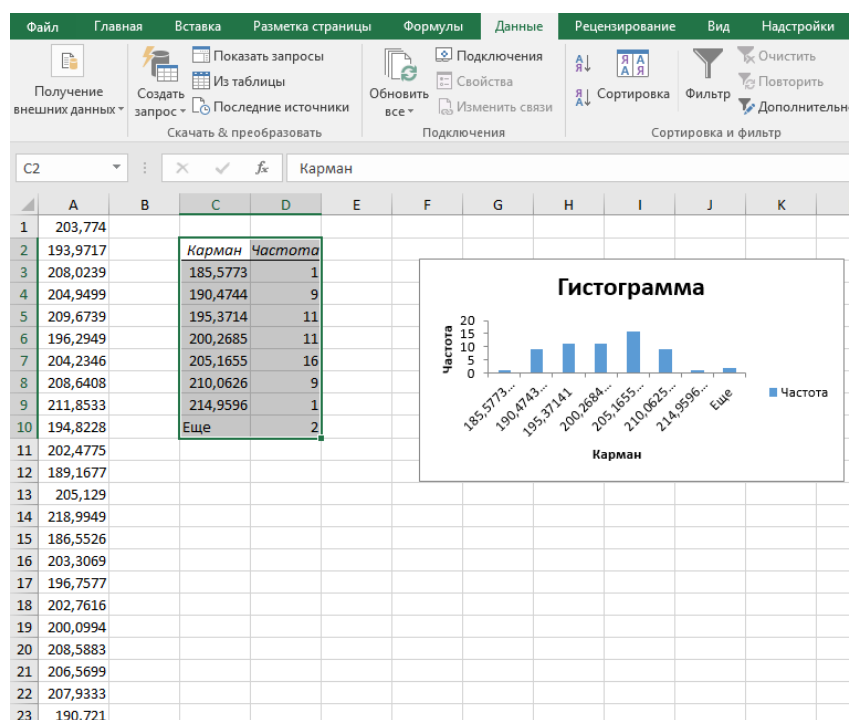
В результате получим, к примеру, следующие данные:

	A
1	203,774
2	193,9717
3	208,0239
4	204,9499
5	209,6739
6	196,2949
7	204,2346
8	208,6408
9	211,8533
10	194,8228
11	202,4775
12	189,1677
13	205,129
14	218,9949
15	186,5526
16	203,3069
17	196,7577
18	202,7616
19	200,0994
20	208,5883
21	206,5699
22	207,9333
23	190,721

3. Для построения графика распределения в меню «Анализ данных» выберем «Гистограмма». В появившемся окне в поле «Входной интервал» выберем все данные, полученные в результате генерации случайных чисел. Заполним остальные поля согласно указаниям на рисунке:



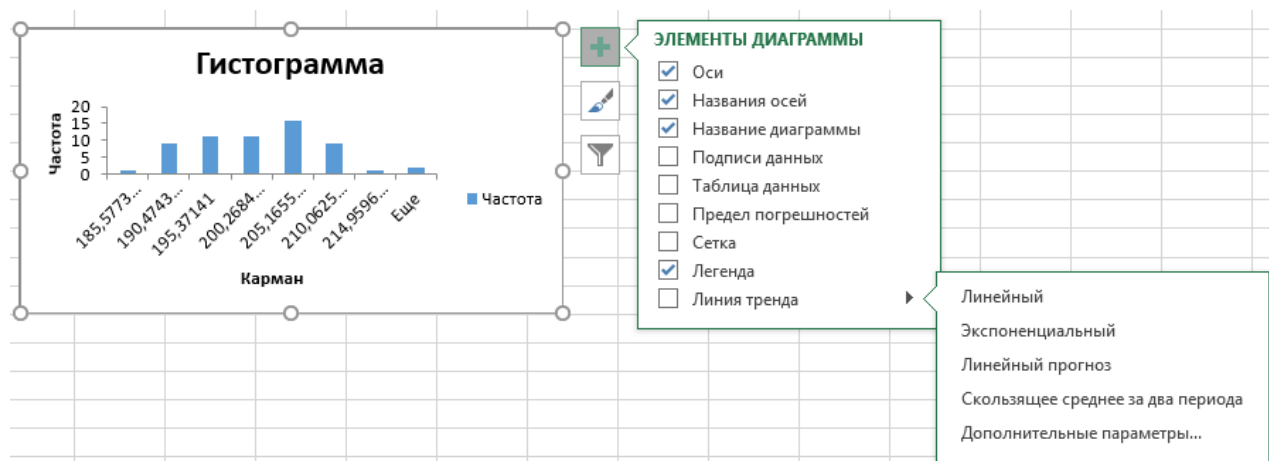
Результат работы модуля представлен в виде таблицы карманов и графика распределения, т. е. полигона частот:



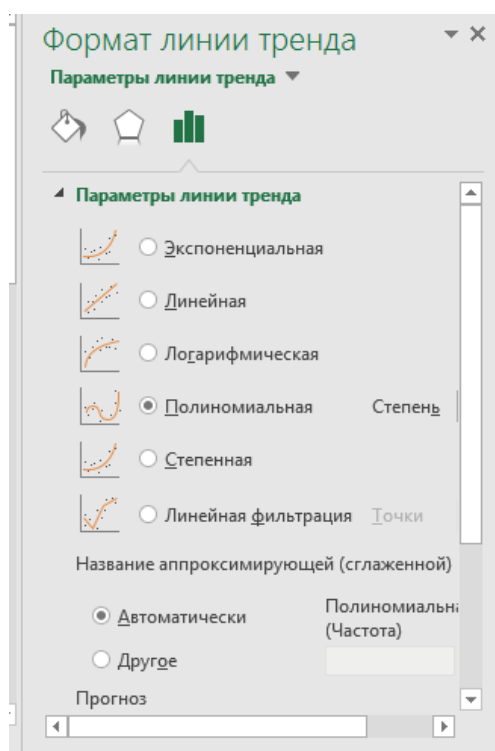
Карманы представляют собой интервалы, в которые попадают соответствующие значения в выборке. Например, карман 185,5773–190,4744 включает все числа, сгенерированные ранее и попадающие в этот интер-

вал, такие как 189,16; 186,55 и т. д. Карманы используются для визуального представления результатов. Они могут быть сформированы вручную или автоматически с помощью программы.

4. Для добавления линии тренда щелчком правой кнопкой мыши по любому столбцу на диаграмме и в появившемся меню выберем «Добавить линию тренда»:



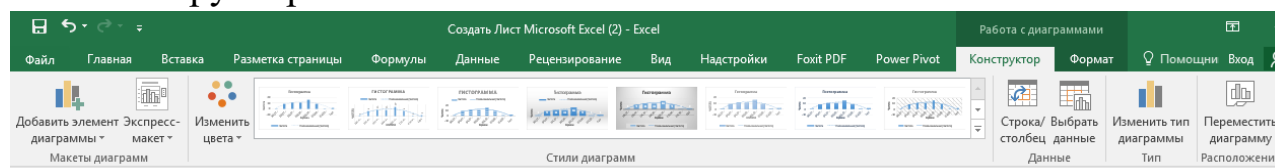
В появившемся окне выберем «Дополнительные параметры», далее – «Полимодалая», затем закроем окно:



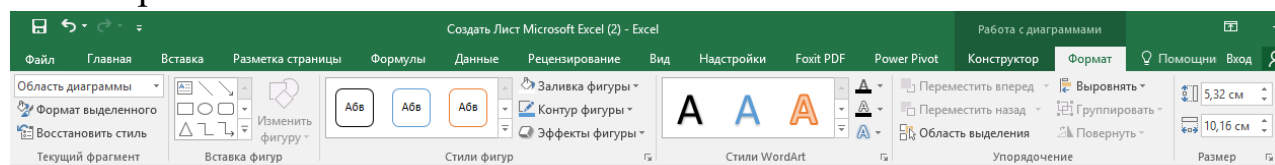
5. Чтобы изменить внешний вид полученной диаграммы и линии тренда, воспользуемся меню «Работа с диаграммами». Это поз-

волит настроить различные аспекты выбранной диаграммы, включая цвет, стиль, размер и многое другое, используя следующие функции MS Excel:

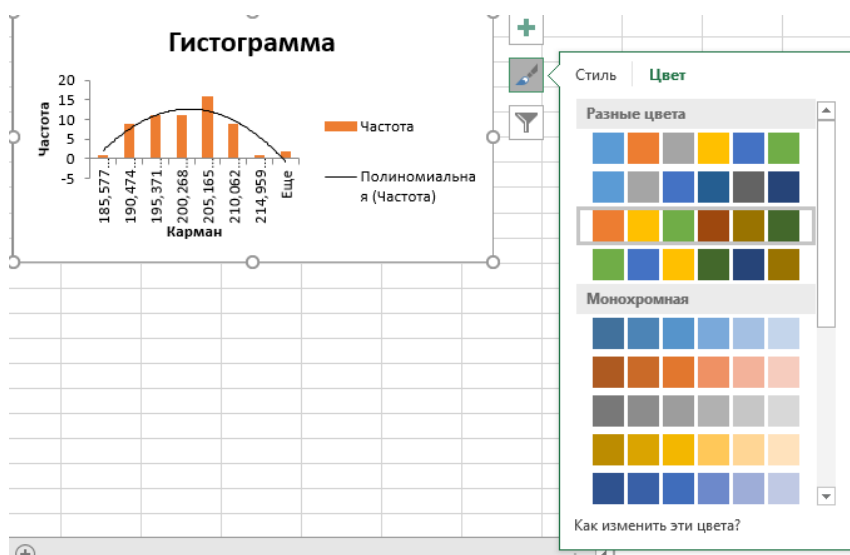
Конструктор:



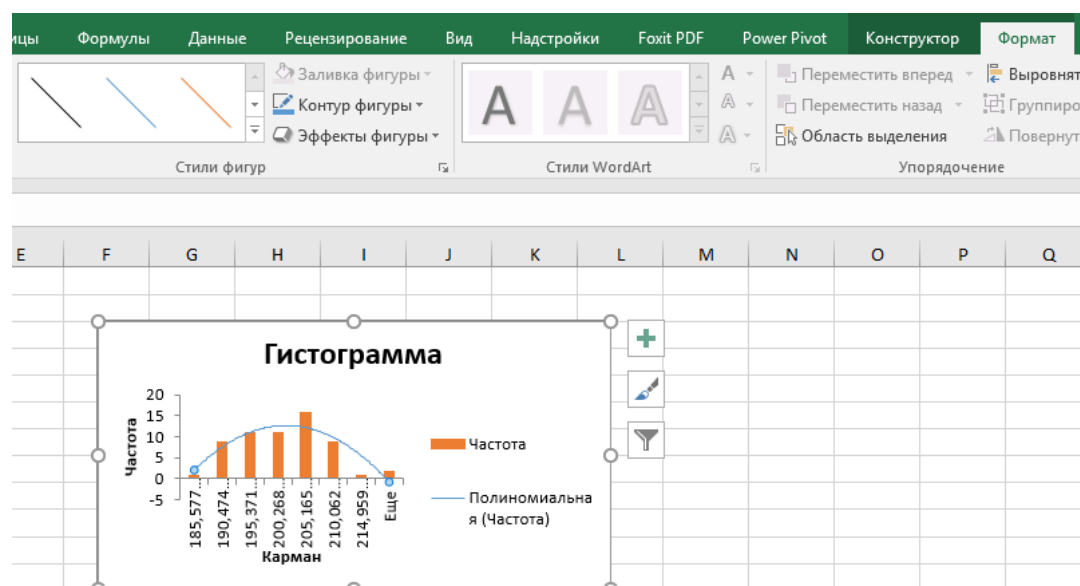
Формат:



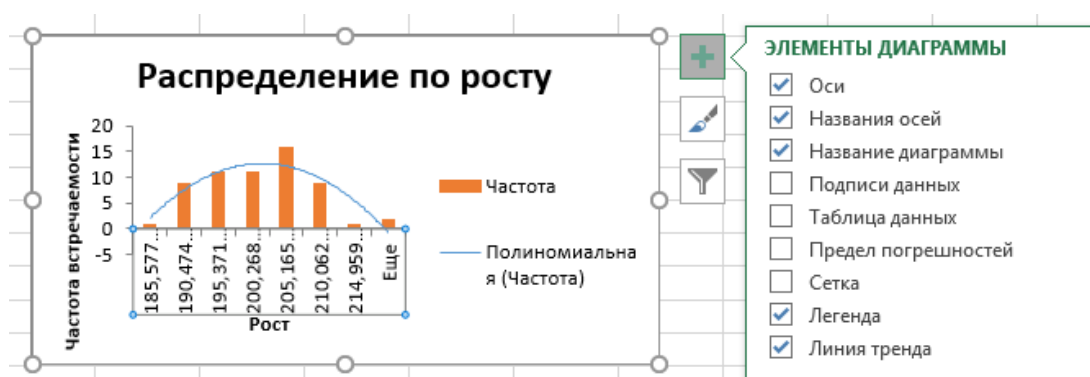
Боковые панели:



Формат линии тренда:



6. Ознакомимся с боковой панелью. Изменим названия осей и название диаграммы: горизонтальная ось – «Рост», вертикальная – «Частота встречаемости»; название гистограммы – «Распределение по росту»:

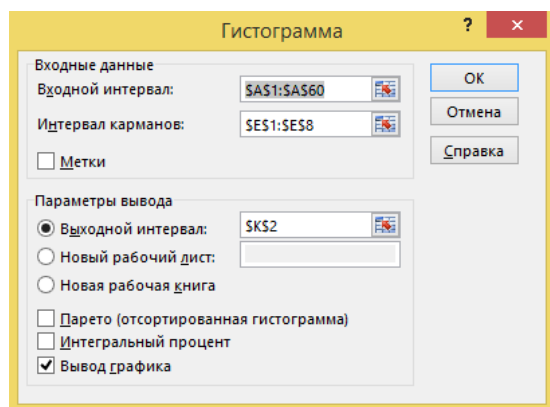


7. Удалим значение границы кармана, обозначенного как «Ещё», и заменим его на подходящее по смыслу значение. Исправим значения интервалов карманов на более приемлемые по внешнему виду: вместо 185,5773058 на 185,57 или 185,6. Доработаем внешнее представление графика до приемлемого вида.

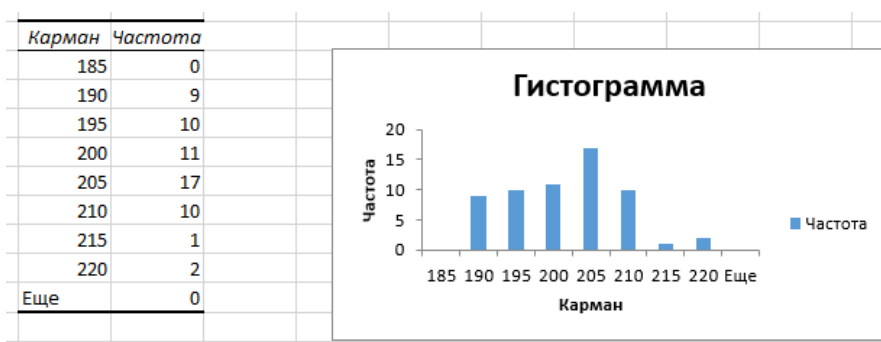
8. Можно создать график самостоятельно, удалив предыдущий и таблицу с интервалами карманов, сформированными программой. Затем нужно создать новую таблицу с интервалами карманов, которые пользователь определит сам. Сначала необходимо отсортировать свои данные от наименьшего к наибольшему. После этого определить минимальное значение первого кармана, округлив минимальное значение в меньшую сторону до ближайшего круглого числа. Затем определить максимальное значение последнего кармана, округлив максимальное значение в большую сторону до ближайшего круглого числа. Пример таблицы с интервалами карманов:

185
190
195
200
205
210
215
220

9. Выполняя построение графика распределения, необходимо указать в графе «Интервал карманов» ячейки полученной таблицы. В данном случае значения карманов находятся в ячейках от E1 до E8, однако при желании можно выбрать другое расположение:

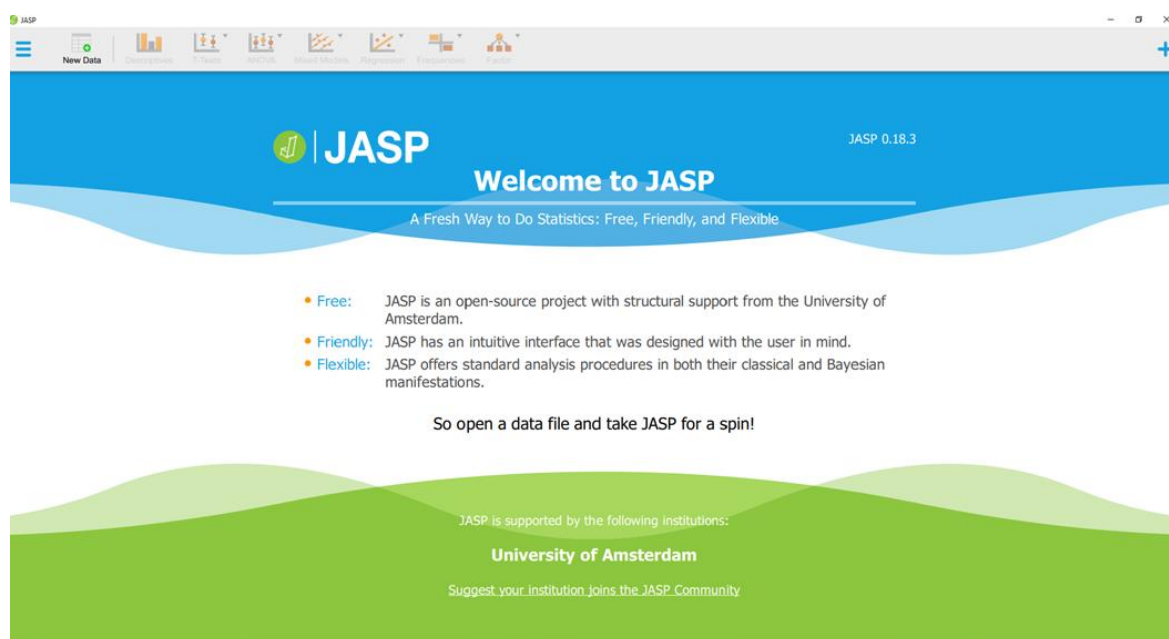


10. Полученный результат:



Алгоритм анализа в JASP

1. Откроем программу JASP – появится следующее окно:



2. Нажмем «New Data»:



3. Появится табличный процессор:

	Column 1							
7								
8								
9								
10								

4. Дважды щелкнем правой кнопкой мышки по «Column 1» – откроется режим редактирования столбца:

5. Переименуем «Column 1» в столбец «№»:

Name:	№	Long name:	№
Column type:	Scale	Description:	...
Computed type:	Not computed		

6. Изменим тип переменной на номинальный тип:

Analyses Synchronisation Insert Remove Undo Redo

Name: № Long name: №

Column type: **Nominal** Description: ...

Computed type: Not computed

7. Создадим последовательность чисел от 1 до 30:

	№	
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	

8. Создадим переменную «Рост, см»:

Analyses Synchronisation Insert Remove Undo Redo

	№	Рост, см				
1	1					
2	2					
3	3					
4	4					
5	5					
6	6					

9. Сгенерируем в столбце «Рост, см» данные по закону нормального распределения. Для этого перейдем в фильтр переменной «Рост, см»:

Analyses Synchronisation Insert Remove Undo Redo

Name: Рост, см Long name: Рост, см

Column type: **Scale** Description: ...

Computed type: Computed with drag-and-drop

Computed column definition: Missing Values

normalDist(185,10)

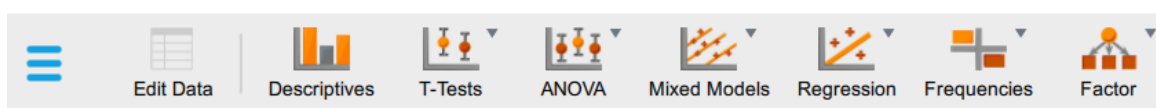
Computed columns code applied

normalDist(y)
tDist(y)
chiSqDist(y)
fDist(y)

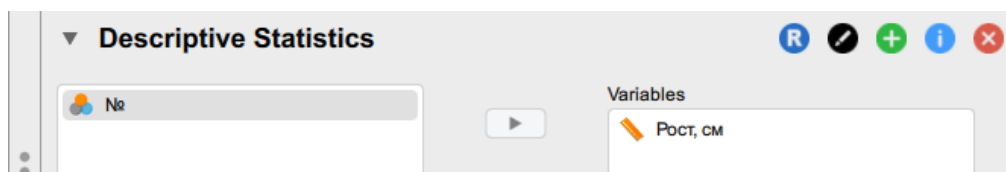
Compute column

	№	Рост, см
1	1	189.6446341
2	2	173.7035956
3	3	169.5478185
4	4	214.4473334
5	5	182.7689524
6	6	173.5694647
7	7	188.8760335

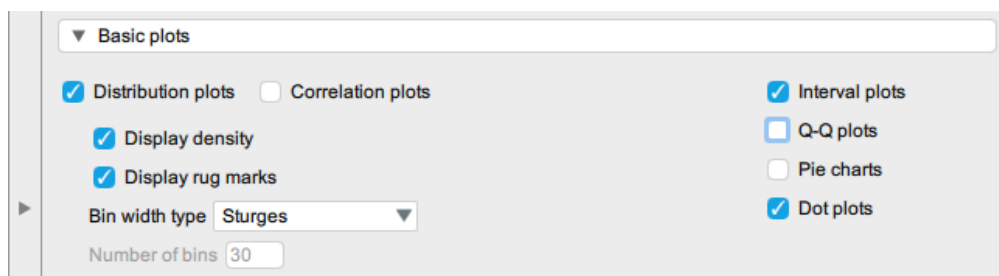
10. Перейдем во вкладку «Analyses» и выберем пункт «Descriptives»:



11. Перенесем переменную в поле Variables. Можно также перенести в поле «split» группирующую переменную для вывода результатов по отдельным группам. В данном случае у нас одна группа:



12. Построим диаграммы, открыв вкладку «Basic plots»:



13. Построим график «Boxplot», открыв вкладку «Customizable plots». Boxplots – коробчатые графики, или графики типа «ящик с усами». Характеризуют распределение с использованием квартилей и медианы.

Выбранные пункты:

Distributions plots – диаграммы (графики) распределения в виде гистограмм. Дополнительно выбран пункт «Display density», чтобы отобразить кривую распределения, которая рассчитывается только для количественных переменных.

Distribution plots / Display rug marks – гистограммы с указанием засечек («ковра») на оси X. Засечки отражают расстояние между отдельными значениями по переменной. Такие графики возможны только для количественных шкал.

Interval plots – интервальные графики, представляющие размах и медиану на одной линии. Такие графики строят для разных групп, так как для одной переменной он неинформативен.

14. Скопируем пошагово полученные данные на рабочий лист MS Excel и проанализируем полученные результаты.

Задание

Сгенерировать данные по закону нормального распределения, задав следующие параметры MS Excel:

Число переменных	Число случайных чисел	Среднее	Стандартное отклонение
1	30	100	20
2	60	150	30
3	90	200	40
4	120	250	60

Контрольные вопросы

1. Какие типы данных вы знаете? Приведите примеры.
2. Какова особенность порядковых данных и их отличие от качественных данных?
3. Что такое генеральная совокупность?
4. Что такое выборка?
5. Почему исследователь чаще всего вынужден использовать выборку?
6. Что означает репрезентативность выборки?
7. Что означает рандомный характер выборки?
8. Почему важна рандомизация выборки?
9. Что такое частота встречаемости?
10. Что такое распределение значений признака?
11. Какие типы распределения наиболее распространены?
12. Приведите примеры графического представления распределения признака.
13. Что такое равномерное распределение?
14. Что такое нормальное распределение?
15. Опишите алгоритм работы в MS Excel и JASP по данной теме.

2.2. Описательная статистика

Краткие сведения из теории

После проведения эксперимента исследователь получает данные, которые нужно обработать для формирования выводов и заключений. Эти данные обычно представлены в виде большого массива чисел, показателей или других значений, которые отражают проявление определенного признака. Например, при исследовании влияния анестетика на артериальное давление во время операции на открытом сердце исследователь получает таблицу с результатами, в которой указаны значения давления у каждого пациента до и во время операции, а также информация о выживаемости после операции. Поскольку работать с большим объемом данных сложно и неудобно, их стремятся представить в более удобном и наглядном виде для дальнейшего анализа.

Первым этапом статистического анализа является краткое описание данных, или описательная статистика.

Описательная часть биостатистики включает в себя следующее.

Формирование таблиц результатов анализа является предварительным этапом. На этом этапе данные проверяются на наличие артефактов (выбросов) и строится график распределения значений признака. Затем рассчитываются основные параметры распределения, такие как среднее значение, медиана, мода, стандартное отклонение и другие.

На основе полученных данных формируются выводы о принадлежности данных к определенному типу распределения. Это позволяет выбрать метод дальнейшего анализа. Например, если данные имеют нормальное распределение, то можно использовать параметрические методы анализа, такие как *t*-тест или анализ дисперсии. Если же данные имеют ненормальное распределение, то могут быть применены непараметрические методы анализа, такие как *U*-тест или критерий Манна–Уитни.

Параметры нормального распределения

При нормальном распределении используют параметрические методы статистики. Параметрами нормального распределения являются среднее значение (математическое ожидание), дисперсия и стандартное отклонение (среднеквадратичное отклонение), стандартная ошибка среднего.

Среднее значение в статистике – это величина, которая показывает среднее значение всех значений в выборке. Оно рассчитывается путем сложения всех значений и деления полученной суммы на количество значений в выборке.

Пример. Предположим, мы проводим исследование на мышах и хотим узнать, как влияет определенный препарат на их вес. Берем 10 мышей и взвешиваем каждую из них. Получаем следующие значения: 200, 220, 210, 230, 240, 250, 260, 270, 280 и 290 г.

Среднее значение веса мышей в этой выборке будет равно:

$$(200 + 220 + 210 + 230 + 240 + 250 + 260 + 270 + 280 + 290)/10 = 250 \text{ г.}$$

Дисперсия в статистике – это мера разброса значений в выборке относительно среднего значения. Она показывает, насколько значения в выборке отклоняются от среднего значения. Дисперсия измеряется в квадратных единицах измеряемого признака.

Для генеральной совокупности дисперсия рассчитывается по формуле

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}.$$

Для выборки дисперсия рассчитывается по формуле

$$s^2 = \frac{\sum (x_i - X)^2}{n - 1},$$

где σ^2 , s^2 – дисперсия; N и n – количество значений в выборке; x_i – значение i -го наблюдения; μ , X – среднее значение генеральной совокупности выборки соответственно.

Пример. Пусть у нас есть 5 значений: 2, 4, 6, 8, 10.

Среднее значение выборки (X) равно: $(2 + 4 + 6 + 8 + 10) / 5 = 6$.

Теперь мы можем рассчитать дисперсию:

$$s^2 = \frac{(2 - 6)^2 + (4 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 + (10 - 6)^2}{5 - 1} = 10.$$

Таким образом, дисперсия выборки равна 10.

Стандартное отклонение в статистике – это мера разброса, или дисперсии, случайной величины относительно ее среднего значения. Оно позволяет оценить, насколько значения случайной величины отклоняются от ее среднего значения.

Стандартное отклонение является корнем из дисперсии и вычисляется по формуле:

а) для генеральной совокупности:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}};$$

б) для выборки:

$$s = \sqrt{\frac{\sum (x_i - X)^2}{n - 1}}.$$

Стандартное отклонение измеряется в тех же единицах, что и сама случайная величина. Например, если случайная величина измеряется в метрах, то стандартное отклонение также будет измеряться в метрах.

Стандартное отклонение используется в различных областях, включая медицину и биологию. Оно позволяет оценить, насколько стабильны или изменчивы данные, а также помогает в принятии решений на основе статистических данных.

Например, если у нас есть данные о росте людей и мы хотим оценить, насколько сильно значения роста отклоняются от среднего значения, мы можем вычислить стандартное отклонение. Если стандартное отклонение большое, это означает, что значения роста сильно отклоняются от среднего значения, т. е. рост людей очень разнообразен. Если стандартное отклонение маленькое, это означает, что значения роста близки к среднему значению, т. е. рост людей относительно стабилен.

Таким образом, стандартное отклонение является важным инструментом в статистике, который помогает оценить разброс данных и принять обоснованные решения на основе этих данных.

Стандартная ошибка среднего (SEM, s_x) – это мера точности оценки среднего значения выборки. Она показывает, насколько точно среднее значение выборки отражает среднее значение генеральной совокупности.

Стандартная ошибка среднего вычисляется по формуле

$$s_x = \frac{s}{\sqrt{n}},$$

где s – стандартное отклонение; n – количество наблюдений в выборке.

Пример. Допустим, у нас есть выборка из 100 человек и мы хотим оценить их средний рост. Мы измеряем рост каждого человека и находим, что среднее значение выборки равно 170 см, а стандартное отклонение равно 5 см.

Стандартная ошибка среднего будет равна:

$$s_x = \frac{5}{\sqrt{100}} = 0,5 \text{ см.}$$

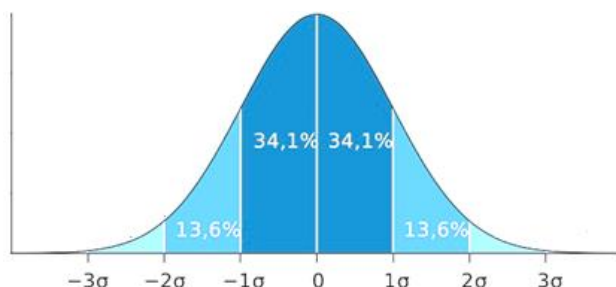
Это означает, что среднее значение выборки (170 см) имеет погрешность около 0,5 см, т. е. если мы приведем еще одну выборку из 100 человек, то среднее значение этой выборки будет в пределах $170 \pm 0,5$ см с вероятностью 68 %.

Таким образом, стандартная ошибка среднего позволяет оценить точность оценки среднего значения выборки и принять обоснованные решения на основе статистических данных.

Закон трех сигм

Закон трех сигм (или *правило трех сигм*) – это эмпирическое правило, которое гласит, что большинство значений случайной величины (около 68 %) лежат в пределах одного стандартного отклонения от среднего значения.

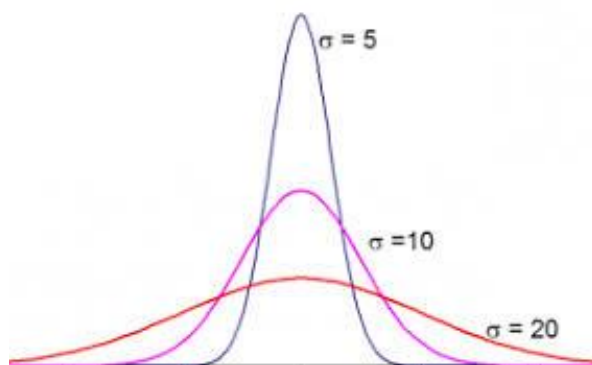
Это правило основано на нормальном распределении, которое является одним из наиболее распространенных распределений в статистике. В нормальном распределении около 68 % значений случайной величины лежат в пределах одного стандартного отклонения от среднего значения, около 95 % значений лежат в пределах двух стандартных отклонений и около 99,7 % значений лежат в пределах трех стандартных отклонений:



Закон трех сигм используется для оценки вероятности того, что значение случайной величины будет находиться в определенном диапазоне. Например, если у нас есть данные о росте людей и мы знаем, что среднее значение роста равно 170 см, а стандартное отклонение равно 5 см, то мы можем использовать закон трех сигм, чтобы оценить вероятность того, что рост человека будет находиться в пределах от 165 до 175 см.

Однако стоит отметить, что закон трех сигм является приближенным и не всегда точно отражает реальное распределение данных. В некоторых случаях, особенно при больших выборках или при наличии выбросов, закон трех сигм может давать неточные результаты. Поэтому при анализе данных всегда следует учитывать конкретные условия и контекст.

Благодаря графику нормального распределения можно представить зависимость ширины «колокола» от стандартного отклонения: чем больше дисперсия (или стандартное отклонение), тем шире «колокол», т. е. разброс значений признака больше:



Ненормальное распределение и его характеристики

Ненормальное распределение – это распределение случайной величины, которое не соответствует нормальному распределению. Оно может иметь различные формы и характеристики, которые отличаются от характеристик нормального распределения.

Ненормальное распределение может иметь различные причины. Например, оно может возникнуть, если данные имеют выбросы или асимметрию. Выбросы – это значения, которые сильно отклоняются от остальных значений в выборке. Асимметрия – это неравномерное распределение значений вокруг среднего значения.

Ненормальное распределение может затруднить анализ данных и принятие обоснованных решений на основе этих данных. Например, если мы используем методы, основанные на нормальном распределении, для анализа данных с ненормальным распределением, то результаты могут быть неточными или неправильными.

Для анализа данных с ненормальным распределением могут использоваться другие методы, такие как непараметрические методы или методы, основанные на других распределениях. Например, для анализа данных с выбросами можно использовать методы, которые учитывают выбросы или исключают их из анализа.

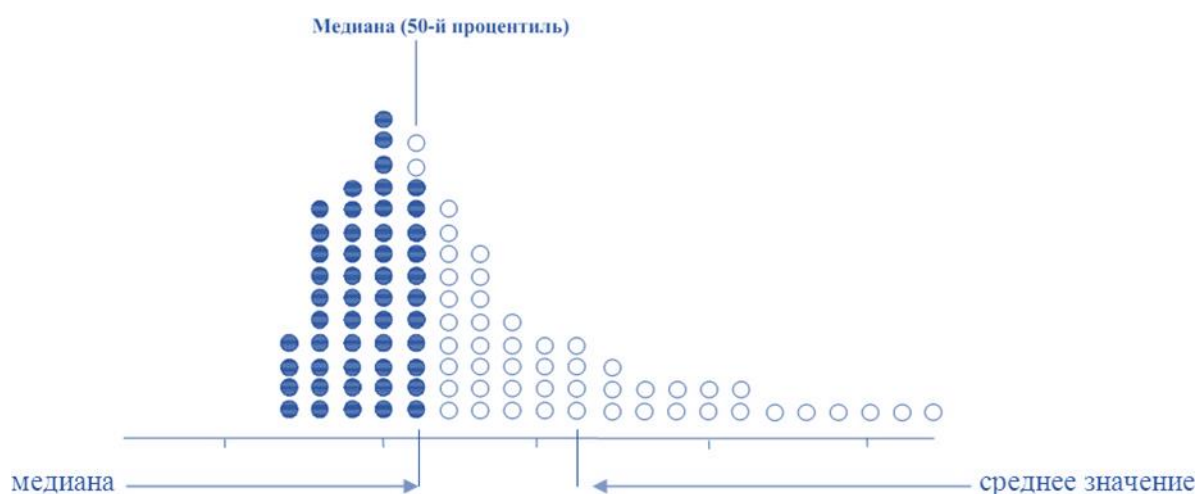
Важно отметить, что ненормальное распределение не является ошибкой или недостатком данных. Оно может быть естественным и ожидаемым в некоторых случаях. В целом ненормальное распределение

является важной концепцией в статистике, которая помогает анализировать данные и принимать обоснованные решения на основе этих данных.

Для описания ненормального распределения не используются параметры (среднее значение, дисперсия, стандартное отклонение, стандартная ошибка среднего), а используют медиану и процентиля (квартили).

Медиана – это значение, которое делит упорядоченный ряд данных на две равные части. Другими словами, половина значений в ряду данных будет больше медианы, а половина – меньше.

Медиана является одним из основных статистических показателей, который используется для описания центральной тенденции данных. Она позволяет оценить, какое значение является типичным или средним для данного ряда данных:



Медиана вычисляется следующим образом:

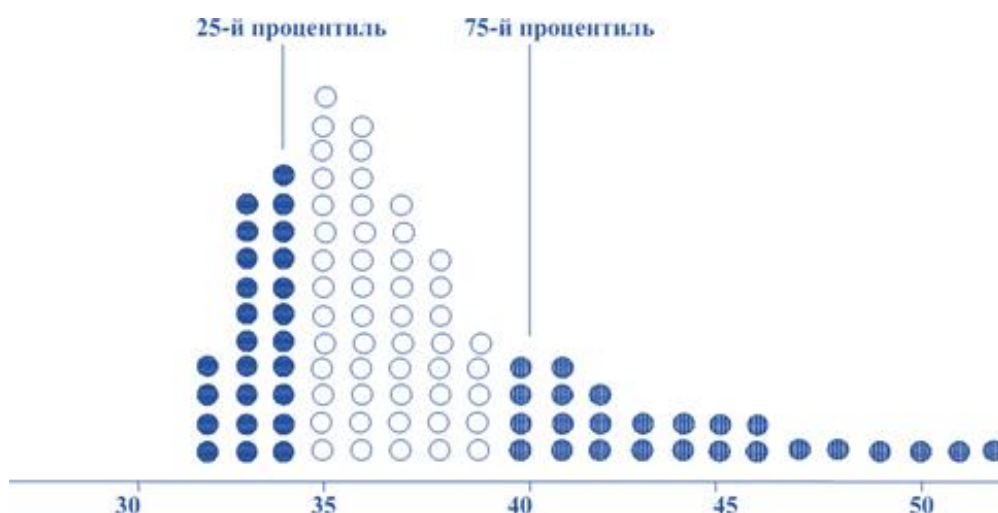
1. Сначала упорядочиваются все значения в ряду данных от наименьшего к наибольшему.
2. Затем определяется количество значений в ряду данных (n).
3. Если n является четным числом, то медиана равна значению, которое делит этот ряд пополам. Например, если у нас есть ряд данных $\{1, 2, 3, 4, 5\}$, то медиана будет равна 3.
4. Если n является нечетным числом, то медиана равна среднему значению между двумя средними значениями в упорядоченном ряду данных. Например, если у нас есть ряд данных $\{1, 2, 3, 4, 5, 6\}$, то медиана будет равна 3,5.

Медиана имеет несколько преимуществ по сравнению с другими показателями центральной тенденции, например, такими как среднее значение. Во-первых, она нечувствительна к выбросам или асимметрии

данных. Это означает, что даже если в ряду данных есть несколько очень больших или очень маленьких значений, медиана все равно будет отражать типичное значение для данного ряда данных. Во-вторых, медиана является более устойчивой к изменениям в данных, чем среднее значение. Это означает, что даже если некоторые значения в ряду данных изменятся, медиана останется относительно стабильной.

В целом, медиана является важным показателем в статистике, который помогает оценить центральную тенденцию данных и принять обоснованные решения на основе этих данных.

Процентили – это значения, которые делят упорядоченный ряд данных на несколько равных частей. Они используются для описания распределения данных и определения, какую долю данных составляют значения, меньшие или большие определенного процентиля:



Процентили вычисляются следующим образом:

1. Сначала упорядочиваются все значения в ряду данных от наименьшего к наибольшему.
2. Затем определяется количество значений в ряду данных (n).
3. Процентиль вычисляется как доля значений, меньших или равных данному процентилю, от общего количества значений в ряду данных. Например, если у нас есть ряд данных $\{1, 2, 3, 4, 5\}$ и мы хотим найти 25-й процентиль, то мы найдем значение, которое меньше или равно 25 % значений в ряду данных. В данном случае, 25-й процентиль будет равен 2.

Процентили используются для описания распределения данных и определения, какую долю данных составляют значения, меньшие или большие определенного процентиля. Например, если у нас есть ряд данных

об уровне холестерина у людей и мы хотим узнать, какой процент людей имеет холестерин меньше определенного уровня, мы можем использовать проценти́ли.

Артефакты, или *выбросы*, в статистике – это значения, которые сильно отклоняются от остальных значений в выборке. Они могут быть вызваны различными причинами, такими как ошибки измерения, неправильная классификация данных или случайные ошибки.

Артефакты могут оказывать значительное влияние на результаты анализа данных. Например, если мы используем методы, основанные на среднем значении или стандартном отклонении, артефакты могут исказить эти показатели.

Для анализа данных с артефактами могут использоваться различные методы. Например, можно исключить артефакты из анализа или использовать методы, которые учитывают их наличие.

Проверка выбросов может производиться по критерию, равному нормированному отклонению выброса:

$$T = \frac{x_i - X}{s} \geq T_{st},$$

где T – критерий выброса; x_i – выделяющееся значение признака (или очень большое, или очень малое); X – среднее значение; s – стандартное отклонение, рассчитанные для группы, включающей артефакт; T_{st} – стандартные значения критерия выбросов, определяемые по таблице:

n	T_{st}	n	T_{st}	n	T_{st}	n	T_{st}
2	2,0	16–20	2,4	47–66	2,8	125–174	3,2
3–4	2,1	21–28	2,5	67–84	2,9	175–349	3,3
5–9	2,2	29–34	2,6	85–104	3,0	350–599	3,4
10–15	2,3	35–46	2,7	105–124	3,1	600–1500	3,5

Если $T \geq T_{st}$, то анализируемое значение признака является выбросом. Альтернатива $T < T_{st}$ не позволяет исключить из анализа значение признака.

Алгоритм анализа в MS Excel

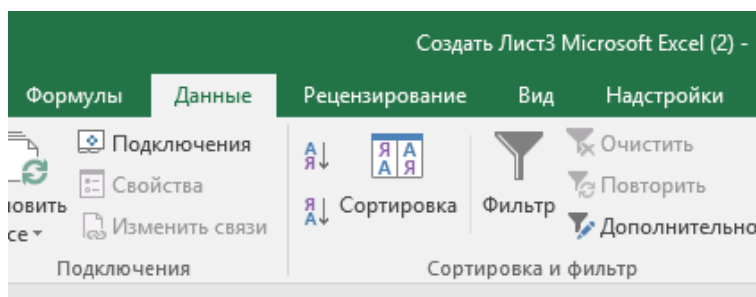
В рамках вымышленного исследования, целью которого является более глубокое изучение новой, ранее не исследованной народности, ученые решили начать с получения информации об антропологических

данных проживающих на данной территории людей. Одним из изучаемых признаков был выбран рост мужчин в возрасте от 20 до 40 лет. Данные представлены в виде таблицы либо можно сгенерировать собственные данные в соответствии с предложенным вариантом:

Рост	
164,96	156,47
169,24	153,75
150,76	181,70
152,55	183,19
187,75	181,25
168,48	178,37
189,61	148,56
181,24	182,35
190,12	159,67
188,34	187,16
169,75	173,61
188,33	158,09
163,91	165,58
206,46	171,36
167,55	127,09
166,88	183,99
186,46	189,89
174,92	180,29
192,60	167,38
172,90	160,18
190,70	170,02
166,10	167,43
151,82	175,06
190,59	199,61
164,51	142,60

1. Описание данных и построение графика распределения частот начинается с копирования таблицы в книгу MS Excel. Первый этап подготовки данных – это сортировка исходных данных. Для этого нужно

выделить анализируемый массив, а затем в меню выбрать сортировку от минимального к максимальному или сортировку от максимального к минимальному и кликнуть левой кнопкой мыши:



Для статистического анализа данных в программе MS Excel воспользуемся «Пакетом анализа».

2. В модуле «Анализ данных» выберем «Описательная статистика», после чего нажмем «ОК».

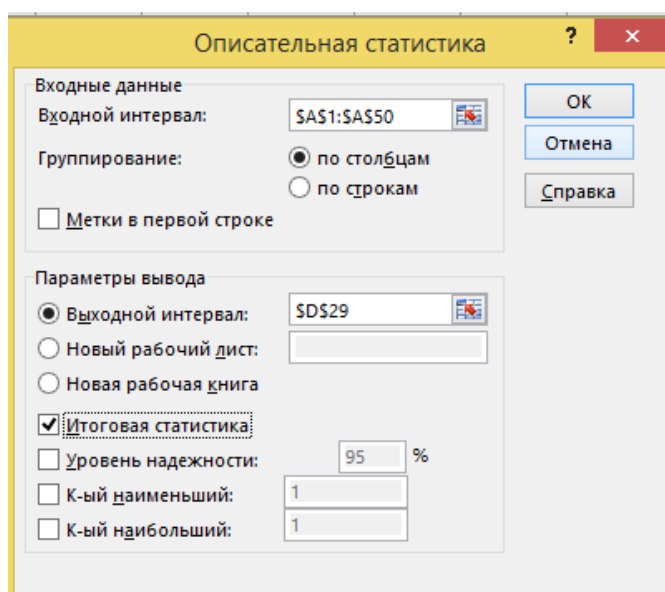
3. В появившемся окне выполним следующие установки:

а) в поле «Метки в первой строке» поставить маркер, если первая строка нашей таблицы содержит метки, а не данные;

б) в поле «Входной интервал» указать диапазон ячеек, содержащих наши данные. В нашем случае это вся таблица «Рост».

в) в поле «Выходной интервал» указать ячейку, куда будет вставлена таблица результатов.

Эти настройки позволят корректно обработать данные и построить график распределения частот:



Полученные результаты представлены в виде таблицы:

D	E
Столбец1	
Среднее	172,8231922
Стандартная ошибка	2,220339918
Медиана	172,1308455
Мода	#Н/Д
Стандартное отклонение	15,70017413
Дисперсия выборки	246,4954676
Эксцесс	0,295338424
Асимметричность	-0,410052748
Интервал	79,36996553
Минимум	127,0921847
Максимум	206,4621502
Сумма	8641,159609
Счет	50

4. Предположим, получены следующие данные:

Рост	
174	171
164	173
179	178
190	173
189	174
194	173
155	190
175	176
188	175
166	172
170	197
160	186
159	201
167	170
169	280
156	110

5. Проведем описательную статистику:

Столбец1	
Среднее	176,6875
Стандартная ошибка	4,407983387
Медиана	173,5
Мода	173
Стандартное отклонение	24,93531955
Дисперсия выборки	621,7701613
Эксцесс	10,36479266
Асимметричность	1,793173783
Интервал	170
Минимум	110
Максимум	280
Сумма	5654
Счет	32

6. Опишем, в чем разница по сравнению с прошлыми данными.

7. Из результатов обработки данных, представленных на графике, видно, что имеются высокие значения эксцесса и асимметрии. Это может свидетельствовать о наличии выбросов в данных. Эти выбросы могут быть представлены значениями 110 и 280 см. Чтобы проверить, являются ли эти значения артефактами, необходимо применить формулу

$$T = \frac{x_i - X}{s} \geq T_{st}$$

и таблицу определения выбросов:

n	T _{st}	n	T _{st}	n	T _{st}	n	T _{st}
2	2,0	16–20	2,4	47–66	2,8	125–174	3,2
3–4	2,1	21–28	2,5	67–84	2,9	175–349	3,3
5–9	2,2	29–34	2,6	85–104	3,0	350–599	3,4
10–15	2,3	35–46	2,7	105–124	3,1	600–1500	3,5

Если значения окажутся выбросами, они должны быть исключены из дальнейшего анализа.

8. Если эти данные являются артефактами, то необходимо провести повторно процедуры статистического описания данных, исключив данные артефакты.

9. Построим график распределения (гистограмму) и определим ширину карманов по формуле

$$k = \frac{X_{max} - X_{min}}{n},$$

где X_{max} – максимальное значение признака; X_{min} – минимальное значение признака; n – количество интервалов: $n = 1 + 3,322 \lg N$, где N – количество наблюдений.

Количество интервалов n округляется до ближайшего целого вверх. После вычисления ширины кармана определяется, какое количество значений признака попало в тот или иной карман. Для упрощения выполнения данной операции можно воспользоваться функцией «частота» программы MS Excel. Для этого необходимо рассчитать интервалы карманов, затем выделить напротив них ячейки, перейти в строку формул, ввести формулу частоты, которая будет включать наш набор чисел, а также диапазон карманов, затем последовательно нажать клавиши Ctrl, Shift и Enter.

Пример расчета с использованием вышеприведенных данных:

$$k = \frac{280 - 110}{6} = 28,3,$$

$$n = 1 + 3,322 \cdot \lg 32 = 6.$$

Округляем полученное значение 28,3 до ближайшего целого числа. Получаем 29. Однако стоит учесть, что сначала надо определить, являются ли артефактами значения 280 и 110. Если да, то производить расчет без них.

Напомним, что если не рассчитывать размеры интервалов, то они будут определены программой автоматически (как показано ниже) и данный шаг можно пропустить:

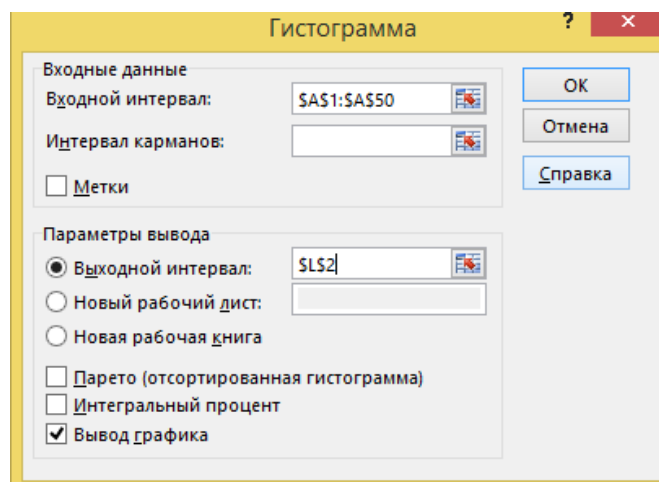
Карманы
150
160
170
180
190
200
210

Для создания данной таблицы на том же листе MS Excel нужно указать диапазоны ячеек, которые будут представлять интервалы карманов. Это можно сделать, введя соответствующие диапазоны ячеек в поле «Интервал карманов».

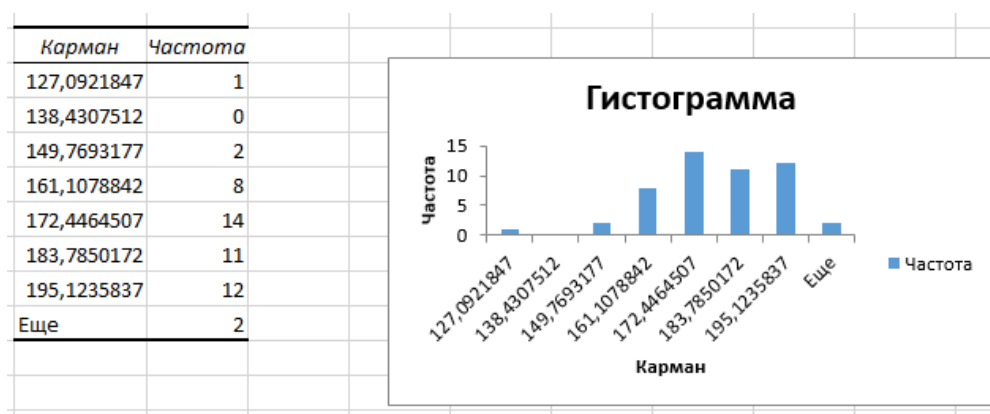
10. Откроем модуль «Анализ данных» во вкладке «Данные»:

- а) в открывшемся окне выберем «Гистограмма» и нажмем «ОК»;
- б) в появившемся диалоговом окне «Гистограмма» введем диапазон своих данных в поле «Входной интервал»;
- в) зададим остальные параметры гистограммы, если это необходимо;
- г) нажмем «ОК», чтобы построить гистограмму.

Повторим эти шаги для каждого из наших примеров:



11. Выполним установки, как показано на рисунке:



12. Построим полимодальную линию тренда.

13. Исправим названия осей на «Рост» (ось X) и «Частота встречаемости» (ось Y).

14. Оформим график. Название оси «Карман» переименуем в название измеряемого параметра – «Рост». Уберем значение «Еще» и заменим его на соответствующее нашей выборке. Если мы использовали таблицу карманов, описанную выше, и при этом появилось значение «Еще» и напротив него в поле «Частота» число 0, удалим обе ячейки.

Алгоритм анализа в JASP

1. Откроем программу JASP и загрузим следующие данные:

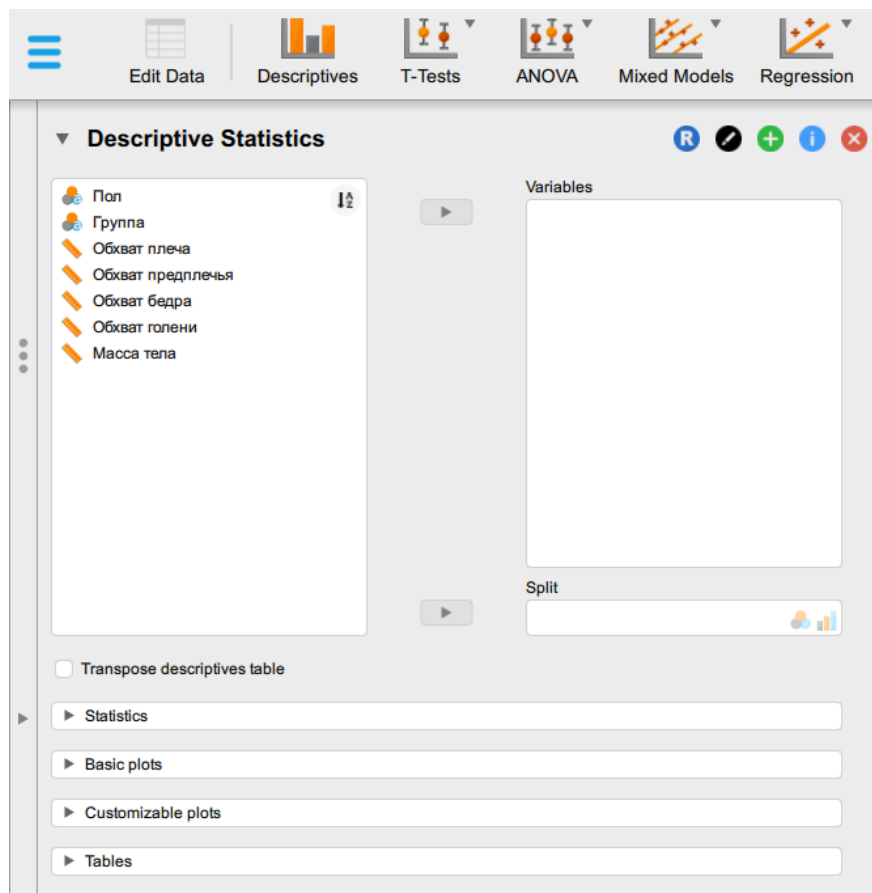
Пол	Группа	Обхват плеча	Обхват предплечья	Обхват бедра	Обхват голени	Масса тела
М	К1	22,8	24,2	48,8	34,0	71,3
М	К1	29,0	23,8	52,0	29,0	72,7
М	К1	28,8	23,6	49,8	31,9	74,8
М	К1	32,6	24,1	48,0	32,3	70,1
М	К1	29,4	23,9	49,0	32,7	74,5
М	К1	28,7	24,4	48,9	31,4	74,3
М	К1	28,6	23,2	48,7	34,4	74,9
М	К1	29,2	23,9	49,4	34,1	67,5
М	К1	30,2	23,2	49,0	30,8	64,5
М	К1	30,1	23,0	53,8	31,6	70,3
М	Э1	30,0	26,3	54,4	33,8	76,2
М	Э1	33,8	25,6	54,7	36,1	76,7
М	Э1	33,2	25,9	55,8	33,5	81,3
М	Э1	32,7	26,4	51,4	33,9	80,2
М	Э1	31,6	26,2	56,5	36,1	77,9
М	Э1	30,4	25,5	55,8	34,3	74,0
М	Э1	35,8	26,7	53,3	33,1	78,6

М	Э1	30,4	25,4	54,3	34,6	79,9
М	Э1	31,0	26,3	54,9	34,2	76,6
М	Э1	32,9	25,6	55,3	34,2	75,4
Ж	К2	24,3	19,4	50,2	27,2	52,8
Ж	К2	25,0	18,1	45,2	28,6	56,6
Ж	К2	24,9	20,0	45,4	29,4	52,3
Ж	К2	24,9	18,8	50,5	30,1	53,6
Ж	К2	25,5	19,3	44,3	29,7	57,9
Ж	К2	24,5	18,5	46,5	27,8	57,2
Ж	К2	25,4	18,9	48,9	26,8	56,5
Ж	К2	25,2	17,6	47,5	27,0	59,3
Ж	К2	25,1	18,1	44,9	29,0	56,4
Ж	К2	24,4	19,1	50,0	29,2	54,6
Ж	Э2	27,4	22,1	60,2	32,3	62,2
Ж	Э2	28,6	22,4	56,4	32,4	64,3
Ж	Э2	27,6	21,7	60,1	30,8	66,7
Ж	Э2	28,4	21,5	59,7	31,3	64,6
Ж	Э2	26,8	20,9	59,5	32,1	64,9
Ж	Э2	28,4	22,5	61,0	31,8	66,2
Ж	Э2	29,1	22,0	61,3	31,6	65,8
Ж	Э2	28,5	22,1	59,4	31,0	63,6
Ж	Э2	27,6	23,7	58,6	31,7	68,1
Ж	Э2	27,8	20,7	58,3	30,4	66,9

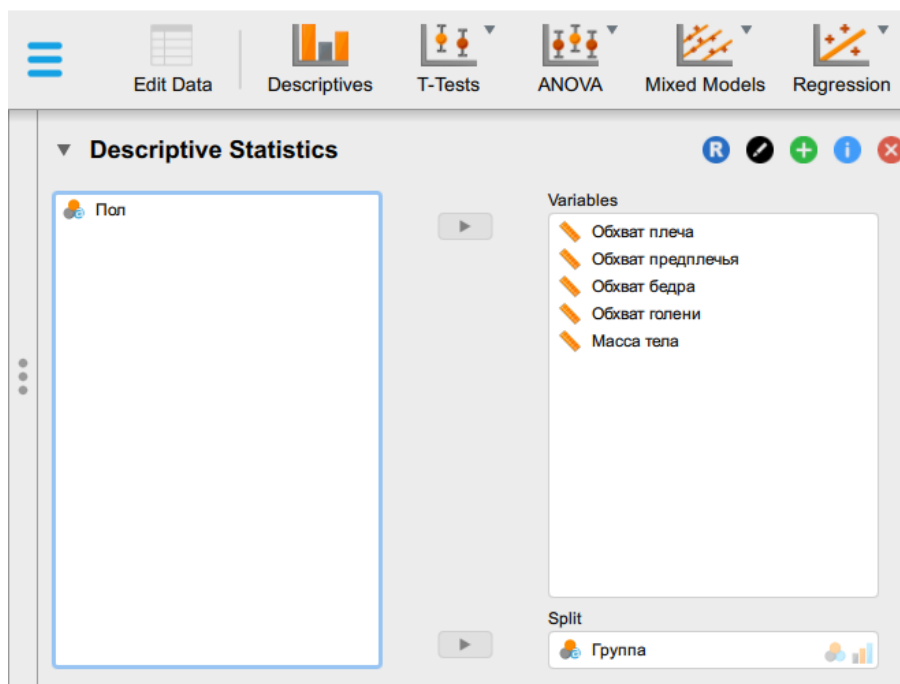
Данные введем в таблицу MS Excel. Затем полученный документ сохраним в формате ods (OpenDocument) через «Сохранить как». После этого откроем его в JASP:

	Пол	Группа	Обхват плеча	Обхват предплечья	Обхват бедра	Обхват голени	Масса тела
1	М	К1	22.83091927	24.24008955	48.77279083	33.96344104	71.29318926
2	М	К1	28.96013275	23.83154339	51.95456812	29.04140919	72.70755357
3	М	К1	28.77587004	23.64425206	49.76311755	31.89111438	74.75292075
4	М	К1	32.55094744	24.14792613	48.00895614	32.33727611	70.11177952
5	М	К1	29.43016394	23.90921037	48.99222382	32.74884406	74.54243184
6	М	К1	28.74405523	24.39317183	48.88546983	31.37186408	74.25141855
7	М	К1	28.61419917	23.23349378	48.68851216	34.41103044	74.93220944
8	М	К1	29.18463106	23.90019273	49.38213443	34.1113276	67.52307906
9	М	К1	30.21438347	23.20036589	48.95355677	30.76750212	64.47881385
10	М	К1	30.147273	22.96421559	53.84307384	31.6270343	70.26817281
11	М	Э1	29.95817985	26.3382072	54.37807854	33.75725348	76.21919971
12	М	Э1	33.83023712	25.60891614	54.71161833	36.08386518	76.70820227
13	М	Э1	33.19258493	25.85910381	55.82536576	33.51904897	81.3108571
14	М	Э1	32.73700085	26.42581291	51.38681584	33.90171011	80.23224402
15	М	Э1	31.6355964	26.20607683	56.51038205	36.06347613	77.87963816
16	М	Э1	30.41317504	25.51031316	55.78559606	34.27406168	74.03606296

2. Перейдем в модуль «Описательные статистики» («Descriptives»):



3. Перенесем переменные, как показано на рисунке:



4. В разделе «Statistics» зададим следующие параметры:

Statistics

Sample size

- ☒ Valid
- ☒ Missing

Central tendency

- ☒ Mode
- ☒ Median
- ☒ Mean

Dispersion

- ☒ Std. deviation
- ☒ Coefficient of variation
- ☒ MAD
- ☒ MAD robust
- ☒ IQR
- ☒ Variance
- ☒ Range
- ☒ Minimum
- ☒ Maximum

Inference

- ☒ S.E. mean
- ☒ Confidence interval for mean

Quantiles

- ☐ Quartiles
- ☐ Cut points for: 4 equal groups
- ☐ Percentiles:

Distribution

- ☒ Skewness
- ☒ Kurtosis
- ☒ Shapiro-Wilk test
- ☒ Sum

Выбранные статистики:

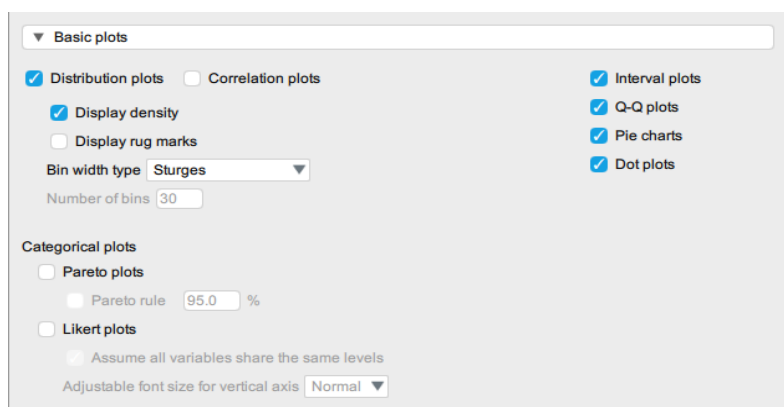
- **Valid** – количество случаев, имеющие значения;
- **Mode** – мода;
- **Median** – медиана;
- **Mean** – среднее;
- **Skewness** – асимметрия;
- **Kurtosis** – эксцесс;
- **Std. deviation** – стандартное отклонение;
- **S.E. mean** – стандартная ошибка среднего.

Другие статистики, которые могут использоваться при обработке данных:

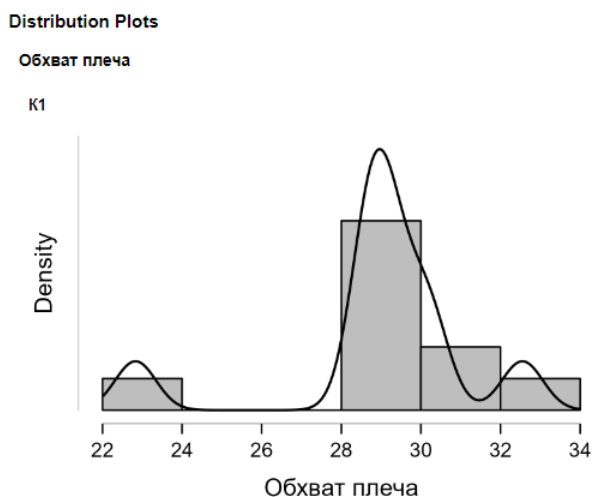
- **Missing** – количество случаев, не имеющих каких-либо значений;
- **Quartiles** – квартили;
- **Percentiles** – процентиля;
- **Shapiro–Wilk test** – критерий Шапиро–Уилка;
- **Sum** – сумма всех значений переменной;
- **Coefficient of variation** – коэффициент изменчивости выборки (вариации), меньшие значения которого свидетельствует о большей однородности или меньшей неоднородности выборки ($C. of v. = \text{стандартное отклонение} / \text{среднее арифметическое}$);
- **MAD** – медианное абсолютное отклонение (аналог стандартного отклонения);
- **MAD robust** – робастное (более устойчивое к параметрам распределения) медианное абсолютное отклонение;
- **IQR** – межквартильный размах;
- **Variance** – дисперсия;

- **Range** – размах;
- **Minimum** – минимальное значение переменной во всей выборке;
- **Maximum** – максимальное значение переменной во всей выборке;
- **Confidence interval for...** – доверительный интервал для соответствующей статистики (среднее, стандартное отклонение и дисперсия). Принято 95 %-е пороговое значение, которое говорит о том, что с 95 %-й вероятностью истинное значение статистики (в генеральной совокупности) находится в рассчитанном диапазоне.

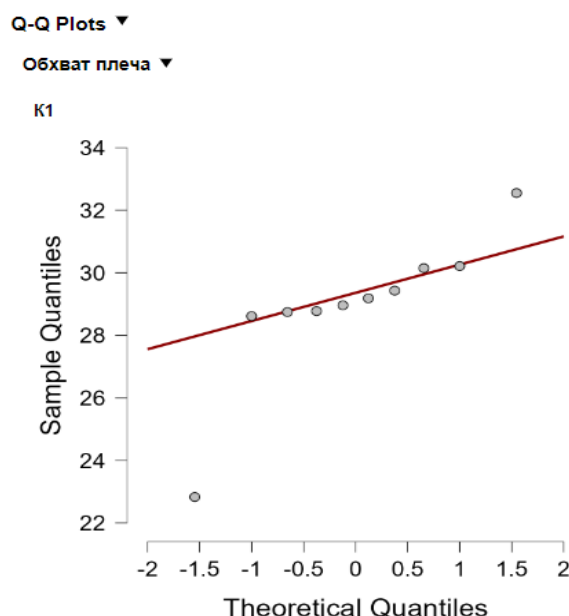
5. Построим диаграммы, используя следующие настройки:



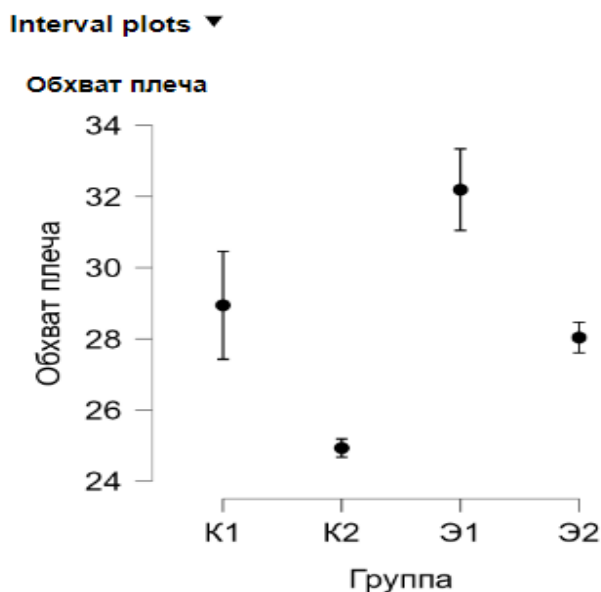
Distributions plots – диаграммы (графики) распределения в виде гистограмм. Дополнительно выбран пункт «Display density» (отобразить кривую распределения, которая рассчитывается только для количественных переменных). График распределения основан на разделении данных на интервалы частоты, которые затем накладываются на кривую распределения. Напомним, что самая высокая полоса – это *мода* (наиболее частое значение в наборе данных). Когда кривая выглядит примерно симметричной, это указывает на то, что данные примерно нормально распределены:



Q–Q plots (график квантиль – квантиль) можно использовать для визуальной оценки того, соответствует ли набор данных нормальному распределению. Графики Q–Q берут выборочные данные, сортируют их в порядке возрастания, а затем происходит визуализация. Если данные распределены нормально, точки будут находиться на опорной линии под углом 45° или близко к ней. Если данные распределены ненормально, точки будут отклоняться от опорной линии:

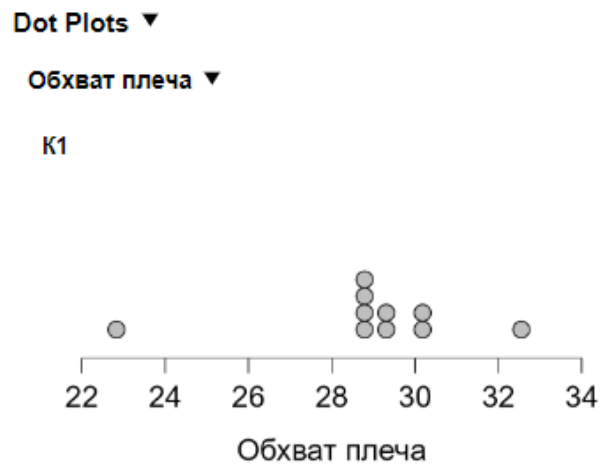


Interval plots – интервальные графики, представляющие размах и медиану на одной линии:



Такие графики строят для разных групп, так как для одной переменной он неинформативен.

Dot Plot – показывает распределение данных на числовой шкале:



6. Построим график «Boxplots», используя вкладку «Customizable plots»:

▼ Customizable plots

Color palette

Colorblind ▼

☒ Boxplots

☒ Boxplot element ☒ Use color palette
☐ Violin element ☐ Label outliers
☐ Jitter element

☒ Scatter plots

Graph above scatter plot

☒ Density
☐ Histogram
☐ None

Graph right of scatter plot

☐ Density
☐ Histogram
☐ None

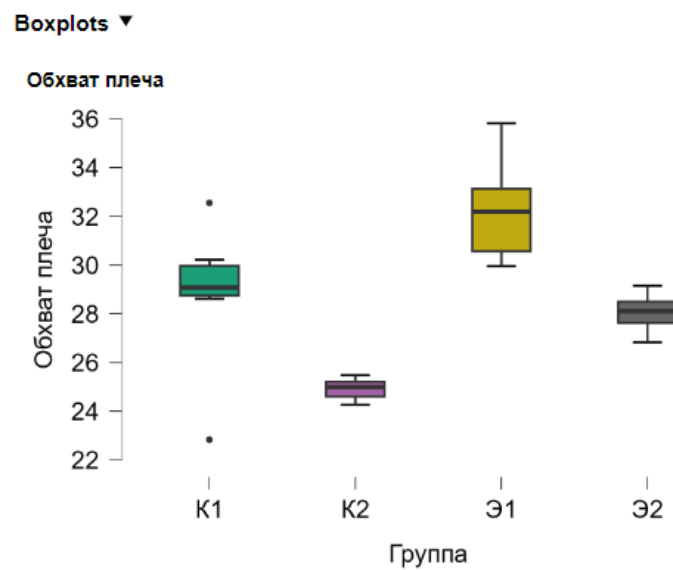
☒ Add regression line

☒ Show legend

☒ Smooth
☐ Linear

☒ Show confidence interval

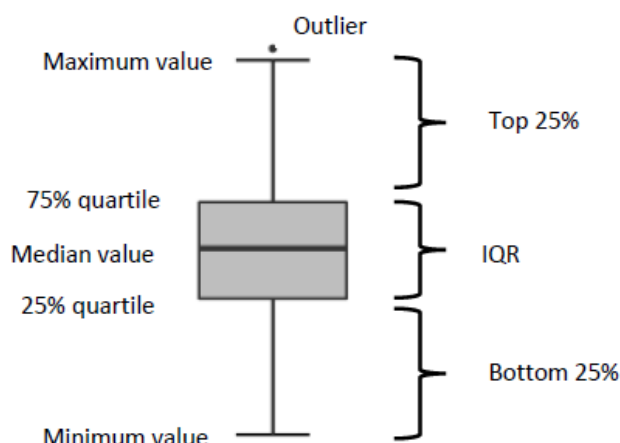
95.0 %



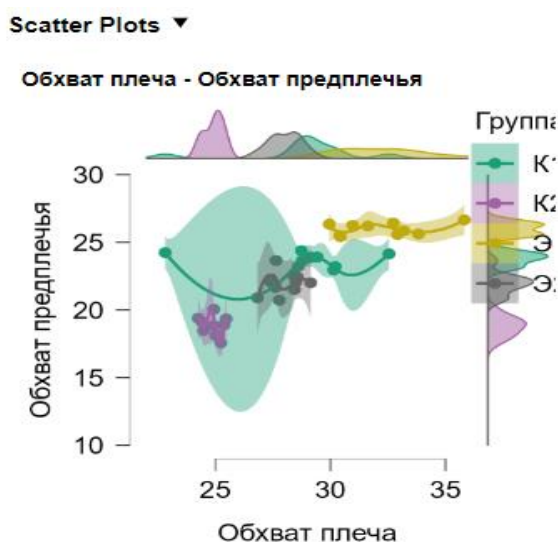
В JASP имеется несколько дополнительных опций (настройка цвета, вида элементов графика, показать номера выбросов (outlier)).

На рисунке ниже отображены следующие данные:

- медианное значение;
- 25 и 75 % квартилей;
- IQR – диапазон между квартилями;
- максимальные и минимальные значения нанесены на график с исключенными выбросами;
- выбросы отображаются по запросу:



Scatter plots – диаграммы рассеивания, характеризуют взаимосвязь двух переменных:



Выбранные пункты:

Pie charts – круговые диаграммы. Рассчитываются только для номинативных и ранговых переменных.

Frequency tables – таблицы частот. Рассчитывается только для текстовых, номинативных и ранговых переменных.

Остальные диаграммы и графики:

Distribution plots / Display rug marks – гистограммы с указанием засечек («ковра») на оси X. Засечки отражают расстояние между отдельными значениями переменной. Такие графики возможны только для количественных шкал.

Correlation plots – комплексная диаграмма, состоящая из гистограмм и диаграмм рассеивания. Представляет собой квадратную матрицу, где на пересечении переменных указываются диаграммы рассеивания, а по диагонали – гистограммы.

Pareto plots – гистограммы категориальных (группирующих) переменных в номинативной или ранговой шкале, на которых добавлены кумулятивные (накопительные) кривые, отражающие прирост количества случаев от одного значения к другому.

Likert plots – диаграммы Ликерта.

Stem and leaf tables – таблицы «стеблей и листьев» (подробнее: <https://www.mathsisfun.com/data/stem-leaf-plots.html>).

Задание

Произвести первичный анализ нижеприведенных в таблицах данных в программных продуктах MS Excel и JASP. Результаты оформить в виде отчета. Отчет должен содержать: исходные данные, полученные таблицы результатов, графики распределения, объяснение полученных результатов, выводы о типе распределения и о том, какими статистическими методами следует пользоваться при дальнейшем исследовании данной выборки.

1. Давление при операции на открытом сердце (галотановая анестезия):

Артериальное давление при галотановой анестезии	
42	66
44	68
46	68
46	68
48	68
48	70
50	70
50	70
52	70
54	70
54	72

56	72
58	72
58	72
58	74
60	74
60	74
60	74
60	76
60	78
62	78
62	78
62	80
62	80
62	82
64	82
64	84
66	86
66	90
66	94
66	98

2. Распределение возраста заболевших менингитом, вызванным гемофильной палочкой:

Возраст	
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	20
1	50
—	70

Контрольные вопросы

1. Параметры нормального распределения.
2. Какие методы статистического анализа используются при нормальном распределении?

3. Что характеризует дисперсия случайной величины?
4. Что характеризует стандартное отклонение?
5. Что такое среднеквадратичное отклонение?
6. Как описать ненормальное распределение?
7. Что такое медиана и процентиля?
8. Сколько переменных было в решаемых задачах?
9. При каком условии можно пользоваться методами параметрической статистики?
10. При каком условии используют методы непараметрической статистики?
11. Что такое артефакты?
12. Зачем необходима проверка на выбросы?
13. Что такое правило трех сигм?

2.3. Дисперсионный анализ

Краткие сведения из теории

Сравнение групп в статистике – это процесс, в котором сравниваются две или более группы данных, чтобы увидеть, есть ли между ними различия. Это может включать в себя сравнение средних значений, медиан, мод, стандартных отклонений, частот и других статистических показателей. Для сравнения групп в статистике используются различные методы, включая t-тест, ANOVA (анализ дисперсии), хи-квадрат тест и др. Выбор метода зависит от типа данных, количества групп и других факторов.

В исследованиях, где сравниваются группы, часто используется контрольная группа, которая служит для сравнения с экспериментальной группой. Контрольная группа, например, может получать стандартное лечение или плацебо, в то время как экспериментальная группа получает новый метод лечения или препарат. Сравнивая показатели между этими двумя группами, исследователь может сделать выводы о эффективности нового метода или препарата.

Гипотеза – это предположение или утверждение, которое требует проверки или доказательства. В статистике гипотеза представляет собой утверждение о свойствах генеральной совокупности, которое нужно проверить на основе выборки.

Гипотезы могут быть двух типов: нулевая гипотеза (H_0) и альтернативная гипотеза (H_1). Нулевая гипотеза обычно представляет собой утверждение о том, что нет различий между группами или нет связи между переменными. Альтернативная гипотеза представляет собой утверждение о наличии различий или связи.

Для проверки гипотезы используются статистические тесты, которые позволяют определить, следует ли отклонить нулевую гипотезу или нет. Если результаты теста указывают на то, что нулевая гипотеза должна быть отклонена, то альтернативная гипотеза принимается как правильная.

Важно отметить, что гипотезы должны быть четко сформулированы и основываться на данных и теории. Они также должны быть проверяемыми, т. е. иметь возможность быть подтвержденными или опровергнутыми на основе доступных данных.

Вероятность ошибки – это вероятность того, что при принятии решения на основе данных будет допущена ошибка. В статистике вероятность ошибки обычно связана с вероятностью отклонения нулевой гипотезы, когда она на самом деле верна (так называемая ошибка первого рода, или тип I), или с вероятностью принятия нулевой гипотезы, когда она на самом деле неверна (ошибка второго рода, или тип II).

Вероятность ошибки первого рода (α) – это вероятность того, что нулевая гипотеза будет отклонена, когда она на самом деле верна. Это также называется *уровнем значимости*. Обычно α устанавливается заранее и составляет 0,05 или 0,01.

Вероятность ошибки второго рода (β) – это вероятность того, что нулевая гипотеза будет принята, когда она на самом деле неверна. Это также называется *мощностью теста*. Мощность теста зависит от размера выборки, разницы между группами и других факторов.

Важно понимать, что невозможно полностью избежать ошибок, но можно минимизировать их вероятность, правильно формулируя гипотезы, выбирая подходящий статистический тест и учитывая размер выборки и другие факторы.

Дисперсия, вычисленная для каждой группы, является оценкой дисперсии совокупности. Таким образом, дисперсию совокупности можно оценить на основе групповых дисперсий. Эта оценка не зависит от различий групповых средних. В то же время разброс выборочных средних также позволяет оценить дисперсию совокупности. Эта оценка зависит от различий выборочных средних.

Если экспериментальные группы являются случайными выборками из одной и той же нормально распределенной совокупности, обе оценки дисперсии совокупности дадут примерно одинаковые результаты. Поэтому, если эти оценки близки, мы не можем отвергнуть нулевую гипотезу. В противном случае мы отвергаем нулевую гипотезу, т. е. заключаем: маловероятно, что мы получили бы такие различия между группами, если бы они были просто случайными выборками из одной нормально распределенной совокупности.

Если верна нулевая гипотеза, то как внутригрупповая, так и межгрупповая дисперсии, служат оценками одной и той же дисперсии и должны быть приближенно равны. Исходя из этого, вычисляется критерий F (критерий Фишера).

Числитель и знаменатель этого отношения – это оценки одной и той же величины – дисперсии совокупности σ^2 , поэтому значение F должно быть близко к 1.

Если F значительно превышает 1, нулевую гипотезу следует отвергнуть. Если же значение F близко к 1, нулевую гипотезу следует принять.

Если извлекать случайные выборки из нормально распределенной совокупности, значение F будет меняться от опыта к опыту.

Критерий, начиная с которого мы отвергаем нулевую гипотезу, называется *критическим значением*. Вероятность ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет, обозначается p . Обычно считается достаточным, чтобы эта вероятность не превышала 5 %. Критическое значение критерия F определяется уровнем значимости (обычно 0,05 или 0,01) и двумя параметрами, которые называются внутригрупповым и межгрупповым числом степеней свободы и обозначаются греческой буквой ν («ню»). *Межгрупповое число степеней свободы* – это число групп минус единица: $\nu_{\text{меж}} = m - 1$. *Внутригрупповое число степеней свободы* – это произведение числа групп на численность каждой из групп минус единица: $\nu_{\text{вну}} = m(n - 1)$. Критическое значение F можно вычислить, используя таблицы критических значений F для разных α , $\nu_{\text{меж}}$ и $\nu_{\text{вну}}$.

Величина критического значения критерия Фишера $F_{\text{кр}}$ определяется на основе уровня значимости и степеней свободы. Для определения статистической значимости различий между группами сравниваются F и $F_{\text{кр}}$, а также p и α . Если $F > F_{\text{кр}}$ и $p < \alpha$, то нулевая гипотеза отвергается

и различия между группами считаются статистически значимыми. В противном случае различия считаются случайными или незначимыми. Важно помнить, что даже при обнаружении статистической значимости различий существует вероятность ошибки, но эта вероятность невелика и равна уровню значимости.

Важно определить, какие данные мы хотим проанализировать. Это могут быть данные о группах, выборках или экспериментальных результатах.

Правила выполнения дисперсионного анализа:

1. Формирование исследуемых групп.
2. Получение результатов измерения.
3. Первичный анализ полученных данных, заключающийся в их статистическом описании.
4. Если данные в группах распределены по нормальному закону распределения (или близкому к таковому), то выбирается метод анализа, в данном случае – дисперсионный анализ.
5. Определение значения уровня значимости (в зависимости от строгости исследования: 0,1 или 0,05, или 0,01). Не стоит забывать об ошибках первого и второго рода.
6. Формулирование нулевой гипотезы.
7. Расчет критерия Фишера F и вероятности ошибочного результата p (вероятности ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет).
8. На основании значения уровня значимости и степеней свободы определение величины критического значения критерия Фишера $F_{кр}$ (значения критерия, начиная с которого мы отвергаем нулевую гипотезу). Сравнение между собой F и $F_{кр}$, а также p и α . Если $F > F_{кр}$ и $p < \alpha$, то нулевая гипотеза отвергается и различия между группами статистически значимы. В противном случае различия случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность равна уровню значимости (т. е. незначительна).

Задача

Низкий уровень холестерина липопротеидов высокой плотности (ХЛПВП) – фактор риска ишемической болезни сердца. Некоторые исследования свидетельствуют, что физическая нагрузка (анализируемый

фактор) может повысить уровень ХЛПВП. Ученые исследовали уровень ХЛПВП у бегунов-марафонцев, бегунов трусцой и лиц, не занимающихся спортом:

Лица, не занимающиеся спортом, %	Бегуны трусцой, %	Бегуны-марафонцы, %
32,8	35,5	83,0
60,9	68,1	50,1
45,7	49,5	36,0
61,1	52,9	73,9
38,4	56,7	63,7
46,0	37,0	67,3
47,8	62,8	46,3
35,0	87,3	60,5
58,4	66,8	58,6
31,4	23,8	59,1
37,7	45,0	62,1
47,5	76,4	53,8
40,9	41,8	42,2
39,1	56,3	59,0
74,4	73,5	51,7
37,5	67,8	81,2
51,3	37,1	78,4
13,1	44,7	61,4
27,5	53,0	51,0
39,1	57,9	63,4
36,2	70,0	50,9
46,2	64,1	77,4
61,1	53,6	66,1
33,9	48,4	45,8
49,8	43,9	68,9
31,8	64,2	78,9
57,6	54,1	53,9
57,1	65,2	48,3
60,4	60,7	41,8
14,6	52,3	62,6

Будем считать, что в каждой группе было по 30 человек. Оценим статистическую значимость различий между группами с помощью однофакторного дисперсионного анализа.

Алгоритм анализа в MS Excel

1. Сформулируем нулевую гипотезу: разные виды бега не влияют на уровень ХЛПВП, различия в группах случайны и статистически не значимы.

2. Скопируем таблицу с данными в MS Excel:

	A	B	C
1	Уровень ХЛПВП, %		
2	Лица, не занимающиеся спортом	Бегуны трусой	Марафонцы
3	32,8	35,5	83,0
4	60,9	68,1	50,1
5	45,7	49,5	36,0
6	61,1	52,9	73,9
7	38,4	56,7	63,7
8	46,0	37,0	67,3
9	47,8	62,8	46,3
10	35,0	87,3	60,5
11	58,4	66,8	58,6
12	31,4	23,8	59,1
13	37,7	45,0	62,1
14	47,5	76,4	53,8
15	40,9	41,8	42,2
16	39,1	56,3	59,0
17	74,4	73,5	51,7
18	37,5	67,8	81,2
19	51,3	37,1	78,4
20	13,1	44,7	61,4
21	27,5	53,0	51,0
22	39,1	57,9	63,4
23	36,2	70,0	50,9

3. С помощью модуля «Описательная статистика» (который нам уже знаком из предыдущих задач) рассчитаем основные статистические параметры (среднее значение, стандартное отклонение, дисперсию) для каждой группы исходных данных. Получим таблицу:

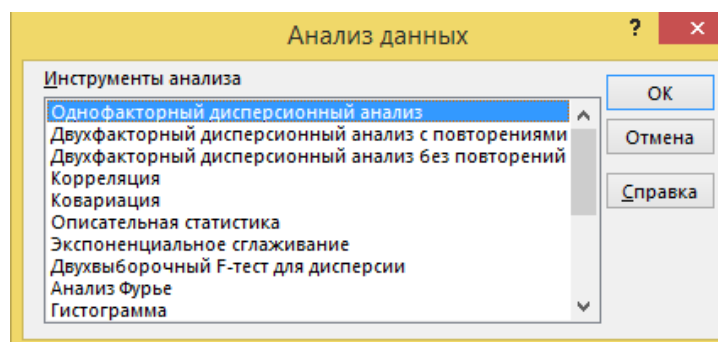
E	F	G	H	I	J
Столбец1		Столбец2		Столбец3	
Среднее	43,81315	Среднее	55,66673	Среднее	59,90786773
Стандартная ошибка	2,55655	Стандартная ошибка	2,514842	Стандартная ошибка	2,28376966
Медиана	43,29174	Медиана	55,19033	Медиана	59,78287627
Мода	#Н/Д	Мода	#Н/Д	Мода	#Н/Д
Стандартное отклонение	14,0028	Стандартное отклонение	13,77436	Стандартное отклонение	12,50872159
Дисперсия выборки	196,0784	Дисперсия выборки	189,733	Дисперсия выборки	156,4681158
Эксцесс	0,078278	Эксцесс	0,125075	Эксцесс	-0,678311602
Асимметричность	-0,1183	Асимметричность	-0,05809	Асимметричность	0,174531823
Интервал	61,24391	Интервал	63,49749	Интервал	47,04110506
Минимум	13,12219	Минимум	23,81916	Минимум	35,98284591
Максимум	74,36609	Максимум	87,31665	Максимум	83,02395097
Сумма	1314,395	Сумма	1670,002	Сумма	1797,236032
Счет	30	Счет	30	Счет	30

Затем сравним средние значения каждой группы и определим цель дисперсионного анализа.

Средние значения отличаются по величине, поэтому необходимо провести дисперсионный анализ, чтобы проверить статистическую значимость этих различий. Основная гипотеза заключается в том, что различия средних значений не являются статистически значимыми.

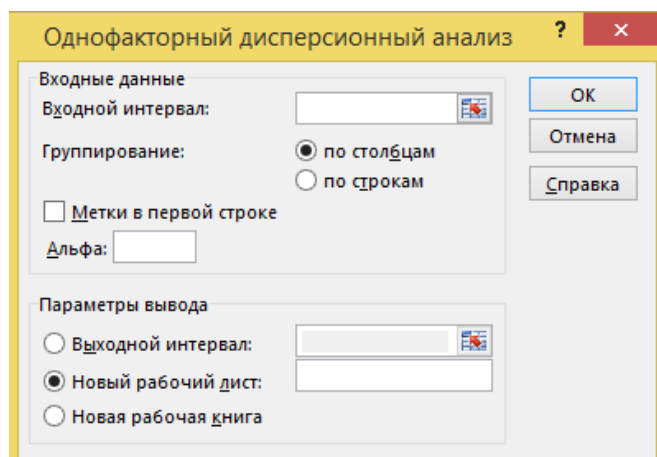
Важно отметить, что перед проведением дисперсионного анализа необходимо убедиться в нормальности распределения значений в обеих группах. Для этого можно построить график распределения. Небольшие отклонения от нормального распределения допустимы. Кроме того, величина дисперсии в первой и второй группах должна быть примерно одинаковой.

4. Так как в задаче проверяется влияние одного фактора (влияет ли спорт на уровень ХЛПВП), то необходимо воспользоваться однофакторным дисперсионным анализом:



Дисперсионный анализ в MS Excel проводится следующим образом:

- а) открыть вкладку «Данные»;
- б) в группе «Анализ данных» выбрать «Однофакторный дисперсионный анализ»;
- в) в открывшемся диалоговом окне «Однофакторный дисперсионный анализ» ввести диапазон ячеек, содержащих наши данные;
- г) нажать «ОК»:



5. Зададим входной интервал – данные 1, 2 и 3-й групп (выделяются при помощи мыши, с заголовками), поставим галочку «Метки в первой строке», далее «Альфа – 0,05» и «Выходной интервал» – им может быть любая свободная ячейка:

Результаты дисперсионного анализа представлены ниже. Интересующие значения оцениваемых параметров выделены маркером:

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	Счет	Сумма	Среднее	Дисперсия		
Столбец 1	30	1314,3946	43,81315	196,0784007		
Столбец 2	30	1670,001788	55,66673	189,7329627		
Столбец 3	30	1797,236032	59,90787	156,4681158		
Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	4175,343	2	2087,672	11,549422	3,56072E-05	3,101295757
Внутри групп	15726,1	87	180,7598			
Итого	19901,45	89				

В рассмотренном примере эмпирический F-критерий (критерий Фишера) показывает, что различие между средними статистически значимо (значимо на уровне $p = 0,000036$, меньше, чем критическое значение 0,05).

Вывод: так как расчетное значение критерия Фишера F больше его критического значения $F_{кр}$ при уровне значимости $\alpha = 0,05$, то нулевая гипотеза отвергается. Вероятность ошибки p меньше уровня значимости. Таким образом, занятия спортом влияют на уровень ХЛПВП.

Алгоритм анализа в JASP *для независимых выборок*

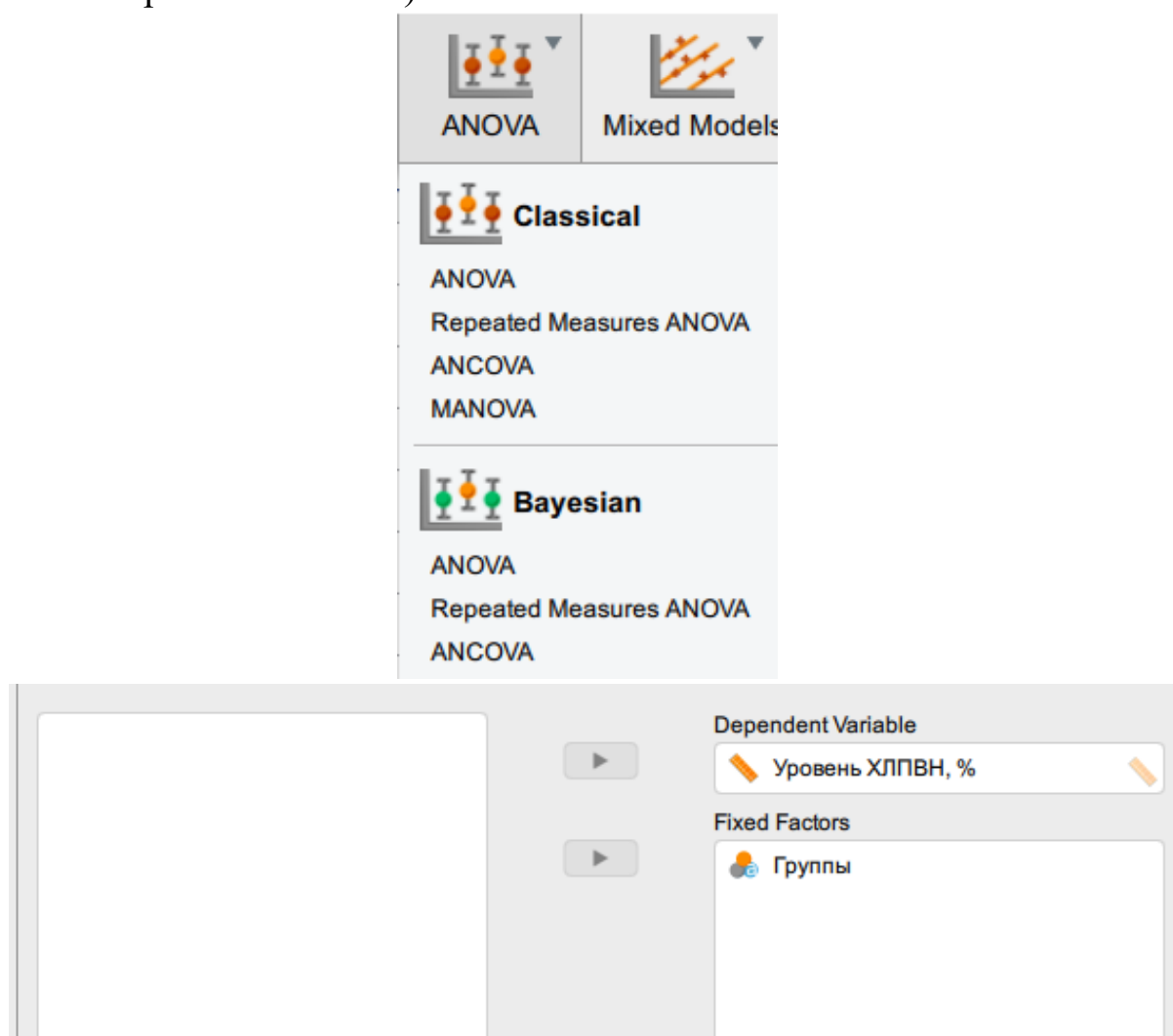
1. Excel-таблицу предварительно преобразуем в вид, представленный ниже, где а, b, с – предикторы для каждой группы: а – группа «Лица, занимающиеся спортом, %», b – группа «Бегуны трусдой, %», с – группа «Бегуны-марафонцы»:

	A	B	C
1	a	32,8	
2	a	60,9	
3	a	45,7	
4	a	61,1	
5	a	38,4	
6	a	46	
7	a	47,8	
8	a	35	
9	a	58,4	
10	a	31,4	
11	a	37,7	
12	a	47,5	
13	a	40,9	
14	a	39,1	
15	a	74,4	
16	a	37,5	
17	a	51,3	
18	a	13,1	
19	a	27,5	
20	a	39,1	
21	a	36,2	
22	a	46,2	
23	a	61,1	
24	a	33,9	
25	a	49,8	
26	a	31,8	
27	a	57,6	
28	a	57,1	
29	a	60,4	
30	a	14,6	
31	b	35,5	
32	b	68,1	
33	b	49,5	
34	b	52,9	
35	b	56,7	
36	b	37	
37	b	62,8	
38	b	87,3	
...			

2. Загружаем данную таблицу в JASP, предварительно сохранив ее в формате ods, либо копируем ее с листа Excel, создаем новую таблицу в JASP и копируем данные в неё:

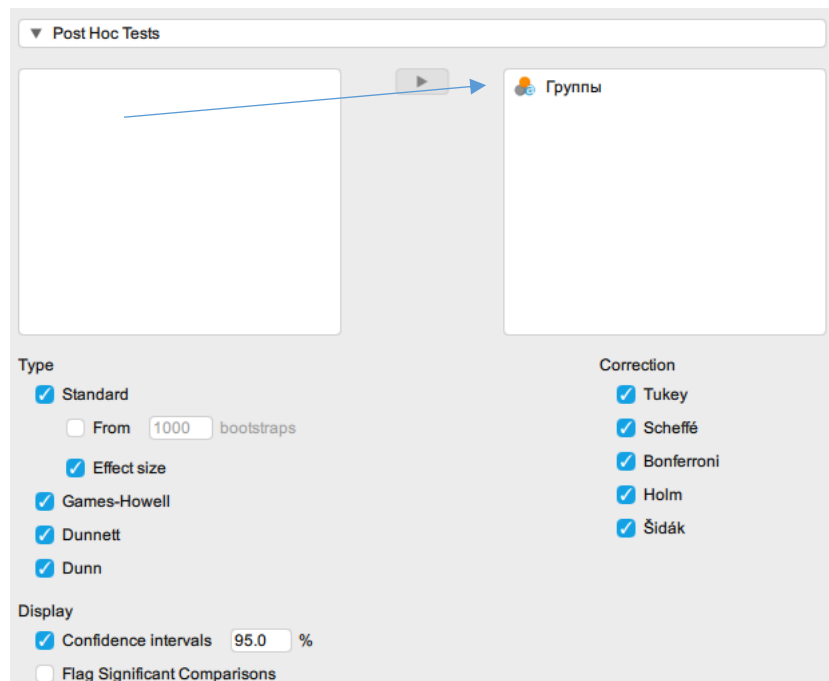
	Группы	Уровень ХЛПВН, %
1	a	32.8
2	a	60.9
3	a	45.7
4	a	61.1
5	a	38.4
6	a	46
7	a	47.8
8	a	35
9	a	58.4
10	a	31.4
11	a	37.7
12	a	47.5
13	a	40.9
14	a	39.1
15	a	74.4
16	a	37.5
17	a	51.3
18	a	13.1
19	a	27.5
20	a	39.1

3. Выбираем модуль ANOVA. Для этого нажимаем на модуль «Дисперсионный анализ» (ANOVA) и в выпадающем списке выбираем «Дисперсионный анализ» (ANOVA). Переносим переменные, как показано на рисунке ниже. В меньшее окно «Dependent Variable» нужно перенести зависимую переменную, а в большее окно «Fixed Factors» перенести факторную переменную, имеющую несколько категорий (номинативная или ранговая шкала):



В данном модуле есть дополнительные виды анализа: дисперсионный анализ с повторными измерениями (аналог сравнения зависимых групп, но для случаев, когда групп более 2); ANCOVA – дисперсионный анализ с использованием факторов в количественной шкале (они носят название «ковариаты»). Двухфакторный дисперсионный анализ возможен в первом меню модуля «ANOVA». Разбор данных видов анализа не предусмотрен в настоящем пособии.

4. Задаем следующие параметры в разделе «Post Hoc Tests», не забывая перенести группирующую переменную, как указано на рисунке:



Обозначения на рисунке:

- **Standard** – критерий Стьюдента;
- **Effect size** – размер эффекта (d Коэна по умолчанию);
- **Games–Howell** – критерий Геймс–Хоулла, непараметрический аналог стандартного t -критерия;
- **Dunnett** – критерий Даннета, параметрический аналог t -критерия (используется при наличии контрольной группы среди других групп);
- **Dunn** – критерий Данна, непараметрический аналог t -критерия;
- **Correction** – коррекция уровня значимости (имеет несколько вариантов; можно выбрать любой, так как для психологических исследований их разница в расчетах несущественна);
- **Confidence intervals** – доверительные интервалы, которые рассчитываются для соответствующих статистик апостериорного сравнения средних.

Используемые пункты меню анализа:

- **Descriptive statistics** – описательные статистики для каждой из категорий факторной переменной;
- **Estimates of effect size** – оценка размера эффекта для F -критерия.

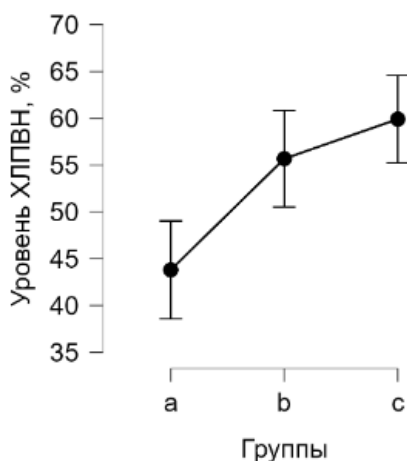
В JASP представлены три варианта:

- квадратичная «эта» (η^2);

- частичный квадрат «эты» (*partial* η^2);
- квадратичная «омега» (ω^2);
- **Homogeneity tests** – критерий Левена;
- **Homogeneity corrections** – коррекции однородности (гомогенности), осуществляется в случае неравенства дисперсий групп, имеет три варианта: None – отсутствие коррекции, коррекция Браун–Форсайта (Brown–Forsythe) и Уэлча (Welch); при выборе одного или нескольких коррекций F-критерий рассчитывается с учетом нарушения допущения об однородности групп факторной переменной;
- **Q–Q plot of residual** – квантиль–квантиль график остатков (остатки рассчитываются для каждого случая после вычета факторной дисперсии).

5. Если распределение ненормальное, проводим непараметрический тест Крускала–Уоллиса.

6. Построим график:



▼ Descriptives Plots

Factors

▶

▶

▶

Horizontal Axis

▶

Группы

Separate Lines

Separate Plots

Display

☒ Display error bars

☒ Confidence interval
95.0 %

☐ Standard error

На данном графике изображены средние арифметические (точки на вертикальных линиях) для каждой из групп и доверительные интервалы среднего значения для каждой из групп.

Главные результаты проверки гипотезы о различиях в группах представлены в таблице:

ANOVA - Column 2

Cases	Sum of Squares	df	Mean Square	F	p
Группы	4179.998	2	2089.999	11.559	< .001
Residuals	15730.742	87	180.813		

Note. Type III Sum of Squares

Столбец Sum of Squares содержит значения меры изменчивости для факторной переменной и для остаточной (общая изменчивость минус изменчивость, обусловленная фактором); df – степень свободы; Mean Square – дисперсия; F – критерий Фишера, p – уровень значимости.

Для оценки степени (силы) различий между группами с помощью размера эффекта необходимо использовать соответствующую таблицу.

Для интерпретации результатов используются эмпирическое значение статистического критерия, уровень значимости и размер эффекта.

При $p < 0,05$ принимается альтернативная гипотеза о наличии различий между группами.

7. Проведем тест Фишера с учетом гомогенности дисперсий у групп:

Assumption Checks							
☒ Homogeneity tests							
Homogeneity corrections							
☒ None ☒ Brown-Forsythe ☒ Welch							
☒ Q-Q plot of residuals							
ANOVA - Уровень ХЛПВН, %							
Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p	η^2
None	Группы	4179.998	2.000	2089.999	11.559	< .001	0.210
	Residuals	15730.742	87.000	180.813			
Brown-Forsythe	Группы	4179.998	2.000	2089.999	11.559	< .001	0.210
	Residuals	15730.742	86.202	182.486			
Welch	Группы	4179.998	2.000	2089.999	11.352	< .001	0.210
	Residuals	15730.742	57.848	271.933			

Note. Type III Sum of Squares

Homogeneity corrections – коррекции однородности (гомогенности). Коррекция осуществляется в случае неравенства дисперсий групп. Существуют три варианта: None – отсутствие коррекции, коррекция Браун–Форсайта (Brown–Forsythe) и Уэлча (Welch). При выборе одного или нескольких вариантов коррекции F-критерий рассчитывается с учетом нарушения допущения об однородности групп факторной переменной. Основная таблица ANOVA показывает, что F-статистика значима ($p < 0,001$) и существует большой размер эффекта. Таким образом, существует значительная разница между группами.

Проведем тест Левена на гомогенность депрессии:

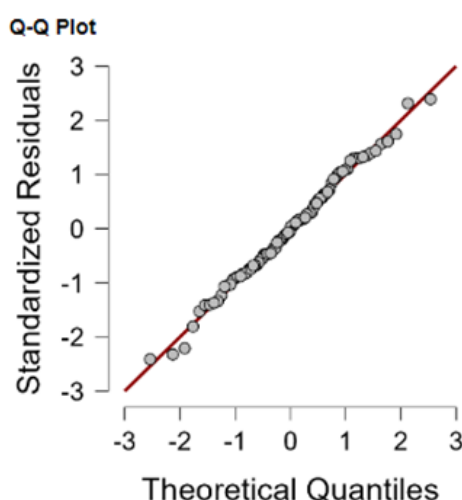
Assumption Checks ▼

Test for Equality of Variances (Levene's)

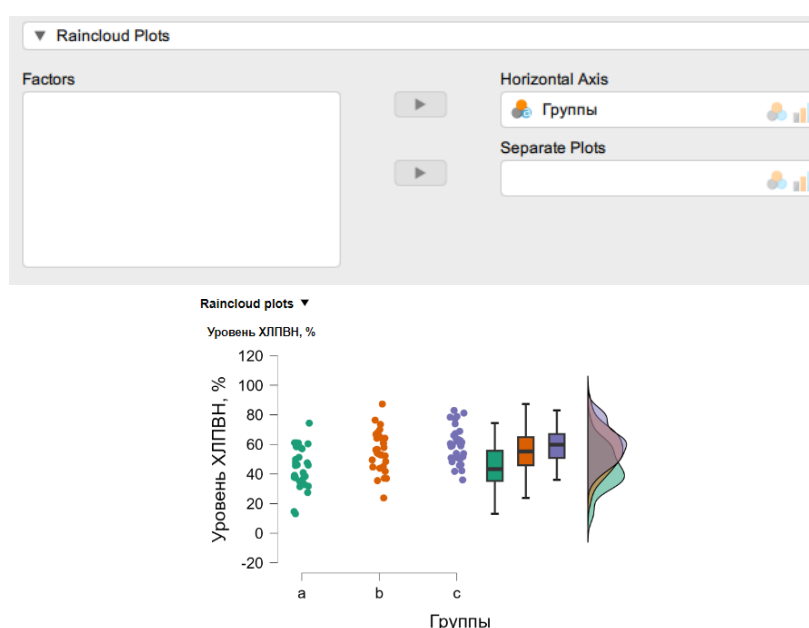
F	df1	df2	p
0.182	2.000	87.000	0.834

Результаты теста Левена показывают, что однородность дисперсии не имеет существенного значения. Однако если критерий Левена обнаруживает значительную разницу в дисперсии, следует указать поправку Брауна–Форсайта или Уэлча. Поскольку приведенное в таблице $p \geq 0,05$, то мы не можем отвергнуть нулевую гипотезу. Другими словами, все три группы имеют равные дисперсии.

График Q–Q показывает, что данные выглядят нормально распределенными и линейными:



8. Построим график «Raincloud plots»:



Теперь проведем дисперсионный анализ в JASP для *зависимых* выборок на примере решения следующей задачи.

Задача

В медицинском исследовании сравнивали эффективность нового метода лечения артериальной гипертензии (АГ). У пациентов сначала измеряли артериальное давление (контроль), затем снимали показатели через 24 часа и через три дня после лечения соответственно. Эффективность лечения оценивалась по уровню систолического артериального давления. Данные представлены в таблице. Необходимо определить, есть ли статистические различия между группами:

Участник	Контроль	Через 24 часа	Через 3 дня
1	145	165	142
2	167	152	127
3	146	135	134
4	166	143	152
5	154	138	133
6	171	143	135
7	158	140	149
8	133	143	143
9	154	128	138
10	147	146	143
11	171	143	136
12	150	147	126
13	150	156	146
14	136	162	143
15	163	135	147
16	142	136	131
17	151	141	141

18	152	158	128
19	147	140	130
20	159	152	122
21	143	128	129
22	169	150	134
23	165	123	130
24	140	156	121
25	138	144	158
26	162	151	123
27	152	131	134
28	149	123	121
29	150	138	130
30	158	143	140
31	157	155	133
32	140	138	142
33	150	143	133
34	165	153	139
35	147	137	144
36	156	148	152
37	168	149	141
38	158	155	162
39	146	148	136
40	152	147	151

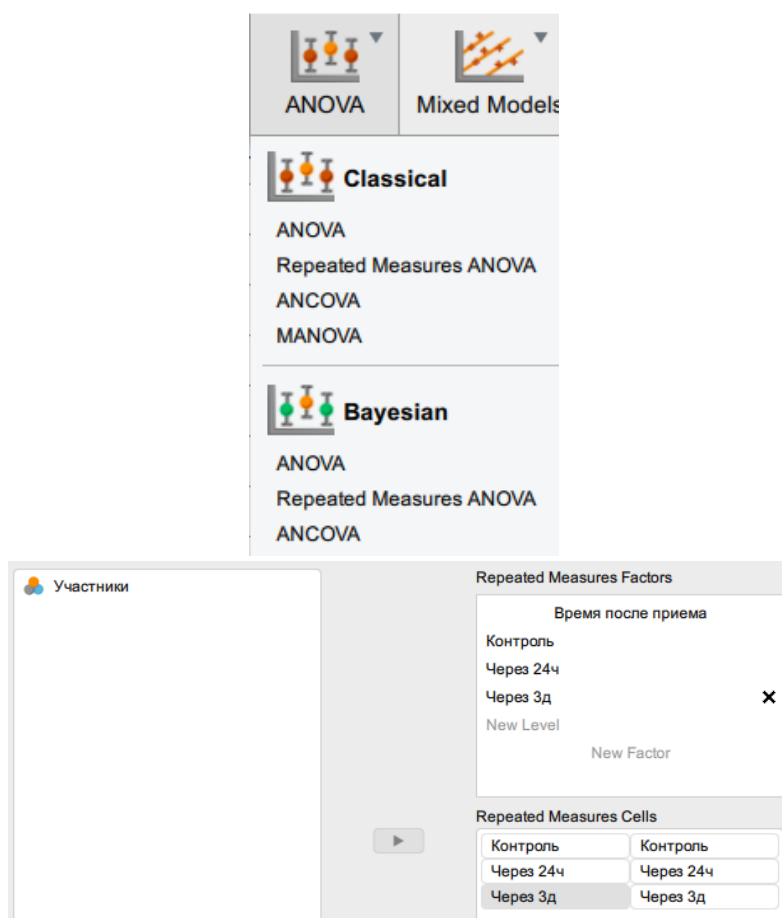
Алгоритм анализа JASP для зависимых выборок

1. Загрузим соответствующий документ в JASP, предварительно сохранив его в формат ods, либо скопируем его с листа MS Excel, создадим новую таблицу в JASP и скопируем данные в неё:

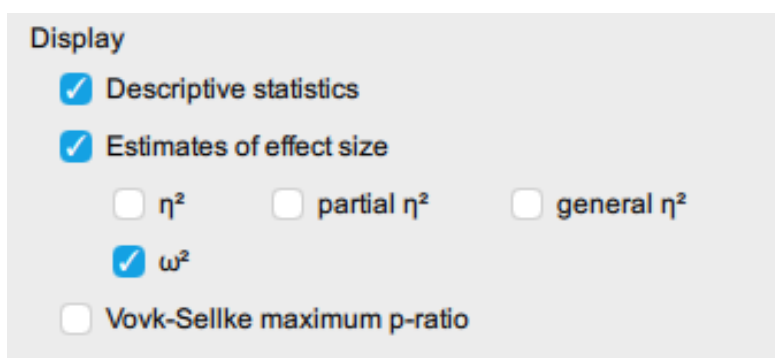
	 Участники	 Контроль	 Через 24 часа	 Через 3 дня
1	1	145	165	142
2	2	167	152	127
3	3	146	135	134
4	4	166	143	152
5	5	154	138	133
6	6	171	143	135
7	7	158	140	149
8	8	133	143	143
9	9	154	128	138
10	10	147	146	143
11	11	171	143	136
12	12	150	147	126

2. Выберем модуль ANOVA. Для этого нажмем на модуль «Дисперсионный анализ» (ANOVA) и в выпадающем списке выберем «Repeated Measures ANOVA». Переименуем «RM Factor 1» в «Время после

приема». Добавим третий уровень и дадим исследуемым группам соответствующие названия. Перенесем переменные в пустые ячейки в разделе «Repeated Measures Cells», как показано на рисунке:



3. Зададим следующие настройки:



Примечание. JASP обеспечивает те же альтернативные расчеты величины эффекта, которые используются с независимой группой.

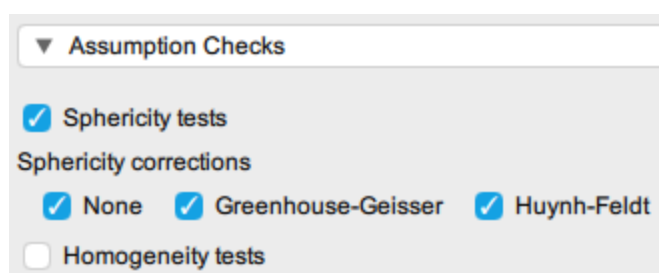
Квадратичная «эта» (η^2) – этот параметр наиболее точен для объясненной выборочной дисперсии, но переоценивает популяционную дисперсию; может затруднить сравнение эффекта одной переменной в разных исследованиях.

Частичный квадрат «эты» ($\partial\eta^2$) – решает проблему, связанную с завышением дисперсии генеральной совокупности; позволяет сравнивать влияние одной и той же переменной в разных исследованиях. По-видимому, это наиболее часто сообщаемый размер эффекта при повторных измерениях ANOVA.

Квадратичная «омега» (ω^2) – обычно статистическая погрешность становится очень малой по мере увеличения размера выборки, но для небольших выборок ($n < 30$) ω^2 обеспечивает несмещенную меру размера эффекта.

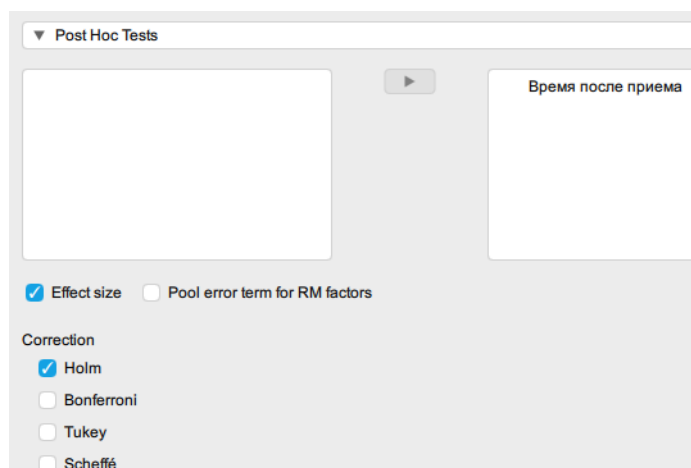
RMANOVA проверяет предположение о сферичности с использованием критерия сферичности Мокли. Это проверяет нулевую гипотезу о том, что дисперсии различий равны. Во многих случаях повторные измерения нарушают предположение о сферичности, что может привести к ошибке первого рода. В этом случае можно внести поправки в F-статистику.

JASP предлагает два метода коррекции F-статистики: Гринхауса–Гейссера и Хюйна–Фельдта:



Рассмотрим поправки эpsilon (ϵ). Общее практическое правило заключается в том, что, если $\epsilon < 0,75$, то необходимо использовать поправку Гринхауса–Гейссера, а если $\epsilon > 0,75$ – поправку Хюйна–Фельдта.

Апостериорное тестирование ограничено в RMANOVA, поэтому JASP предоставляет две альтернативы: Bonferroni и Holm.



Поправка Бонферрони (Bonferroni) может быть очень консервативной, но дает гарантированный контроль над ошибкой первого рода с риском снижения статистической мощности.

Тест Холма–Бонферрони (Holm) представляет собой последовательный метод Бонферрони, который менее консервативен, чем исходный тест Бонферрони. Если запросить апостериорные исправления Тьюки (Tukey) или Шеффе (Scheffe), JASP вернет ошибку NaN (а не число).

4. Получим следующие результаты:

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	ω^2
Время после приема	None	4971.467	2.000	2485.733	26.164	< .001	0.287
	Greenhouse-Geisser	4971.467	1.970	2523.261	26.164	< .001	0.287
	Huynh-Feldt	4971.467	2.000	2485.733	26.164	< .001	0.287
Residuals	None	7410.533	78.000	95.007			
	Greenhouse-Geisser	7410.533	76.840	96.441			
	Huynh-Feldt	7410.533	80.897	91.605			

Note. Type III Sum of Squares

В данной таблице представлены главные результаты проверки гипотезы о различиях в группах.

Столбец «Sum of Squares» содержит значения меры изменчивости для факторной переменной и для остаточной (общая изменчивость минус изменчивость, обусловленная фактором); df – степень свободы; Mean Square – дисперсия; F – критерий Фишера, p – уровень значимости.

Для оценки степени (силы) различий между группами с помощью размера эффекта необходимо использовать соответствующую таблицу.

Для интерпретации данных используются эмпирическое значение статистического критерия, уровень значимости и размер эффекта.

При $p < 0,05$ принимается альтернативная гипотеза о наличии различий между группами.

Рассмотрим описательную статистику ранее полученных результатов:

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Residuals	4191.992	39	107.487		

Note. Type III Sum of Squares

Descriptives

Descriptives	N	Mean	SD	SE	Coefficient of variation
Время после приема					
Контроль	40	153.175	10.002	1.581	0.065
Через 24ч	40	144.075	9.944	1.572	0.069
Через 3д	40	137.475	9.928	1.570	0.072

Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. χ^2	df	p-value	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ	Lower Bound ϵ
Время после приема	0.985	0.578	2	0.749	0.985	1.000	0.500

Описательная статистика (Descriptives) показывает количество участников в группах (N), среднее значение (Mean), стандартное отклонение (SD), стандартную ошибку среднего (SE).

Если дисперсионный анализ статистически значим, то можно провести апостериорное тестирование. Данное тестирование показывает, есть ли различия между конкретными группами:

Post Hoc Tests

Post Hoc Comparisons - Время после приема

		Mean Difference	SE	t	Cohen's d	P_{holm}
Контроль	Через 24ч	9.100	2.218	4.102	0.914	< .001
	Через 3д	15.700	2.269	6.919	1.577	< .001
Через 24ч	Через 3д	6.600	2.045	3.227	0.663	0.003

Note. Computation of Cohen's d based on pooled error.

Note. P-value adjusted for comparing a family of 3

5. Если распределение ненормальное, проводим непараметрический тест Фридмана:

Nonparametrics

Factors

RM Factor: Время после приема

Optional Grouping Factor

☐ Conover's post hoc tests

Nonparametrics

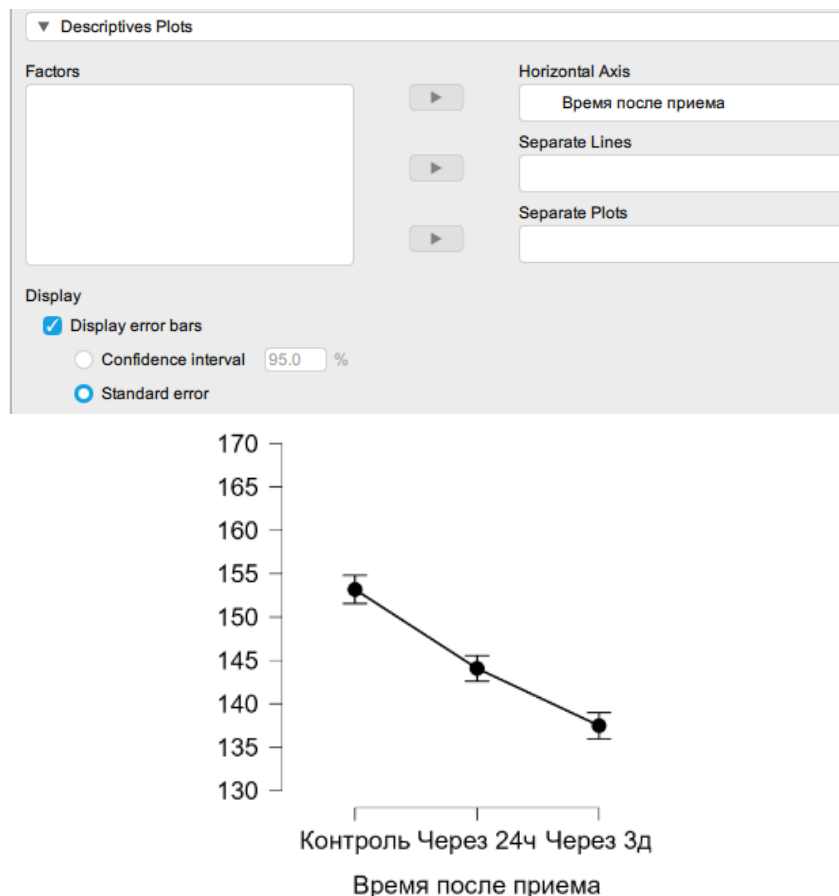
Friedman Test

Factor	Chi-Squared	df	p	Kendall's W
Время после приема	30.532	2	< .001	0.382

Если параметрические предположения нарушаются или данные являются порядковыми, следует рассмотреть возможность использования непараметрической альтернативы – критерия Фридмана. Подобно критерию Крускала–Уоллиса, критерий Фридмана используется для одностороннего анализа дисперсии с повторными измерениями по рангам, и он

не предполагает, что данные имеют нормальное распределение. Тест Фридмана представляет собой еще один комплексный тест, в котором не указывается, какие конкретно группы независимой переменной статистически значимо отличаются друг от друга. Для этого JASP предоставляет возможность запуска апостериорного теста Коновера, если результаты теста Фридмана значимы.

6. Построим график:



На данном графике изображены средние арифметические значения (точки на вертикальных линиях) для каждой из групп и стандартная ошибка среднего.

Задание

Проведите дисперсионный анализ в MS Excel и JASP для следующей задачи.

Галотан – эффективный препарат для общей анестезии, который удобно применять и контролировать. Он вводится через респиратор и быстро действует, что позволяет оперативно управлять анестезией. Однако одним из его недостатков является снижение артериального давления, вызванное угнетением сократимости миокарда и расширением вен.

В связи с этим предлагается использовать морфин вместо галотана, так как он не влияет на артериальное давление.

Т. Конахан и коллеги провели сравнительное исследование галотановой и морфиновой анестезии у пациентов, перенесших операцию на открытом сердце. Пациенты, у которых не было противопоказаний к обоим препаратам, были случайным образом распределены на две группы: одна получала галотан, другая – морфин. Были зарегистрированы следующие показатели: параметры гемодинамики на разных этапах операции, длительность пребывания в реанимационном отделении и общая длительность пребывания в больнице после операции, а также послеоперационная летальность.

Особое внимание было уделено артериальному давлению между началом анестезии и началом операции. В этот период артериальное давление наиболее адекватно отражает гипотензивное действие анестетика, поскольку в дальнейшем начинает сказываться гипотензивный эффект самой операции. Артериальное давление измеряли многократно, каждый раз вычисляя среднее артериальное давление.

В исследовании приняли участие 122 пациента. У половины пациентов использовали галотан (1-я группа), у половины – морфин (2-я группа). Результаты представлены в таблице:

Среднее артериальное давление при анестезии	
галотановая	морфиновая
42	42
44	46
46	46
46	52
48	56
48	58
50	58
50	58
52	58
54	60
54	60
56	60
58	62
58	62
58	62
60	62
60	62

60	64
60	64
60	64
62	66
62	66
62	66
62	66
62	68
64	68
64	68
66	68
66	70
66	70
66	70
66	72
68	74
68	74
68	76
68	76
70	76
70	78
70	80
70	80
70	80
72	82
72	82
72	82
72	84
74	84
74	84
74	84
74	86
76	86
78	86
78	88
78	88
80	90
80	90
82	92
82	96
84	96
86	98
90	98
94	98
98	100

Контрольные вопросы

1. Когда надо выполнять сравнение групп?
2. Какие группы называются контрольными и экспериментальными?
3. Что такое нулевая гипотеза?
4. Что такое альтернативная гипотеза?
5. Что называется уровнем значимости?
6. Какие значения уровня значимости используются в биологических и медицинских исследованиях?
7. Какие типы ошибок в биостатистике вам известны?
8. Что подразумевается, когда говорят, что уровень значимости равен 0,05?
9. Опишите пошагово выполнение дисперсионного анализа.
10. Что такое критическое значение критерия F?

2.4. Критерий Стьюдента

Краткие сведения из теории

Критерий Стьюдента (t) – это статистический метод, который используется для проверки гипотезы о равенстве средних значений двух выборок. Этот критерий базируется на предположении, что две выборки имеют нормальное распределение и одинаковую дисперсию.

Следует помнить, что критерий Стьюдента предназначен для сравнения только двух групп, а не нескольких групп попарно.

Критерий Стьюдента основан на вычислении t-статистики, которая определяется как отношение разности средних значений к стандартной ошибке разности средних:

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Стандартная ошибка разности средних вычисляется как квадратный корень из суммы квадратов стандартных отклонений, деленной на количество наблюдений в каждой выборке.

Если две выборки извлечены из одной нормально распределенной совокупности, то отношение их средних значений, как правило, будет близко к нулю. Это означает, что средние значения двух выборок, извлеченных из одной совокупности, будут близки друг к другу.

Чем меньше (по абсолютной величине, по модулю) t -статистика, тем больше вероятность справедливости нулевой гипотезы, т. е. гипотезы о равенстве средних значений двух выборок. Чем больше t -статистика, тем больше оснований отвергнуть нулевую гипотезу и считать, что различия статистически значимы.

Критическое значение критерия Стьюдента ($t_{кр}$) – это величина, начиная с которой отвергается нулевая гипотеза. Это значение выбирается в зависимости от уровня значимости (обычно составляющее 0,05) и количества степеней свободы (обычно вычисляемое как количество наблюдений в каждой выборке минус 1). Если t -статистика больше критического значения, то гипотеза о равенстве средних значений отвергается.

Если t -статистика больше критического значения t -распределения с заданным уровнем значимости и количеством степеней свободы, то гипотеза о равенстве средних значений отвергается. В противном случае гипотеза принимается.

Ошибки в использовании критерия Стьюдента могут возникнуть, если не выполняются предположения, на которых основан этот метод. Некоторые из наиболее распространенных ошибок:

1. Неправильное предположение о нормальности данных: критерий Стьюдента предполагает, что данные имеют нормальное распределение. Если данные не имеют нормального распределения, то результаты могут быть искажены. В этом случае следует использовать другие методы статистического анализа, например, критерий Манна–Уитни.

2. Неправильное предположение о равенстве дисперсий: критерий Стьюдента предполагает, что две выборки имеют одинаковую дисперсию. Если дисперсии выборок различны, то результаты могут быть искажены.

Чтобы избежать ошибок в использовании критерия Стьюдента, необходимо тщательно проверять предположения, на которых основан этот метод, и использовать другие методы статистического анализа, если эти предположения не выполняются.

Алгоритм действий при использовании критерия Стьюдента:

1. Создание таблиц с результатами измерений.
2. Анализ полученных данных с помощью описательной статистики.
3. Проверка соответствия распределения данных нормальному распределению.
4. Выбор значения уровня значимости (в зависимости от строгости исследования: 0,1 или 0,05, или 0,01).

5. Формулировка нулевой гипотезы (различие между группами незначимо или является следствием случайности).

6. Расчет значения критерия Стьюдента (значение критерия, начиная с которого мы отвергаем нулевую гипотезу) и вероятности ошибочного результата p (вероятность ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет).

7. Определение величины критического значения критерия $t_{кр}$ (значение критерия, начиная с которого мы отвергаем нулевую гипотезу) на основании значения уровня значимости и количества элементов выборки.

8. Сравнение t и $t_{кр}$, а также p и α , и если $t > t_{кр}$ и $p < \alpha$, то нулевая гипотеза отвергается и различия между группами статистически значимы, в противном случае различия случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность равна уровню значимости.

Задача

Поиск новых антибиотиков или улучшение существующих, чтобы они были более эффективными против устойчивых бактерий, является важным направлением в современной науке. Поэтому группа исследователей решила изучить влияние экспериментального антибиотика на гибель определенного вида бактерий. Были сформированы две группы (контрольная и экспериментальная): бактерии, выращенные в среде без антибиотика (группа 1), и бактерии, выращенные в среде с антибиотиком (группа 2). Оценим статистическую значимость различий между группами. Были получены следующие данные:

Количество погибших бактерий, %	
Группа 1	Группа 2
9,2	13,7
8,6	14,4
7,8	16,8
7,5	14,8
9,5	14,2
6,4	15,3
6,7	16,3
7,4	12,8
8,2	13,3
7,0	14,8
8,3	14,2
7,7	14,7

9,7	13,3
7,6	14,4
7,9	14,1
8,2	14,2
8,1	13,7
7,0	15,2
8,0	15,7
5,9	13,7
8,6	13,6
9,0	12,7
7,8	15,3
8,2	13,6
7,8	14,4

Сформулируем нулевую гипотезу (H_0): применение экспериментального антибиотика не влияет на гибель бактерий.

Алгоритм анализа в MS Excel

1. Скопируем таблицу с данными в MS Excel:

	А	В
1	Группа 1	Группа 2
2	9,2	13,7
3	8,6	14,4
4	7,8	16,8
5	7,5	14,8
6	9,5	14,2
7	6,4	15,3
8	6,7	16,3
9	7,4	12,8
10	8,2	13,3
11	7,0	14,8
12	8,3	14,2
13	7,7	14,7
14	9,7	13,3
15	7,6	14,4
16	7,9	14,1
17	8,2	14,2
18	8,1	13,7
19	7,0	15,2
20	8,0	15,7
21	5,9	13,7
22	8,6	13,6
23	9,0	12,7
24	7,8	15,3
25	8,2	13,6
26	7,8	14,4

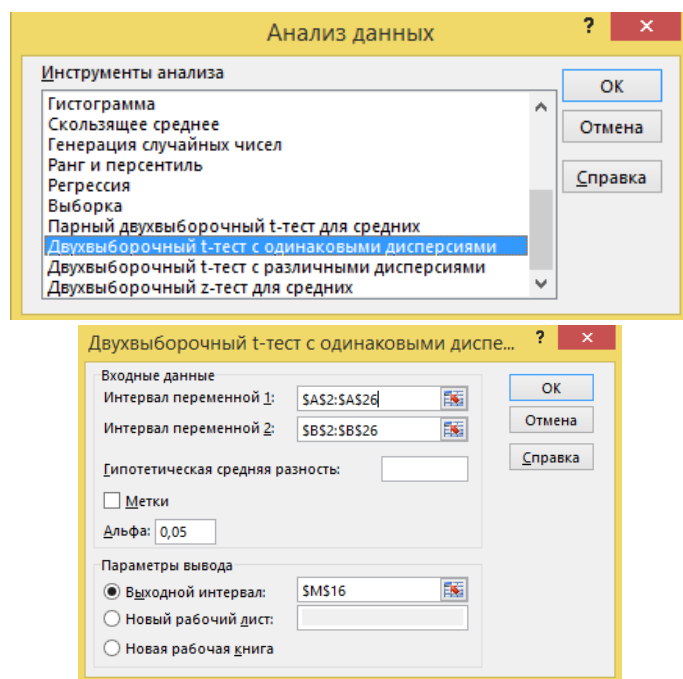
2. Используем модуль «Описательная статистика», чтобы получить основные статистические параметры для каждой группы данных:

Столбец1		Столбец2	
Среднее	7,92431	Среднее	14,36693
Стандартная ошибка	0,18275	Стандартная ошибка	0,202621
Медиана	7,87024	Медиана	14,22675
Мода	#Н/Д	Мода	#Н/Д
Стандартное отклонение	0,913752	Стандартное отклонение	1,013103
Дисперсия выборки	0,834944	Дисперсия выборки	1,026379
Эксцесс	0,196897	Эксцесс	0,252286
Асимметричность	-0,1448	Асимметричность	0,619333
Интервал	3,804143	Интервал	4,098688
Минимум	5,87773	Минимум	12,72361
Максимум	9,681874	Максимум	16,8223
Сумма	198,1078	Сумма	359,1732
Счет	25	Счет	25

Рассчитаем средние значения для каждой группы и сравним их. Цель анализа – определить, есть ли значимые различия между средними значениями каждой группы, а также понять, есть ли какая-либо тенденция или закономерность в данных и как они могут быть использованы для принятия решений или разработки стратегий.

Средние значения отличаются друг от друга по величине, поэтому необходимо использовать критерий Стьюдента для проверки статистической значимости этих различий. Нулевая гипотеза заключается в том, что различия средних значений не являются статистически значимыми. Однако будем помнить, что критерий Стьюдента применим только в случае, если распределение данных в группах нормальное и их дисперсии приблизительно равны.

3. Воспользуемся двухвыборочным t-тестом для одинаковых дисперсий:



4. Результат выводится в виде таблицы:

Двухвыборочный t-тест с одинаковыми дисперсиями		
	Переменная 1	Переменная 2
Среднее	7,924310214	14,36692703
Дисперсия	0,834943568	1,026378509
Наблюдения	25	25
Объединенная дисперсия	0,930661039	
Гипотетическая разность средних	0	
df	48	
t-статистика	-23,61138932	
P(T<=t) одностороннее	2,26103E-28	
t критическое одностороннее	1,677224196	
P(T<=t) двухстороннее	4,52207E-28	
t критическое двухстороннее	2,010634758	

Интересующие исследователя значения выделены цветом. Из таблицы видно, что расчетное значение критерия Стьюдента (t-статистика) равно $-23,51$, что больше (по модулю) его критического значения (t критическое двухстороннее), равного $4,4$. Вероятность ошибки p намного меньше заданного уровня значимости α , равного $0,05$. Следовательно, нулевая гипотеза отвергается, и различия средних значений показателей статистически значимы: экспериментальный антибиотик увеличивает гибель исследуемых бактерий.

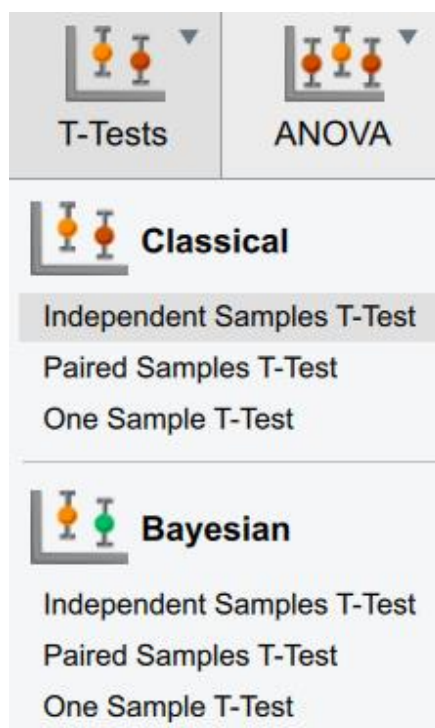
Алгоритм анализа в JASP для независимых групп

1. Откроем требуемые данные в JASP, сохранив документ MSExcel в формате ods. Однако предварительно Excel-таблицу необходимо преобразовать в следующий вид (где a, b – предикторы для каждой группы):

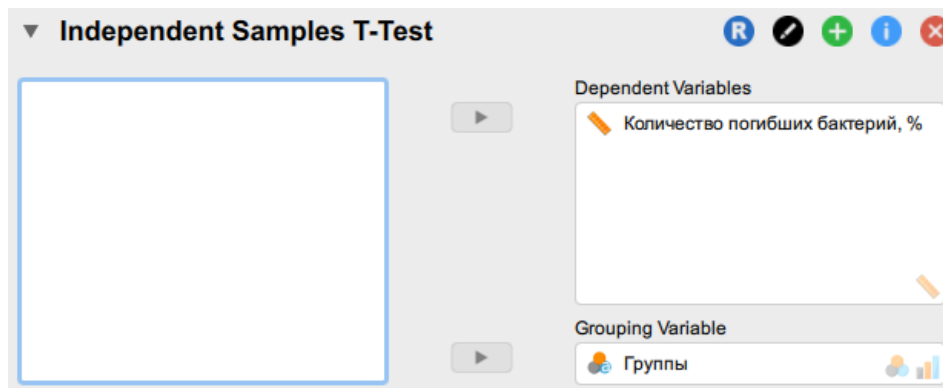
	A	B
16	a	8,2
17	a	8,1
18	a	7
19	a	8
20	a	5,9
21	a	8,6
22	a	9
23	a	7,8
24	a	8,2
25	a	7,8
26	b	13,7
27	b	14,4
28	b	16,8
29	b	14,8
30	b	14,2
31	b	15,3
32	b	16,3
33	b	12,8

	Группы	Количество погибших бактерий, %
1	a	9.2
2	a	8.6
3	a	7.8
4	a	7.5
5	a	9.5
6	a	6.4
7	a	6.7
8	a	7.4
9	a	8.2
10	a	7
11	a	8.3
12	a	7.7
13	a	9.7
14	a	7.6
15	a	7.9
16	a	8.2
17	a	8.1
18	a	7
19	a	8
20	a	5.9

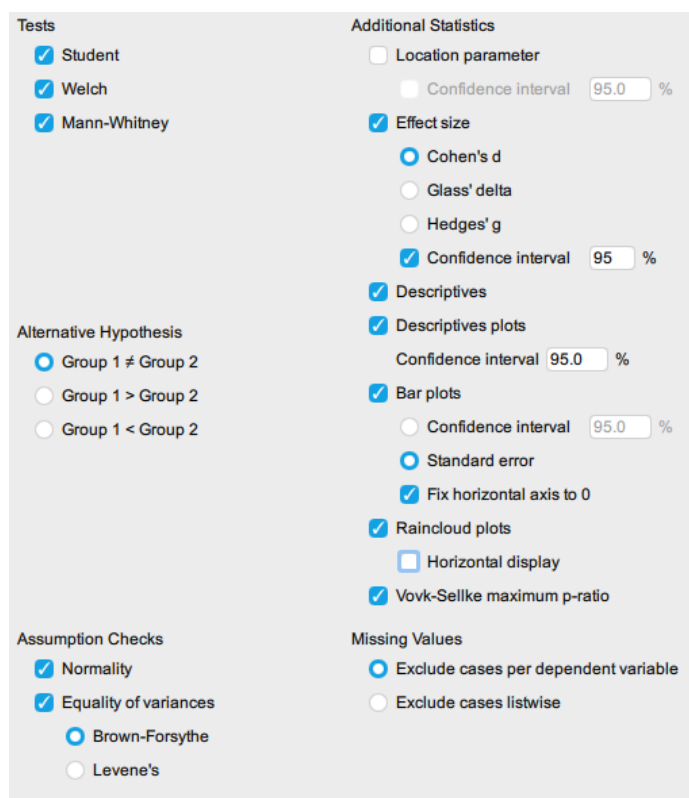
2. Нажмем на модуль «Сравнение групп» («T-tests»). В выпадающем списке выбрать «Сравнение независимых групп» («Independent Samples T-test»):



3. В большее окно «Dependent Variables» перенесем «целевые» переменные или переменную, а в меньшее окно «Grouping Variable» перенесем категориальную (группирующую) переменную, имеющую только две категории (группы):



4. В открывшемся меню настроим необходимые параметры для проверки гипотезы, расчета необходимых статистик, построения графиков и пр. Шаблон приведен на рисунке:



Выбор критерия зависит от результатов проверки допущений.

Используемые пункты меню анализа:

Student – критерий Стьюдента (параметрический, для независимых групп).

Welch – критерий Уелча (Вэлча) (параметрический, для независимых групп). По сути, является доработкой критерия Стьюдента для ситуаций, где сравниваемые группы являются неоднородными, например, с сильно различающейся дисперсией.

Mann–Whitney – критерий Манна–Уитни (непараметрический, для независимых групп). Независимый от типа распределения.

Effect size – размер эффекта.

Cohen's D – d Козна.

Confidence interval – доверительный интервал для размеров эффекта.

Descriptive – таблица описательных статистик.

Descriptive plot – график средних значений.

Alternative Hypothesis – указывает на направление проверки статистической гипотезы. Менять не надо.

Assumption Checks – для проверки допущений (условий) использования критериев.

Normality – для проверки допущения о нормальном распределении в каждой из групп используется критерий Шапиро–Уилка. Если хотя бы в одной группе имеется нарушение допущения (при условии маленьких выборок – менее 60), следует отказаться от использования параметрических методов. При $p < 0,05$ следует принять гипотезу о различии между теоретически нормальным распределением и эмпирическим распределением соответствующей переменной и группы. Из этого следует, что распределение не соответствует нормальному.

Equality of variances – для проверки допущений об однородности групп через оценку нулевой гипотезы о равенстве (отсутствии различий) дисперсий. Для этого представлены два взаимозаменяемых критерия – критерий Браун–Форсайта (Brown–Forsythe) и критерий Ливиня (Levene's). Аналогично критериям согласия при $p < 0,05$, следует принять альтернативную гипотезу о неравенстве (наличии различий) дисперсий групп.

Missing values – настройка обработки пропущенных значений (при их наличии в матрице данных). Есть два варианта: исключить случаи для каждой «целевой» переменной по отдельности (Exclude cases per dependent variable) и исключить случаи целиком (Exclude cases listwise). Второй отличается от первого тем, что при отсутствии значения по одной переменной исключается случай полностью и по другим переменным и группам.

Анализ данных идет в три этапа: 1) оценка допущений, 2) выбор критерия, 3) описание результатов.

В таблице ниже столбец *W* содержит эмпирические значения критерия Шапиро–Уилка, а столбец *p* – уровень значимости (вероятность ошибки первого рода или вероятность ошибиться при принятии нулевой гипотезы):

Assumption Checks ▼

Test of Normality (Shapiro-Wilk) ▼

		W	p
Количество погибших бактерий, %	a	0.982	0.913
	b	0.965	0.533

Note. Significant results suggest a deviation from normality.

В таблице ниже важными значениями являются те, что находятся в столбце *F* и столбце *p*:

Test of Equality of Variances (Brown-Forsythe)

	F	df ₁	df ₂	p
Количество погибших бактерий, %	0.245	1	48	0.623

Аналогично предыдущей таблице *F* – значение статистического критерия, а *p* – уровень значимости. Для критерия Ливиня интерпретация аналогична. Данная проверка необходима в случае нормального распределения.

При $p < 0,05$ необходимо принять альтернативную гипотезу, в остальных случаях – нулевую гипотезу.

Так как тип распределения является нормальным для обеих групп, то не следует использовать непараметрический критерий Манна–Уитни. Для иллюстрации далее будут представлены результаты *t*-критерия Стьюдента и *U*-критерия Манна–Уитни.

В таблице ниже представлены главные результаты проверки гипотезы о различии в группах:

Independent Samples T-Test ▼

Independent Samples T-Test

	Test	Statistic	df	p	VS-MPR*	Effect Size	SE Effect Size	95% CI for Effect Size	
								Lower	Upper
Количество погибших бактерий, %	Student	-23.646	48.000	< .001	$1.377 \times 10^{+25}$	-6.688	0.987	-8.126	-5.238
	Welch	-23.646	47.514	< .001	$9.302 \times 10^{+24}$	-6.688	0.987	-8.132	-5.231
	Mann-Whitney	0.000		< .001	$1.304 \times 10^{+7}$	-1.000	0.163	-1.000	-1.000

Note. For the Student t-test and Welch t-test, effect size is given by Cohen's d. For the Mann-Whitney test, effect size is given by the rank biserial correlation.

* Vovk-Sellke Maximum p-Ratio: Based on a two-sided p-value, the maximum possible odds in favor of H_1 over H_0 equals $1/(-e p \log(p))$ for $p \leq .37$ (Sellke, Bayarri, & Berger, 2001).

Столбец «Statistic» содержит информацию об эмпирическом значении t-критерия для соответствующей переменной, df – степень свободы, p – уровень значимости, столбец «SE Effect Size» – стандартная ошибка размера эффекта. Для оценки степени (силы) различий между группами с помощью размера эффекта используем соответствующую таблицу.

Визуализируем полученные результаты статистического анализа в виде следующих рисунков:

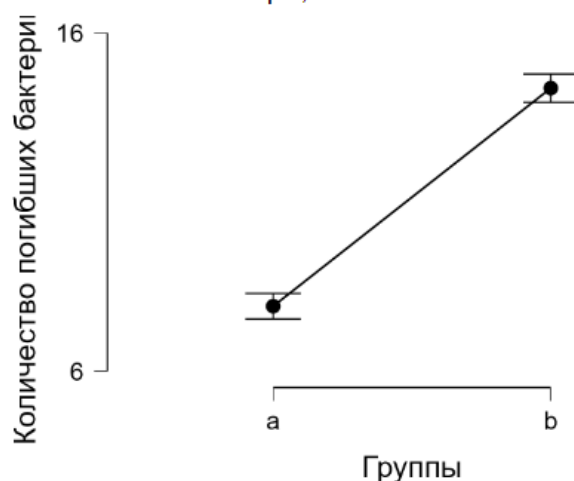
Descriptives ▼

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Количество погибших бактерий, %	a	25	7.924	0.913	0.183	0.115
	b	25	14.368	1.011	0.202	0.070

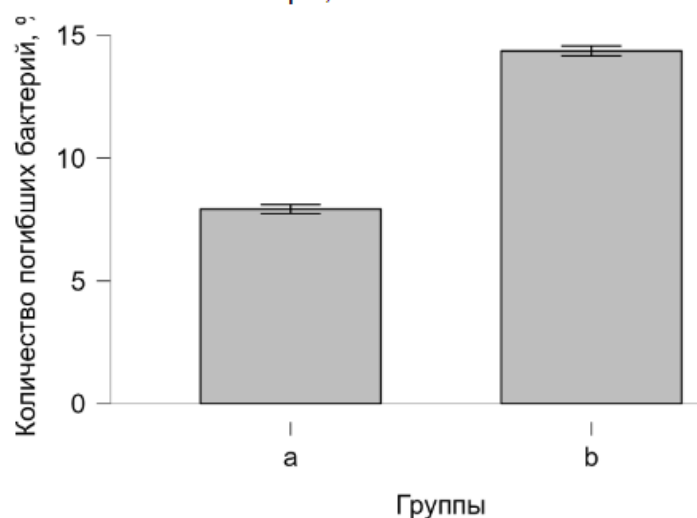
Descriptives Plots ▼

Количество погибших бактерий, %

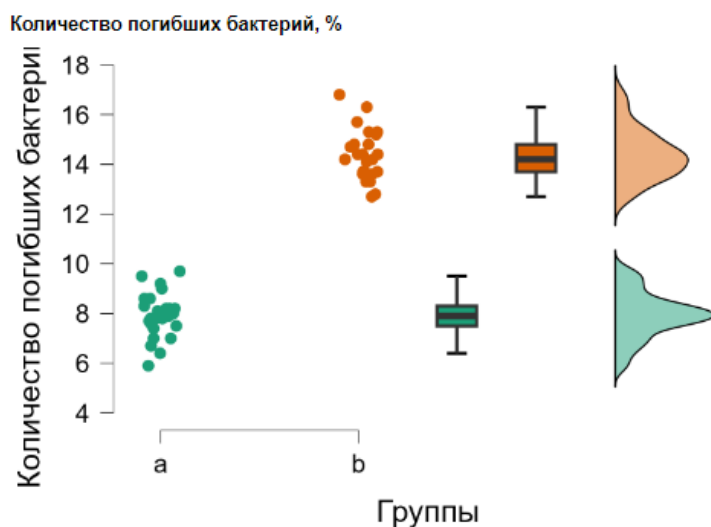


Bar Plots ▼

Количество погибших бактерий, %



Raincloud Plots

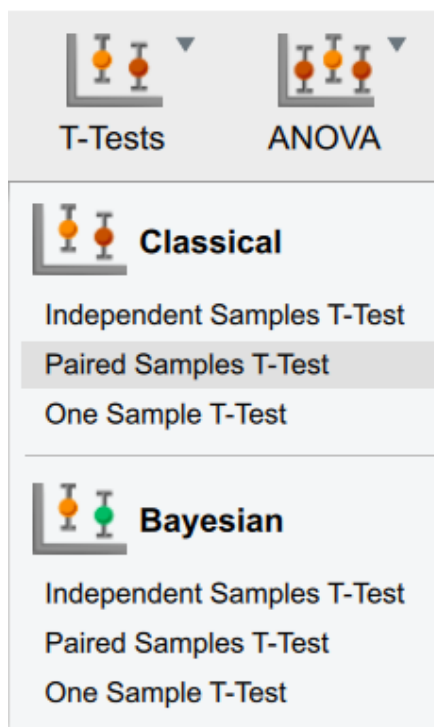


Алгоритм анализа в JASP для зависимых групп

1. Загрузим таблицу (в формате ods) в JASP:

Пол	Систолическое давление до	Диастолическое давление до	Систолическое давление после	Диастолическое давление после
М	177	126	123	101
М	174	171	144	102
М	161	125	145	111
М	163	128	142	92
М	164	142	146	92
М	153	86	140	111
М	154	128	138	95
М	170	112	134	111
М	174	137	144	108
М	155	142	125	114
Ж	161	117	107	88
Ж	157	130	122	97
Ж	155	135	91	93
Ж	159	124	120	86
Ж	155	127	112	95
Ж	121	128	111	91
Ж	155	125	108	89
Ж	135	112	106	89
Ж	136	123	114	88
Ж	121	123	110	99

2. Нажмем на модуль «Сравнение групп» («T-tests»). В выпадающем списке выберем «Сравнение зависимых групп» («Paired Samples T-test»):



Зависимые группы (выборки) – это такие группы случаев (людей), у которых измерения переменной связаны. А измерения считаются связанными, так как оценивался один и тот же признак, но в разные дни у одних и тех же людей. То есть измерения одной переменной у одной выборки, но в разное время, являются зависимыми. В связи с вышеизложенным, переменные указываются попарно. В пару переносятся зависимые измерения (выборки).

3. В открывшемся меню введем параметры для проверки гипотезы, расчета необходимых статистик, построения графиков и пр.

Используемые пункты меню анализа:

Student – критерий Стьюдента (параметрический, для зависимых групп).

Wilcoxon signed-rank – критерий Вилкоксона (непараметрический, для зависимых групп). Независимый от типа распределения.

Effect size – размер эффекта. Для параметрического критерия используется d Коэна, а для непараметрического – ранговая бисериальная корреляция.

Confidence interval – доверительный интервал для размеров эффекта.

Descriptive – таблица описательных статистик.

Descriptive plot – график средних значений.

Alternative Hypothesis – указывает на направление проверки статистической гипотезы. Менять не надо.

Assumption Checks – для проверки допущений (условий) использования критериев.

Normality – для проверки допущения о нормальном распределении в каждой из групп используется критерий Шапиро–Уилка. Если хотя бы в одной группе имеется нарушение допущения (при условии маленьких выборок – менее 60), следует отказаться от использования параметрических методов. При $p < 0,05$ следует принять гипотезу о различии между теоретически нормальным распределением и эмпирическим распределением соответствующей переменной и группы. Из этого следует, что распределение не соответствует нормальному.

Анализ данных идет в три этапа: 1) оценка допущений, 2) выбор критерия, 3) описание результатов.

В таблице ниже столбец W содержит эмпирические значения критерия Шапиро–Уилка, а столбец p – уровень значимости (вероятность ошибки первого рода или вероятность ошибиться при принятии нулевой гипотезы):

Test of Normality (Shapiro-Wilk)

			W	p
Систолическое давление до	-	Систолическое давление после	0.946	0.308
Диастолическое давление до	-	Диастолическое давление после	0.878	0.016

Note. Significant results suggest a deviation from normality.

При $p < 0,05$ необходимо принять альтернативную гипотезу, в остальных случаях – нулевую гипотезу.

Следует сделать вывод о нормальном типе распределения.

Можно заметить, что в проверке допущений отсутствует оценка однородности групп (выборок или измерений). Данное обстоятельство связано с теоретическим условием зависимости выборок – того факта, что они взяты из одной генеральной совокупности и имеют равную дисперсию. Однако это может быть не так эмпирически. И если условия зависимости выполнены, то слишком большая дисперсия связана с условиями измерения признака, а не свойствами объектов (случаев/людей).

В меню анализа необходимо выбрать соответствующий критерий. Далее будут представлены результаты параметрического и непараметрического критериев:

Paired Samples T-Test ▼

Paired Samples T-Test

Measure 1		Measure 2	Test	Statistic	z	df	p
Систолическое давление до	-	Систолическое давление после	Student	8.953		19	< .001
			Wilcoxon	210.000	3.920		< .001
Диастолическое давление до	-	Диастолическое давление после	Student	7.176		19	< .001
			Wilcoxon	205.000	3.733		< .001

4. Построим графики по следующему плану:

Alternative Hypothesis

☒ Measure 1 ≠ Measure 2

☐ Measure 1 > Measure 2

☐ Measure 1 < Measure 2

Confidence interval 95.0 %

☒ Bar plots

☐ Confidence interval 95.0 %

☒ Standard error

☒ Fix horizontal axis to 0

☒ Raincloud plots

☒ Raincloud difference plots

☐ Horizontal display

☒ Vovk-Sellke maximum p-ratio

Assumption Checks

☐ Normality

Missing Values

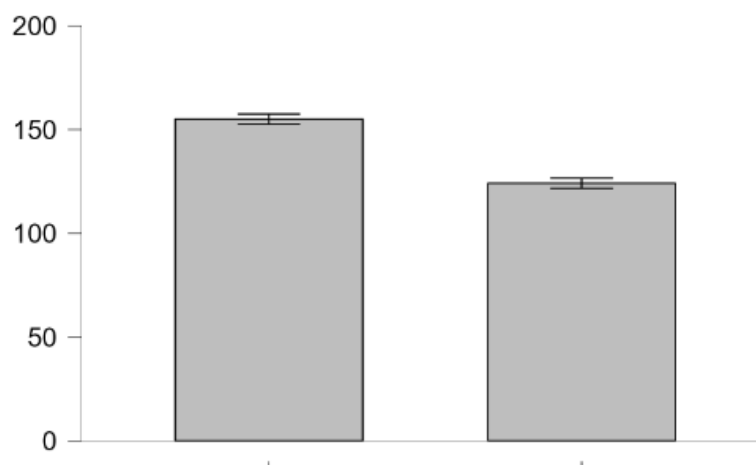
☒ Exclude cases per variable

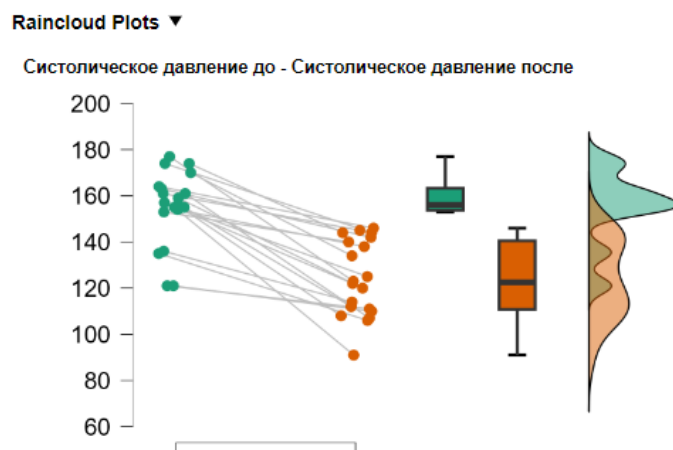
☐ Exclude cases listwise

Descriptives ▼

Bar Plots ▼

Систолическое давление до - Систолическое давление после ▼





Контрольные вопросы

1. Алгоритм использования критерия Стьюдента.
2. Что такое критическое значение критерия Стьюдента?
3. Основные ошибки применения критерия Стьюдента.
4. Если расчетное значение критерия t больше чем $t_{кр}$ при заданном уровне значимости, что это значит?

2.5. Корреляционный и регрессионный анализ

Краткие сведения из теории

Регрессионный анализ – это статистический метод, который используется для изучения взаимосвязи между одной или более независимыми переменными (предикторами) и зависимой переменной (откликом). Основная цель регрессионного анализа – определить, насколько хорошо изменение одной или более независимых переменных объясняет изменение зависимой переменной.

Регрессионный анализ предполагает построение линии регрессии методом наименьших квадратов и описание имеющейся зависимости уравнением регрессии с определенной точностью.

В уравнении регрессии одна из переменных (x) называется независимой переменной, а другая (y) – зависимой. Это не означает, что одна переменная однозначно определяет другую. По значению одного признака предсказывается значение второго. В условиях эксперимента можно изменять независимую переменную и отслеживать, как меняется зависимая. При достаточном количестве данных эксперимента можно делать выводы о наличии взаимосвязи и адекватности полученной модели.

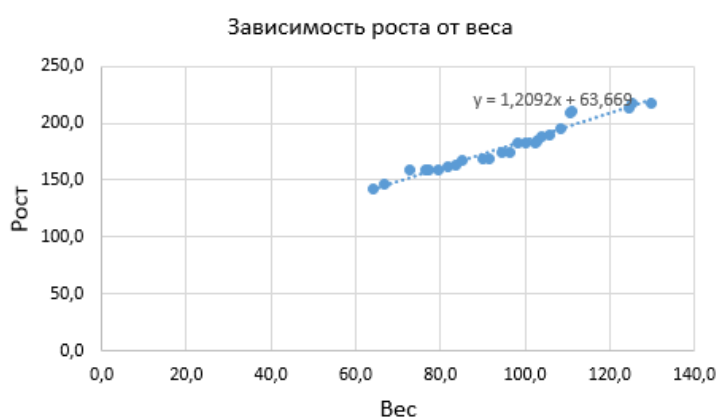
График уравнения регрессии представляет собой линию регрессии, для построения которой по оси X откладываются значения независимой переменной (x), а по оси Y – значения зависимой переменной (y). Координаты точек на графике соответствуют значениям переменных в уравнении регрессии.

Пример уравнения регрессии, описывающего прямую линию:

$y = a + bx$, где a – константа; b – коэффициент регрессии.

Уравнение регрессии может быть использовано для описания зависимости, например, между ростом и весом человека. В этом случае: y – рост человека в сантиметрах, x – вес человека в килограммах.

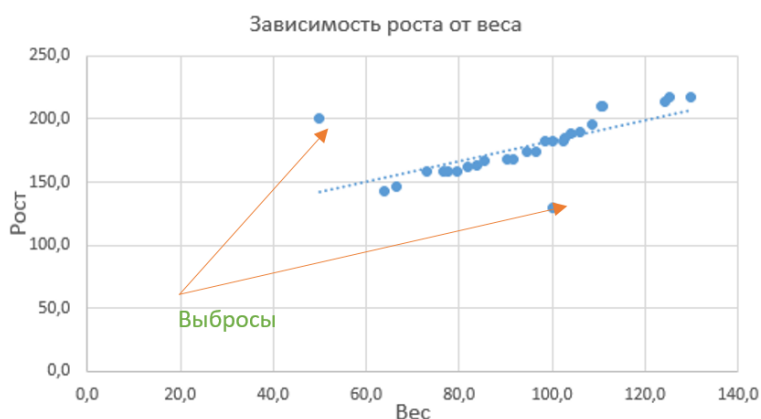
На графике уравнения регрессии по оси X будут откладываться значения веса человека, а по оси Y – рост человека:



Если мы хотим предсказать рост человека на основе его веса, то, принимая $a = 170$ и $b = 0,5$, получим следующее уравнение регрессии:

$$y = 170 + 0,5x.$$

Кроме этого, не стоит забывать о выбросах. Выбросы в регрессионном анализе – это значения, которые значительно отличаются от остальных данных:



Они могут быть как положительными, так и отрицательными. Выбросы могут исказить результаты анализа и приводить к неправильным выводам.

Для определения выбросов можно использовать различные методы, например, графический анализ или статистические тесты. Если выбросы обнаружены, их можно удалить из данных или учесть при построении модели.

Таким образом, регрессионный анализ позволяет построить модель, которая описывает зависимость между переменными. Эта модель может быть использована для прогнозирования значений зависимой переменной на основе значений независимых переменных.

Существует несколько видов регрессионного анализа, включая линейную регрессию, полиномиальную регрессию, логистическую регрессию и другие. Каждый вид регрессии имеет свои особенности и применяется в зависимости от характеристик данных и исследовательских задач.

Регрессионный анализ также позволяет оценить значимость влияния каждой независимой переменной на зависимую переменную. Это делается с помощью статистических тестов, которые проверяют, насколько хорошо модель объясняет данные.

Корреляционный анализ – это статистический метод, который используется для определения степени связи между двумя переменными. Он позволяет оценить, насколько сильно одна переменная влияет на другую.

Корреляционный анализ основан на вычислении коэффициента корреляции, который может принимать значения от -1 до $+1$. Если коэффициент корреляции равен $+1$, то это означает, что две переменные полностью коррелируют друг с другом, т. е. изменение одной переменной полностью объясняет изменение другой. Если коэффициент корреляции равен -1 , то это означает, что две переменные полностью коррелируют друг с другом, но в противоположном направлении. Если коэффициент корреляции равен 0 , то это означает, что две переменные не коррелируют друг с другом.

Например, если мы имеем данные о росте и весе 100 человек, то мы можем вычислить коэффициент корреляции между этими двумя переменными. Если коэффициент корреляции близок к $+1$, то это означает, что рост и вес человека сильно связаны друг с другом. Если коэффициент корреляции близок к 0 , то это означает, что рост и вес человека слабо связаны друг с другом.

Корреляционный анализ может быть полезен для определения связи между различными переменными и для прогнозирования будущих значений одной переменной на основе значений другой переменной.

Коэффициент корреляции Пирсона (r) является мерой линейной зависимости между двумя переменными. Он позволяет оценить, насколько тесно связаны две переменные, и определить, в каком направлении эта связь идет.

Если коэффициент корреляции Пирсона положителен ($r > 0$), то это говорит о том, что с увеличением значения одной переменной значение другой переменной также увеличивается. Такая связь называется *положительной линейной зависимостью*.

Если коэффициент корреляции Пирсона отрицателен ($r < 0$), то это говорит о том, что с увеличением значения одной переменной значение другой переменной уменьшается. Такая связь называется *отрицательной линейной зависимостью*.

Если коэффициент корреляции Пирсона равен нулю ($r = 0$), то это говорит о том, что между двумя переменными нет линейной зависимости.

Если данный коэффициент возвести в квадрат, получим *коэффициент детерминации* (r^2). Он представляет собой долю вариации, общую для двух переменных (иными словами, степень зависимости или связанности двух переменных). Чтобы оценить зависимость между переменными, нужно знать как величину корреляции, так и ее значимость.

Несколько правил использования коэффициента корреляции Пирсона:

1. Коэффициент корреляции Пирсона может быть использован только для измерения линейной зависимости между переменными. Если связь между переменными нелинейная, то коэффициент корреляции Пирсона может быть равен нулю, даже если связь между переменными существует.

2. Коэффициент корреляции Пирсона не говорит о причинно-следственной связи между переменными. Он только показывает, насколько тесно связаны две переменные.

3. Коэффициент корреляции Пирсона может быть использован для оценки силы связи между переменными. Если коэффициент корреляции Пирсона близок к 1 или -1 , то это говорит о сильной связи между переменными.

4. Коэффициент корреляции Пирсона может быть использован для оценки значимости связи между переменными. Если коэффициент корреляции Пирсона значим (p -значение меньше заданного уровня значимости), то это говорит о том, что связь между переменными является статистически значимой.

5. Коэффициент корреляции Пирсона может быть использован для построения регрессионной модели. Если коэффициент корреляции

Пирсона значим, то можно построить регрессионную модель, которая будет объяснять зависимость между переменными.

6. Коэффициент корреляции Пирсона, как и регрессионный анализ, требует нормальности распределения.

Коэффициент корреляции Спирмена (или ранговый коэффициент корреляции) – это статистический показатель, который используется для измерения степени линейной зависимости между двумя переменными. Он был предложен сэром Рональдом Эйлмером Спирменом, британским статистиком, в начале XX века.

Коэффициент корреляции Спирмена обозначается греческой буквой ρ («ро»).

Значения коэффициента корреляции Спирмена могут варьироваться от -1 до 1 . Если $\rho = 1$, то это говорит о полной положительной линейной зависимости между переменными (чем больше значение одной переменной, тем больше значение другой). Если $\rho = -1$, то это говорит о полной отрицательной линейной зависимости (чем больше значение одной переменной, тем меньше значение другой). Если $\rho = 0$, то это говорит об отсутствии линейной зависимости между переменными.

Коэффициент корреляции Спирмена является мерой непараметрической зависимости между двумя переменными. Он может быть использован в случаях, когда данные не соответствуют нормальному распределению или когда данные имеют пропущенные значения.

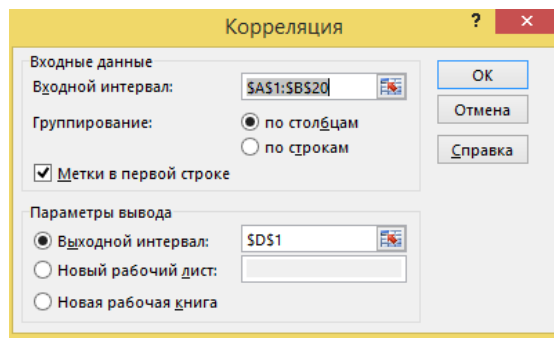
Порядок выполнения корреляционного, а затем регрессионного анализа покажем на примере измерения количества погибших бактерий (Y) и дозы антибиотика (X). Исходные данные представлены в таблице:

X	Y
1	5
2	7
3	9
4	11
5	13
6	15
7	17
8	19
9	21
10	23
11	25
12	27
13	22
14	31
15	33
16	35
17	37
18	39
19	41

Алгоритм корреляционного анализа в MS Excel

1. Откроем модуль «Анализ данных», перейдем во вкладку «Данные». В группе «Анализ» выберем «Анализ данных...».

2. Выберем тип анализа. Для этого в появившемся диалоговом окне выберем «Корреляция» из списка доступных типов анализа. В открывшемся окне выполним операции и установки, как показано на рисунке:



На стартовой панели наши значения в ячейке «Входной интервал» и «Выходной интервал» могут отличаться, метки в первой строке ставятся при условии, что входной интервал включает заголовки таблицы X и Y.

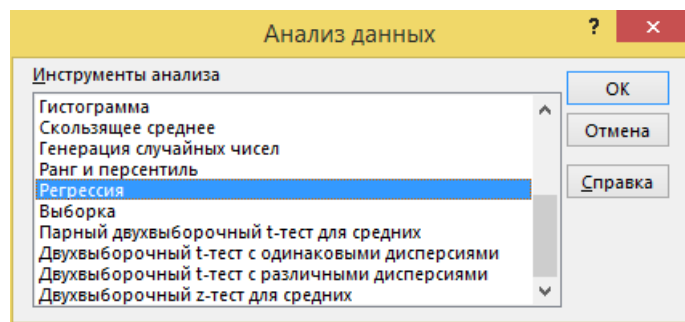
3. Щелкнем мышкой «ОК». Результат обработки появится в поле «Выходной интервал \$D\$1»:

D	E	F
	X	Y
X	1	
Y	0,99	1

В полученной таблице нас интересует значение в ячейке на пересечении X и Y – 0,99. Это и есть искомое значение коэффициента корреляции.

Алгоритм регрессионного анализа в MS Excel

1. Откроем модуль «Анализ данных», выберем опцию «Регрессия» и нажмем «ОК»:



2. В появившемся окне выполним операции и установки, как показано на рисунке:

3. Щелчком мышкой «ОК». Результат обработки появится в поле «Выходной интервал \$I\$1».

Результат анализа:

Вывод итогов								
Регрессионная статистика								
Множественный R	0,98977							
R-квадрат	0,979644							
Нормированный R-квадрат	0,978446							
Стандартная ошибка	1,638639							
Наблюдения	19							
Дисперсионный анализ								
	df	SS	MS	F	Значимость F			
Регрессия	1	2196,773684	2196,773684	818,1227949	8,09585E-16			
Остаток	17	45,64736842	2,685139319					
Итого	18	2242,421053						
	Коэффициент	Стандартная ошибка	t-статистика	P-значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	3	0,782560027	3,833571735	0,001330278	1,348942665	4,651057335	1,348942665	4,651057335
Переменная X 1	1,963158	0,068635055	28,60284592	8,09585E-16	1,818350587	2,107965202	1,818350587	2,107965202

Таким образом, корреляционная связь между дозой препарата и количеством погибших бактерий характеризуется высоким и достоверным коэффициентом корреляции ($r = 0,99$). Достоверность проверяется критерием Фишера из данной таблицы: критерий Фишера $F = 818,122$ при уровне значимости данного критерия, существенно меньшем $0,05$. Получена очень надежная регрессия, о чем свидетельствует t -статистика из таблицы (уровень значимости критерия F и p -значения существенно меньше $0,05$).

Уравнение линейной регрессии Пирсона:

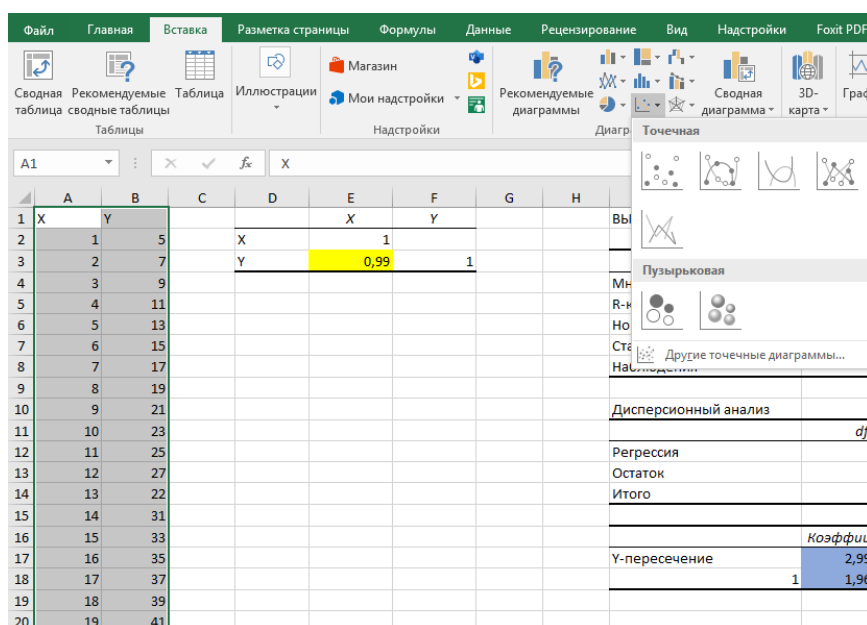
$Y = 3 + 1,96 \cdot X$, где 3 и 1,96 получены в результате обработки данных в MS Excel.

Альтернативный способ выполнения анализа в MS Excel

Другой, более простой и менее информативный, способ выполнения операций регрессии и корреляции в MS Excel – использование модуля «Мастер диаграмм».

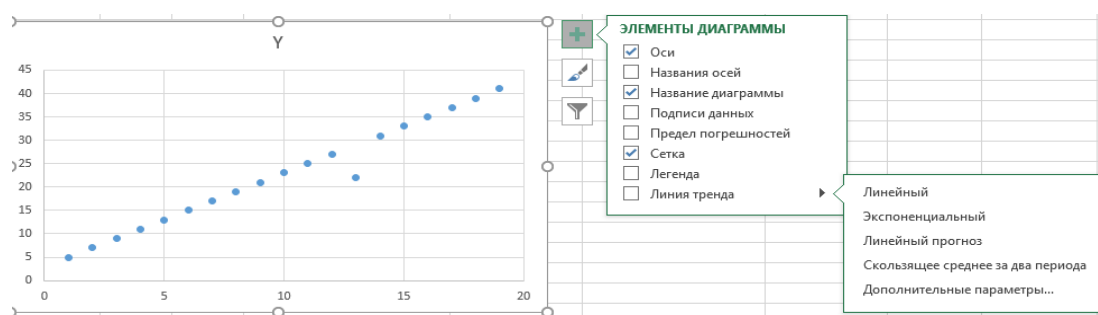
1. В MS Excel выделить значения переменных X и Y, которые мы хотим проанализировать. Затем открыть вкладку «Вставка», опция «Диаграммы».

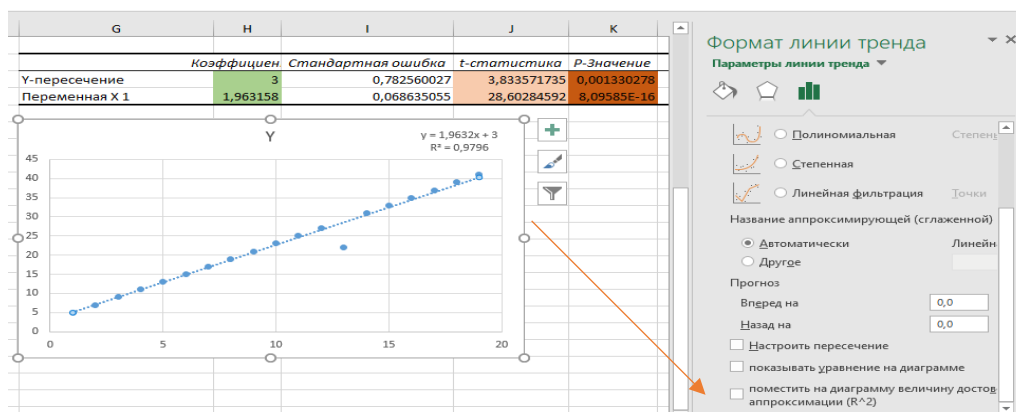
2. Выбрать тип диаграммы, который мы хотим создать, например, «Точечная» для Excel:



Чтобы оформить график в соответствии со своими предпочтениями, можно воспользоваться инструментами для работы с диаграммами в MS Excel. Они позволяют настроить такие элементы графика, как цвета, шрифты, размеры и стили. Выбираем нужный элемент на графике и применяем желаемые изменения.

3. Добавить линейную линию тренда, а также уравнение регрессии и коэффициент детерминации через «Линии тренда», как показано на рисунке:





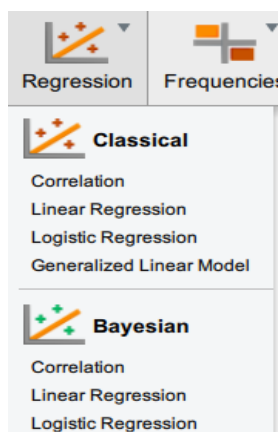
Уравнение регрессии и коэффициент детерминации r^2 находятся в правом нижнем углу диаграммы. Как видно, они такие же, как и при выполнении регрессионного анализа в пакете «Анализ данных – Регрессия».

Алгоритм регрессионного анализа в JASP

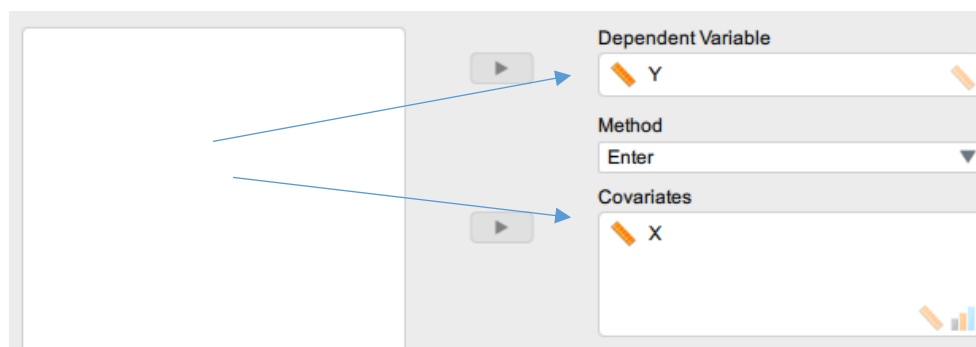
1. Загрузим данные в систему JASP:

	X	Y
1	1	5
2	2	7
3	3	9
4	4	11
5	5	13
6	6	15
7	7	17
8	8	19
9	9	21
10	10	23
11	11	25
12	12	27
13	13	22
14	14	31
15	15	33
16	16	35
17	17	37
18	18	39
19	19	41

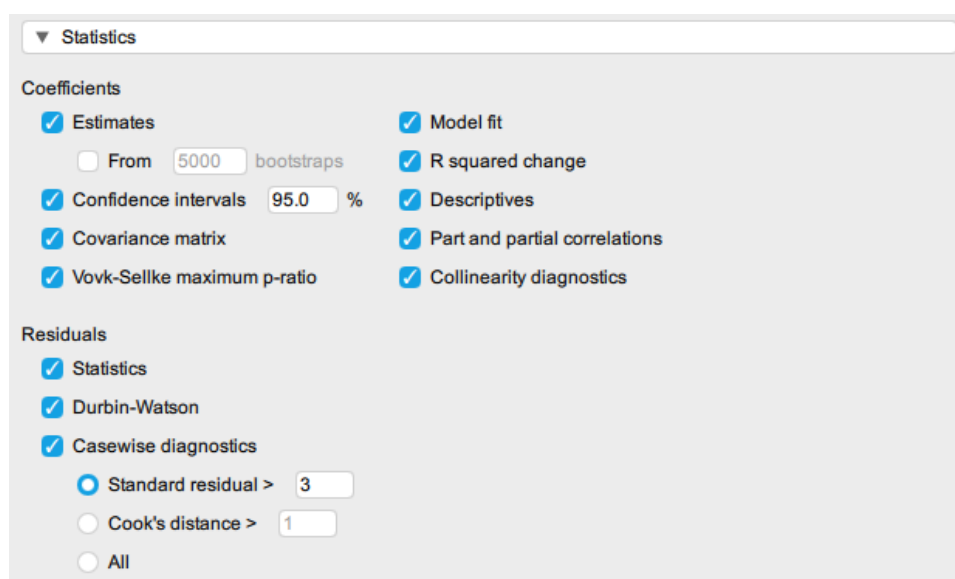
2. Выберем «Linear Regression»:



3. Перенесем переменные x и y :



4. Выставим параметры, как предложено на рисунке:



5. Проведя статистический анализ и построив соответствующие графики, получим следующие данные:

Results ▾

Linear Regression ▾

Model Summary - Y

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p	Durbin-Watson		
										Autocorrelation	Statistic	p
H ₀	0.000	0.000	0.000	11.161	0.000		0	18		0.818	0.076	< .001
H ₁	0.990	0.980	0.978	1.639	0.980	818.123	1	17	< .001	-0.079	2.147	0.949

Здесь видно, что корреляция (r) между двумя переменными высока: $r = 0,990$, а $r^2 = 0,980$. Тест Дурбин–Ватсона проверяет корреляции между остатками, которые могут сделать тест недействительным. Значение r должно быть выше 1 и ниже 3, а в идеале равным около 2.

Таблица «ANOVA» показывает все суммы квадратов, упомянутые ранее. Моделью является регрессия. F-статистика значима: $p < 0,001$:

ANOVA ▼

Model		Sum of Squares	df	Mean Square	F	p	VS-MPR*
H ₁	Regression	2196.774	1	2196.774	818.123	< .001	1.308×10 ⁺¹³
	Residual	45.647	17	2.685			
	Total	2242.421	18				

Note. The intercept model is omitted, as no meaningful information can be shown.

* Vovk-Sellke Maximum p -Ratio: Based on the p -value, the maximum possible odds in favor of H₁ over H₀ equals $1/(-e \cdot p \log(p))$ for $p \leq .37$ (Sellke, Bayarri, & Berger, 2001).

В таблице ниже приведены нестандартизованные коэффициенты, которые можно ввести в линейное уравнение. Здесь показаны как H₀, так и H₁ модели, а также константы (перехват) и коэффициенты регрессии (нестандартизованные) для всех предикторов, включенных в модель:

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	VS-MPR*	95% CI		Collinearity Statistics	
								Lower	Upper	Tolerance	VIF
H ₀	(Intercept)	22.632	2.561		8.838	< .001	382887.643	17.252	28.011		
H ₁	(Intercept)	3.000	0.783		3.834	0.001	41.759	1.349	4.651		
	X	1.963	0.069	0.990	28.603	< .001	1.308×10 ⁺¹³	1.818	2.108	1.000	1.000

* Vovk-Sellke Maximum p -Ratio: Based on the p -value, the maximum possible odds in favor of H₁ over H₀ equals $1/(-e \cdot p \log(p))$ for $p \leq .37$ (Sellke, Bayarri, & Berger, 2001).

ANOVA показывает, что модель значима. Столбцы «Collinearity Statistics» (Tolerance и VIF) проверяют предположение о мультиколлинеарности. Как правило, если $VIF > 10$ и толерантность $< 0,1$, эти предположения сильно нарушаются. Если средний $VIF > 1$ и толерантность $< 0,2$, то модель может быть необъективной.

Таблица описательной статистики показывает количество наблюдений, среднее значение, стандартное отклонение и стандартную ошибку среднего:

Descriptives ▼

	N	Mean	SD	SE
Y	19	22.632	11.161	2.561
X	19	10.000	5.627	1.291

В таблице регистровой диагностики будут выделены все случаи (строки), в которых остатки отклоняются на 3 или более стандартных отклонения от среднего значения:

Part And Partial Correlations

Model		Partial	Part
H ₁	X	0.990	0.990

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients Covariance Matrix

Model		X
H ₁	X	0.005

Note. The intercept model is omitted, as no meaningful information can be shown.

Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions	
				(Intercept)	X
H ₁	1	1.877	1.000	0.061	0.061
	2	0.123	3.907	0.939	0.939

Note. The intercept model is omitted, as no meaningful information can be shown.

Casewise Diagnostics ▼

Case Number	Std. Residual	Y	Predicted Value	Residual	Cook's Distance
13	-4.123	22.000	28.521	-6.521	0.624

Residuals Statistics

	Minimum	Maximum	Mean	SD	N
Predicted Value	4.963	40.300	22.632	11.047	19
Residual	-6.521	0.700	4.163×10^{-17}	1.592	19
Std. Predicted Value	-1.599	1.599	4.101×10^{-17}	1.000	19
Std. Residual	-4.123	0.476	0.005	1.009	19

Эти случаи с наибольшими ошибками вполне могут быть выбросами. Слишком большое количество выбросов окажет влияние на модель.

6. Построим графики, выбрав следующие параметры:

▼ Plots

Residuals Plots

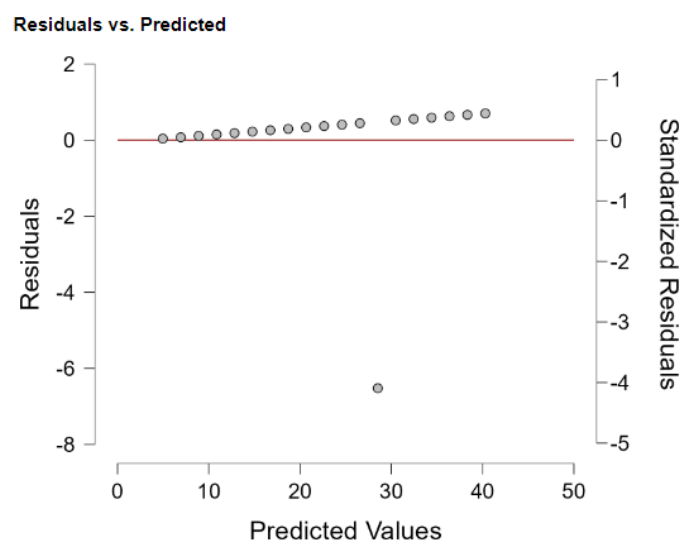
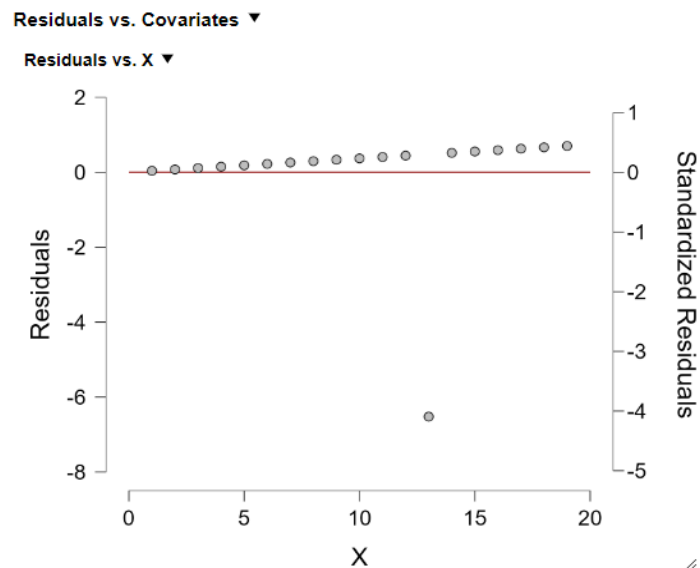
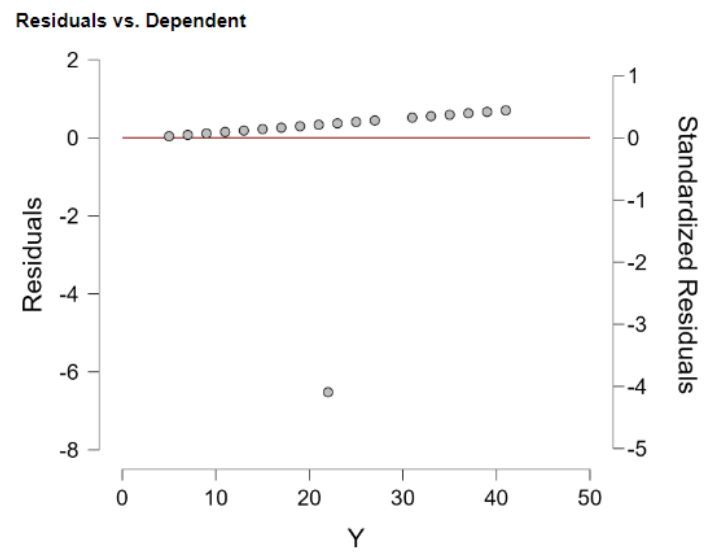
☒ Residuals vs. dependent
☒ Residuals vs. covariates
☒ Residuals vs. predicted
☒ Residuals histogram
☒ Standardized residuals
☒ Q-Q plot standardized residuals
☒ Partial plots

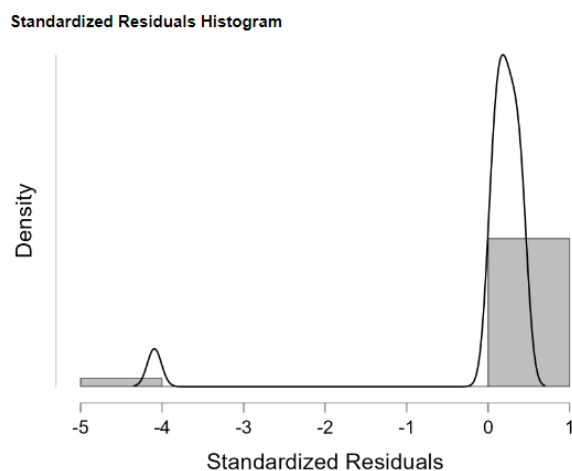
☒ Confidence intervals 95.0 %
☒ Prediction intervals 95.0 %

Other Plots

☒ Marginal effects plots
☒ Confidence intervals 95.0 %
☒ Prediction intervals 95.0 %

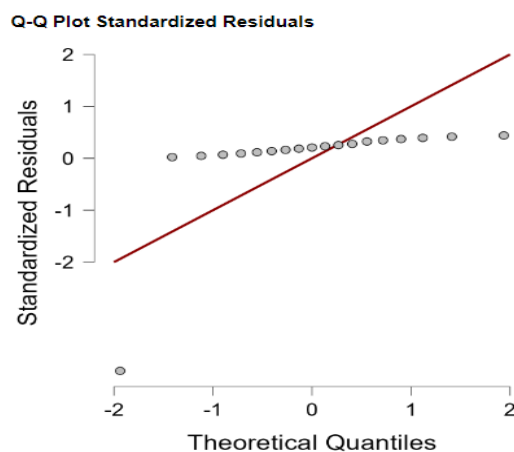
7. Получим следующие графики:





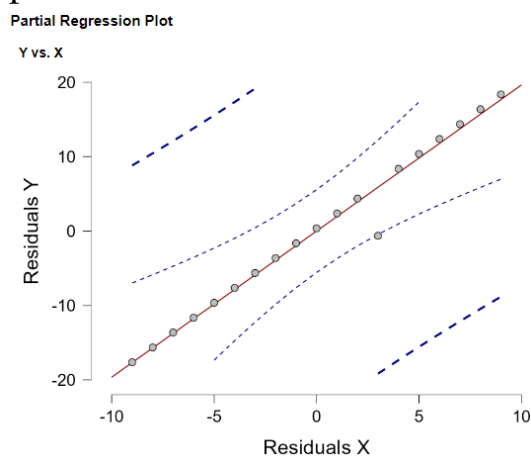
Как видно из рисунков, остатки расположены вокруг базовой линии. Это означает, что можно провести линейный регрессионный анализ.

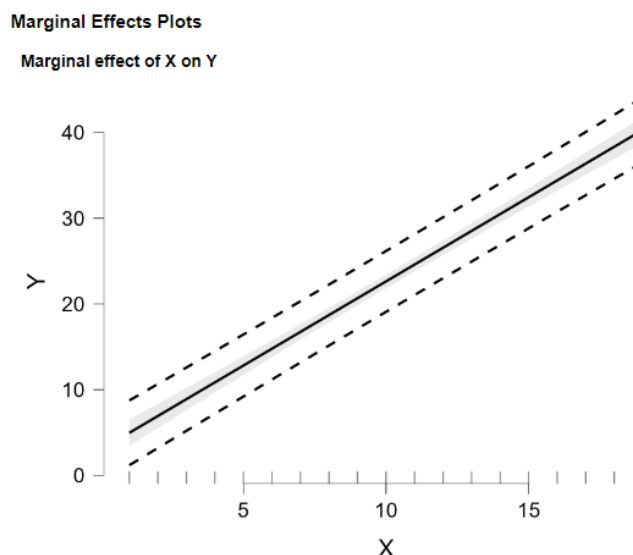
График Q-Q показывает, что стандартизованные остатки не соответствуют диагонали:



Это позволяет предположить, что нормальность или линейность были нарушены.

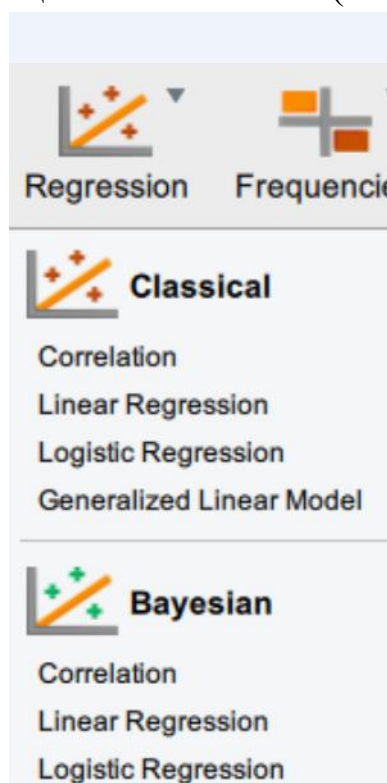
Оценить корректность выполнения регрессионного анализа позволяют следующие графики:





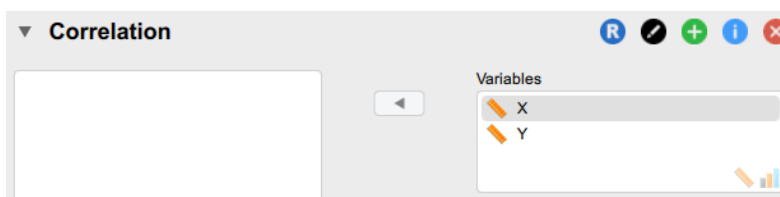
Алгоритм корреляционного анализа в JASP

1. Загрузим исходные данные в JASP.
2. Нажмем на модуль «Регрессия» («Regression»). В выпадающем списке выберем «Корреляционный анализ» («Correlation»):



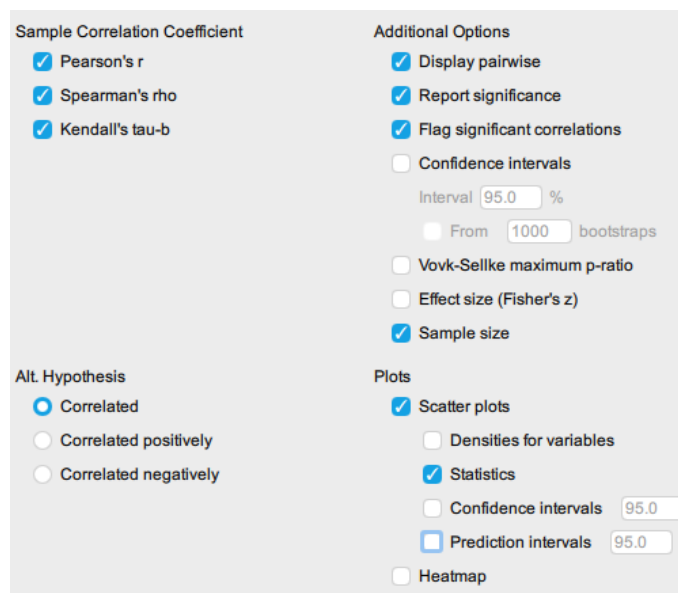
В данном модуле есть дополнительные виды анализа: линейный регрессионный анализ, логистический регрессионный анализ, обобщенное линейное моделирование, разбор которых выходит за пределы настоящего пособия и возможен при самостоятельном освоении.

3. Перенесем переменные x и y .

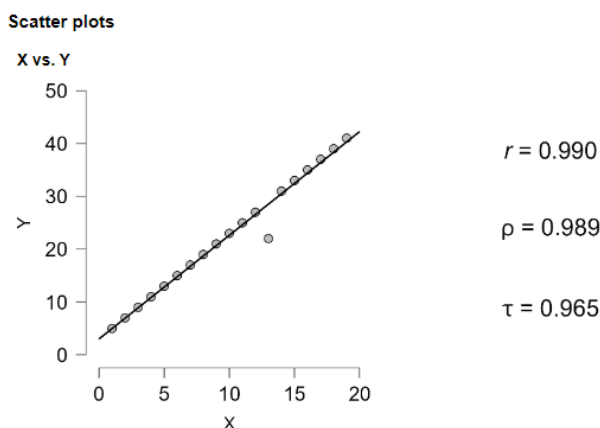


Для выполнения корреляционного анализа необходимы как минимум две переменные. Однако обычно добавляются все необходимые переменные. Дополнительно окно «Partial out» используется совместно с окном «Variables» для выполнения частного корреляционного анализа. В окно выше добавляются переменные, связь с которой нужно «вычесть» из связи между переменными в окне «Variables».

4. Выберем следующие настройки для проведения корреляционного анализа:



5. Получим следующий график:



Выбор критерия зависит от результатов проверки допущений.

Используемые пункты меню анализа:

Sample Correlation Coefficient – позволяет выбрать коэффициенты корреляции. В JASP представлены три варианта:

– **Pearson's r** – коэффициент корреляции Пирсона (параметрический, линейный);

– **Spearman's rho** – коэффициент корреляции Спирмена (непараметрический, ранговый). Лучше использовать для ранговых шкал;

– **Kendall's tau-b** – коэффициент корреляции Кендалла (непараметрический, ранговый). Может использоваться для номинативных шкал.

Display pairwise – позволяет отобразить таблицу, в которой корреляции представлены попарно в виде строчки. Стандартный вариант представления данных корреляционного анализа – квадратная матрица корреляций.

Report significance – позволяет отобразить точные уровни значимости для каждой корреляции. Данный вариант используется редко. Обычно используются астериксы (*) для отображения значимости корреляций.

Flag significance correlation – позволяет пометить значимые корреляции с помощью астерисков (*).

Confidence intervals – доверительные интервалы для корреляций.

Sample size – размер выборки.

Scatter plots – диаграммы рассеивания.

Heatmap – способ визуализации корреляций под названием «Тепловая карта».

Multivariate normality – многомерная оценка эмпирического распределения на соответствие нормальному закону с помощью критерия Шапиро–Уилка.

Pairwise normality – двумерная оценка эмпирического распределения на соответствие нормальному закону пары переменных с помощью критерия Шапиро–Уилка.

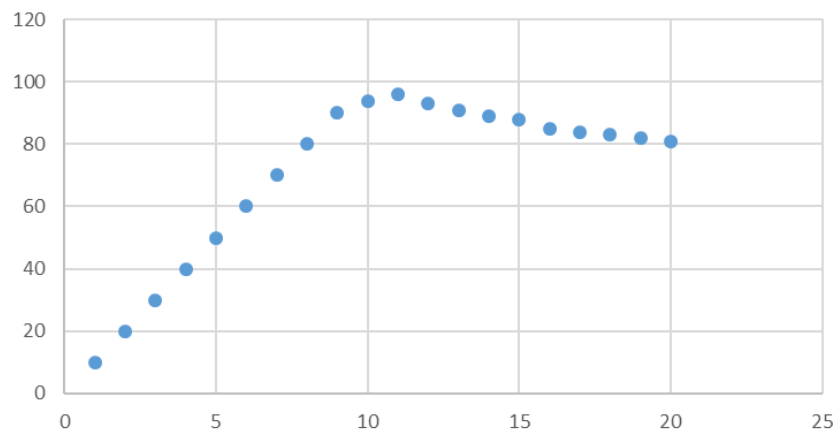
Контрольные вопросы

1. Что такое корреляция?
2. Каковы основные виды корреляции?
3. Как определить наличие корреляции между двумя переменными?
4. Что такое коэффициент корреляции Пирсона?
5. Что такое коэффициент корреляции Спирмена?
6. Каковы правила корреляционного и регрессионного анализа?

2.6. Криволинейная регрессия

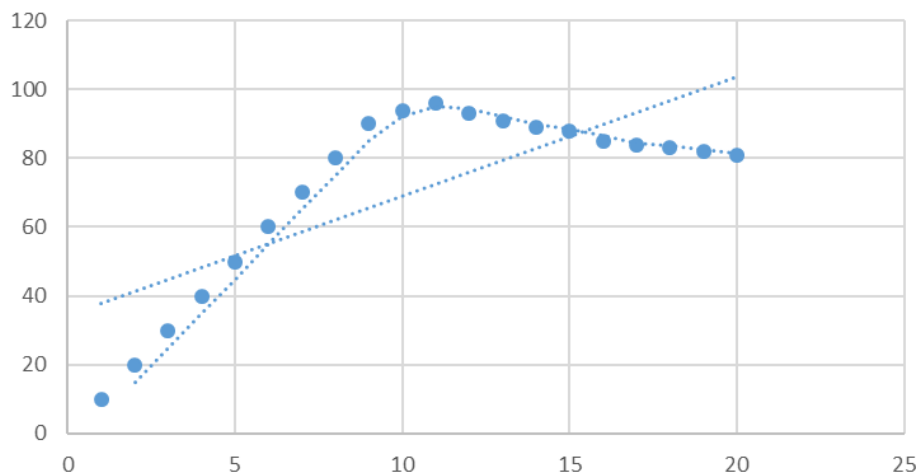
Краткие сведения из теории

Если связь между изучаемыми явлениями не прямая, а кривая (как на рисунке), то коэффициент корреляции не подходит для измерения такой связи:



В такой ситуации коэффициент корреляции может показать отсутствие связи («криволинейности»), хотя на самом деле существует сильная криволинейная зависимость.

Очевидно также, что коэффициент корреляции, рассчитанный на основе предположения о линейной зависимости, а также сама прямая линия не будут отражать реальную зависимость:



Поэтому нужен новый способ измерения, который будет учитывать криволинейную зависимость. Таким способом является корреляционное отношение, обозначаемое греческой буквой η («эта»).

Корреляционное отношение – это статистический показатель, который используется для измерения степени связи между двумя переменными, когда эта связь не является линейной. Оно позволяет учитывать более сложные формы зависимости, такие как кривые линии или нелинейные функции. Корреляционное отношение η может принимать значения от 0 до 1. Чем ближе значение к 1, тем сильнее связь между переменными.

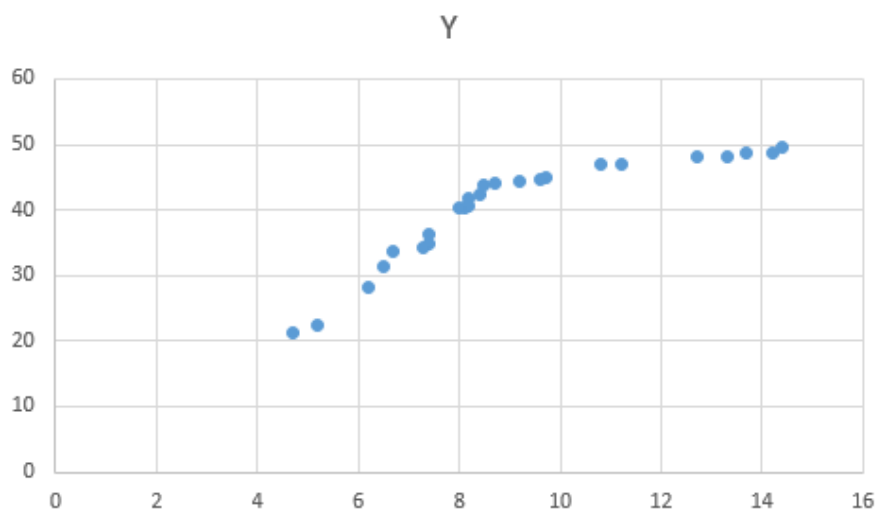
Криволинейная связь между признаками – это тип связи, при котором зависимость между двумя переменными не является линейной. Вместо прямой линии, описывающей связь, криволинейная связь может быть представлена кривой линией или нелинейной функцией. Это означает, что изменение одной переменной не приводит к прямому и пропорциональному изменению другой переменной. Криволинейная связь может быть более сложной и непредсказуемой, чем линейная связь, и требует использования специальных методов для ее измерения и анализа.

Алгоритм анализа в MS Excel

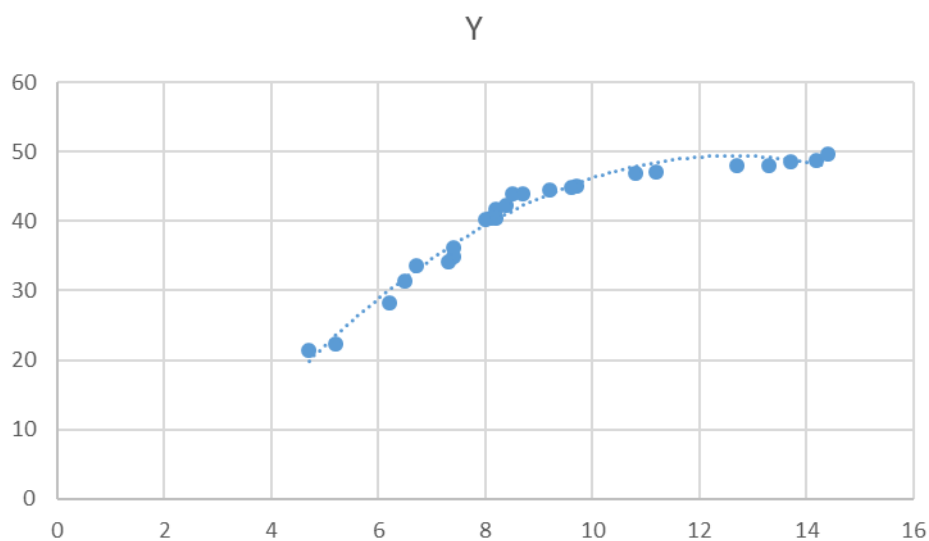
Проведем анализ криволинейной корреляции между дозой (X) экспериментального препарата и гибелью бактерий (Y):

X	Y
14,5	49,8
5,1	22,5
6,1	28,2
6,5	31,2
6,7	33,5
8,2	41,8
7,4	34,9
10,7	46,9
14,3	48,8
8,1	40,3
8,1	40,3
4,6	21,3
8,4	42,2
9,6	44,9
8,8	43,9
9,1	44,3
9,6	44,8
7,4	36,2
8,5	43,9
11,2	47,1
12,7	47,9
13,4	48,2
13,8	48,7
7,2	34,1
8	40,3

1. Скопируем данные в MS Excel. Выделим все значения переменных (всю таблицу с заголовками) и построим график (тип диаграммы – «Точечная»):



2. Выберем «Добавить линию тренда»:



В меню доступны четыре различных типа аппроксимации: логарифмическая, полиномиальная, степенная и экспоненциальная. В данном случае используется полином второй степени, однако стоит отметить, что вид кривой может быть разным и зависит от конкретной задачи. Например, вместо полиномиальной аппроксимации может использоваться логарифмическая зависимость.

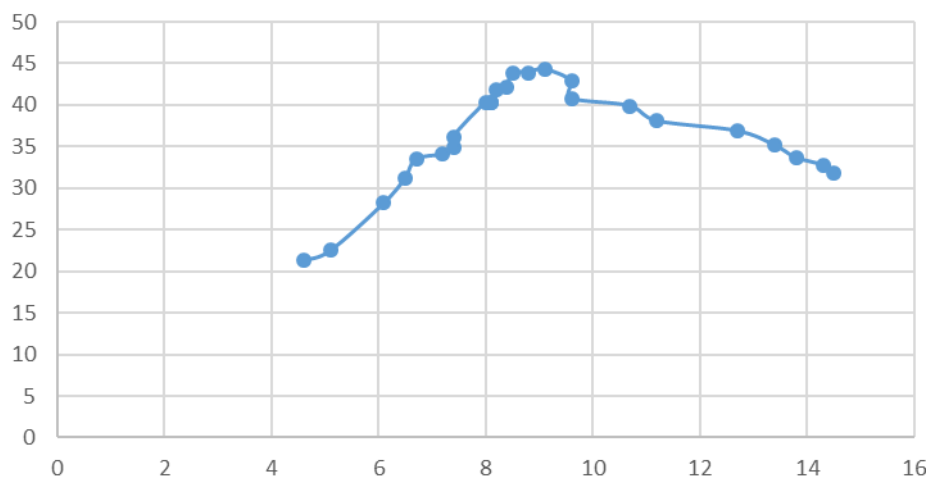
3. Добавим уравнение регрессии и квадрат корреляционного отношения.

Алгоритм анализа в JASP

1. Прежде чем открыть данные в JASP, сделаем криволинейную регрессию более выраженной. Внесем в JASP следующие данные:

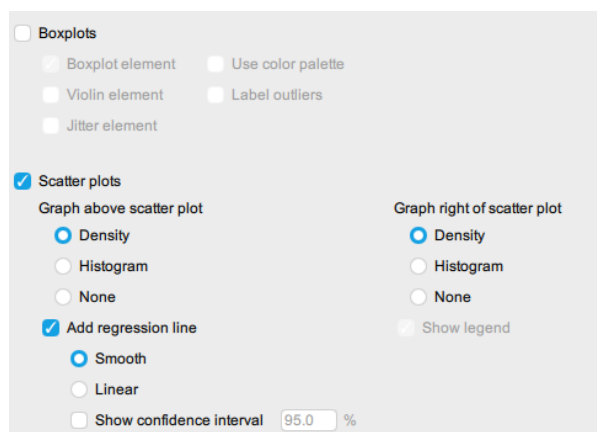
X	Y
4,6	21,3
5,1	22,5
6,1	28,2
6,5	31,2
6,7	33,5
7,2	34,1
7,4	34,9
7,4	36,2
8	40,3
8,1	40,3
8,1	40,3
8,2	41,8
8,4	42,2
8,5	43,9
8,8	43,9
9,1	44,3
9,6	42,9
9,6	40,8
10,7	39,9
11,2	38,1
12,7	36,9
13,4	35,2
13,8	33,7
14,3	32,8
14,5	31,8

На рисунке видно, что мы усилили криволинейную тенденцию:



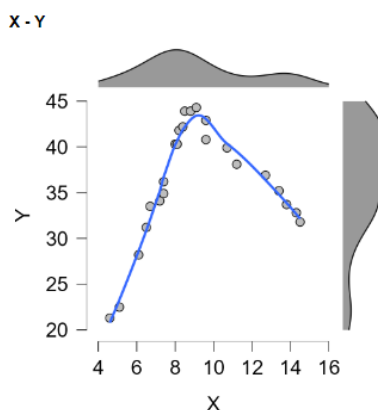
2. Надо учитывать, что в JASP пока нет анализа для нелинейной регрессии. Но мы можем так преобразовать данные, чтобы высчитать данный тип регрессии.

3. Сначала рассмотрим кривую линии регрессии, используя модуль описательной статистики «Descriptives». Выставим следующие настройки:

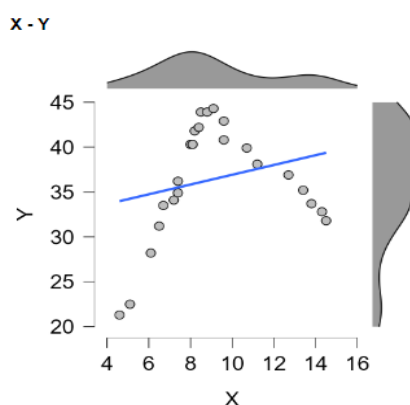


The screenshot shows the 'Descriptives' settings window in JASP. Under the 'Boxplots' section, 'Boxplot element', 'Violin element', and 'Jitter element' are unchecked. 'Use color palette' and 'Label outliers' are also unchecked. Under the 'Scatter plots' section, 'Scatter plots' is checked. For 'Graph above scatter plot', 'Density' is selected. For 'Graph right of scatter plot', 'Density' is also selected. 'Add regression line' is checked, and 'Smooth' is selected under the regression line options. 'Show confidence interval' is checked, and the value is set to 95.0 %.

4. Получим график, который отображает криволинейную зависимость:



5. Если ошибочно выставить иные неверные параметры, можно получить линейную зависимость, которая не будет соответствовать действительности:



6. Теперь проведем классический (линейный) регрессионный анализ. Используя модуль «Regression», выберем линейную регрессию.

7. Выставим соответствующие параметры и перенесем переменные, как показано на рисунке:

The screenshot shows the SPSS Regression dialog box and the Statistics sub-dialog box. In the main dialog, the 'Dependent Variable' is set to 'Y', the 'Method' is 'Enter', and the 'Covariates' are 'X'. The Statistics sub-dialog box is open, showing options for Coefficients and Residuals. Under Coefficients, 'Estimates' is checked, 'Model fit' is checked, 'R squared change' is checked, and 'Descriptives' is checked. Under Residuals, 'Statistics' is checked, 'Durbin-Watson' is checked, and 'Casewise diagnostics' is checked. The 'Standard residual >' is set to 3 and 'Cook's distance >' is set to 1.

8. Получим соответствующие результаты, которые показывают, что нет достоверной связи между переменными:

Model Summary - Y									
Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p
H ₀	0.000	0.000	0.000	6.240	0.000		0	24	
H ₁	0.245	0.060	0.019	6.180	0.060	1.471	1	23	0.238

Однако даже визуально хорошо прослеживается тенденция к росту и последующему спаду. Линейный регрессионный анализ не справляется с задачей, когда данные криволинейны. Однако в рамках линейного анализа мы можем получить на криволинейной модели верный вывод.

Здесь видно, что корреляция (r) между двумя переменными слабая, так как $r = 0,245$ и $r^2 = 0,060$. Таблица «ANOVA» показывает все суммы квадратов, упомянутые ранее; моделью является регрессия:

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H ₁	Regression	56.168	1	56.168	1.471	0.238
	Residual	878.472	23	38.194		
	Total	934.640	24			

Note. The intercept model is omitted, as no meaningful information can be shown.

F-статистика не значима, так как $p = 0,238$:

Coefficients ▼

Model		Unstandardized	Standard Error	Standardized	t	p
H ₀	(Intercept)	36.440	1.248		29.197	< .001
H ₁	(Intercept)	31.472	4.279		7.356	< .001
	X	0.545	0.449	0.245	1.213	0.238

Descriptives

	N	Mean	SD	SE
Y	25	36.440	6.240	1.248
X	25	9.120	2.809	0.562

9. Создадим переменную X^2 :

Name: X2 Long name: X2

Column type: Scale Description: ...

Computed type: Computed with drag-and-drop

Computed column definition Missing Values

$+ - * \div / ^ \sqrt \% = \neq < \leq > \geq \wedge \vee | \neg$

X
Y
X2

X * X

Computed columns code applied

	X	Y	f_x X2
4	6.5	31.2	42.25
5	6.7	33.5	44.89
6	7.2	34.1	51.84
7	7.4	34.9	54.76
8	7.4	36.2	54.76
9	8	40.3	64
10	8.1	40.3	65.61
11	8.1	40.3	65.61

10. После этого вернемся в раздел линейной регрессии и перенесем новую переменную, как показано на рисунке:

The screenshot shows the SPSS 'Linear Regression' dialog box. On the left is a large empty box for the model equation. On the right, the 'Dependent Variable' is set to 'Y'. The 'Method' is set to 'Enter'. Under 'Covariates', the variables 'X' and 'X2' are listed. There are two arrow buttons between the left box and the right panel, and a small bar chart icon at the bottom right of the covariates list.

11. Получим следующие результаты:

Model Summary - Y

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p
H ₀	0.000	0.000	0.000	6.240	0.000		0	24	
H ₁	0.951	0.904	0.895	2.025	0.904	103.013	2	22	< .001

Здесь видно, что корреляция (r) между двумя переменными высока, так как $r = 0,951$ и $r^2 = 0,904$. Таблица «ANOVA» показывает все суммы квадратов, упомянутые ранее; моделью является регрессия:

ANOVA ▼

Model		Sum of Squares	df	Mean Square	F	p
H ₁	Regression	844.465	2	422.233	103.013	< .001
	Residual	90.175	22	4.099		
	Total	934.640	24			

Note. The intercept model is omitted, as no meaningful information can be shown.

F-значение статистически значимо, так как $p < 0,001$:

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H ₀	(Intercept)	36.440	1.248		29.197	< .001
H ₁	(Intercept)	-32.660	4.832		-6.759	< .001
	X	14.680	1.030	6.607	14.255	< .001
	X2	-0.714	0.051	-6.428	-13.868	< .001

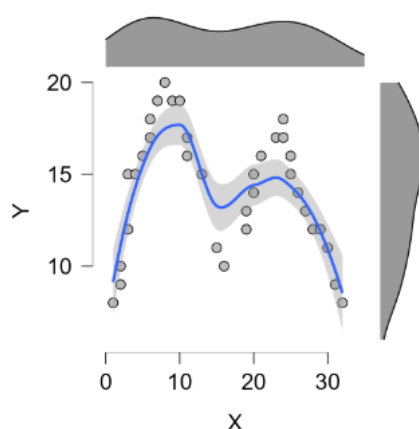
Descriptives ▼

	N	Mean	SD	SE
Y	25	36.440	6.240	1.248
X	25	9.120	2.809	0.562
X2	25	90.747	56.189	11.238

12. Усложним задачу и создадим криволинейную регрессию, которая имеет две точки перегиба. Для этого введем в JASP следующие данные:

X	Y
1	8
2	9
2	10
3	12
3	15
4	15
5	16
6	17
6	18
7	19
7	19
8	20
9	19
10	19
11	17
11	16
13	15
15	11
16	10
19	12
19	13
20	14
20	15
21	16
23	17
24	18
24	17
25	16
25	15
26	14
27	13
28	12
29	12
30	11
31	9
32	8

13. Перейдем в модуль описательной статистики и построим график:



14. Теперь перейдем в модуль линейной регрессии и произведем расчет:

The image shows the SPSS Linear Regression dialog box and the Statistics sub-dialog box. In the main dialog, the Dependent Variable is 'Y' and the Covariates are 'X'. The Method is set to 'Enter'. In the Statistics sub-dialog, the following options are checked: Estimates, Model fit, R squared change, and Confidence intervals (95.0%). Other options like Descriptives, Part and partial correlations, and Collinearity diagnostics are unchecked.

15. Получим следующие данные. Однако будем помнить, что линейная регрессия не подходит для описания данных, имеющих криволинейную зависимость. Поэтому необходимо внести следующую корректировку:

Model Summary - Y

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p
H ₀	0.000	0.000	0.000	3.424	0.000		0	35	
H ₁	0.222	0.049	0.021	3.387	0.049	1.767	1	34	0.193

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H ₁	Regression	20.272	1	20.272	1.767	0.193
	Residual	390.033	34	11.472		
	Total	410.306	35			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H ₀	(Intercept)	14.361	0.571		25.166	< .001
H ₁	(Intercept)	15.565	1.067		14.588	< .001
	X	-0.077	0.058	-0.222	-1.329	0.193

Descriptives

	N	Mean	SD	SE
Y	36	14.361	3.424	0.571
X	36	15.611	9.871	1.645

16. Создадим переменные X^2 и X^3 :

	X	Y	X^2	X^3
1	1	8	1	1
2	2	9	4	8
3	2	10	4	8
4	3	12	9	27
5	3	15	9	27
6	4	15	16	64
7	5	16	25	125

Name: X3 Long name: X3

Column type: Scale Description: ...

Computed type: Computed with drag-and-drop

Computed column definition

Missing Values

$+ \cdot \div / \wedge \sqrt{\%} = \neq < \leq > \geq \wedge \vee | \neg$

X

Y

X2

X3

X ^ 3

Computed columns code applied

17. Вернемся в модуль линейной регрессии и добавим переменные, как показано на рисунке. Зададим настройки для регрессионного анализа:

Dependent Variable

Y

Method

Enter

Covariates

X

X2

X3

18. Получим следующие результаты:

Model Summary - Y

Model	R	R ²	Adjusted R ²	RMSE	R ² Change	F Change	df1	df2	p
H ₀	0.000	0.000	0.000	3.424	0.000		0	35	
H ₁	0.624	0.390	0.333	2.797	0.390	6.818	3	32	0.001

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H ₁	Regression	159.998	3	53.333	6.818	0.001
	Residual	250.308	32	7.822		
	Total	410.306	35			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients ▼

Model		Unstandardized	Standard Error	Standardized	t	p
H ₀	(Intercept)	14.361	0.571		25.166	< .001
H ₁	(Intercept)	9.364	1.948		4.806	< .001
	X	1.479	0.530	4.264	2.791	0.009
	X2	-0.082	0.037	-7.723	-2.203	0.035
	X3	0.001	7.414×10 ⁻⁴	3.276	1.548	0.131

Descriptives

	N	Mean	SD	SE
Y	36	14.361	3.424	0.571
X	36	15.611	9.871	1.645
X2	36	338.444	323.893	53.982
X3	36	8309.778	9772.368	1628.728

Контрольные вопросы

1. Что такое криволинейная регрессия?
2. Чем она отличается от линейной регрессии?
3. Какие существуют методы криволинейной регрессии?
4. Что такое полиномиальная регрессия?
5. В каких случаях используется логарифмическая регрессия?

2.7. Логистическая регрессия

Краткие сведения из теории

Логистическая регрессия – это статистический метод, который оценивает вероятность возникновения события (например, выжить или умереть) на основе заданного набора данных независимых переменных. Этот тип статистической модели (также известный как логит-модель) часто используется для классификации и прогнозной аналитики. Поскольку результат является вероятностью, зависимая переменная ограничена диапазоном от 0 до 1. В логистической регрессии к шансам применяется логит-преобразование, т. е. отношение вероятности успеха к вероятности неудачи. Это также широко известно как *логарифм шансов* или *натуральный логарифм шансов*:

$$\text{Logit}(pi) = \frac{1}{(1 + \exp(-pi))};$$

$$\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1 * X_1 + \dots + \text{Beta}_k * X_k,$$

где $\text{logit}(pi)$ – зависимая переменная, или переменная отклика, а x – независимая переменная.

Бета-параметр, или бета-коэффициент, в этой модели обычно оценивается с помощью метода максимального правдоподобия (MLE). Этот метод тестирует различные бета-значения посредством нескольких итераций, чтобы оптимизировать наилучшее соответствие шансов. Все эти итерации создают логарифмическую функцию правдоподобия, а логистическая регрессия стремится максимизировать эту функцию, чтобы найти наилучшую оценку параметра. Как только оптимальный коэффициент (или коэффициенты, если существует более одной независимой переменной) найден, условные вероятности для каждого наблюдения можно рассчитать, записать в журнал и суммировать, чтобы получить прогнозируемую вероятность. Для двоичной классификации вероятность менее 0,5 будет предсказывать 0, а вероятность больше 0,5 – 1. После расчета модели рекомендуется оценить, насколько хорошо модель предсказывает зависимую переменную. Тест Хосмера–Лемешоу – популярный метод оценки соответствия модели.

Логарифмические шансы могут быть трудными для понимания в рамках анализа данных логистической регрессии. В результате возведение в степень бета-оценок является обычным явлением для преобразования результатов в отношение шансов (Odds Ratio – OR), что упрощает интерпретацию результатов. OR представляет собой вероятность того, что

результат произойдет при определенном событии, по сравнению с вероятностью того, что результат произойдет в отсутствие этого события. Если OR больше 1, то событие связано с более высокими шансами на получение определенного результата. И наоборот, если OR меньше 1, то событие связано с более низкой вероятностью наступления этого результата. Основываясь на приведенном выше уравнении, интерпретацию отношения шансов можно обозначить следующим образом: шансы на успех изменяются в $\exp(sB_1)$ раз для каждого увеличения x на s -единицу. В качестве примера предположим, что нам нужно оценить шансы человека на выживание на «Титанике», учитывая, что этот человек — мужчина, а отношение шансов для мужчин, к примеру, составляло 0,0810. Мы бы интерпретировали отношение шансов так, что шансы на выживание мужчин уменьшились в 0,0810 раза по сравнению с женщинами, оставив все остальные переменные постоянными.

Как линейная, так и логистическая регрессия, являются одними из самых популярных моделей в области науки о данных, а инструменты с открытым исходным кодом, такие как Python и R, делают вычисления для них быстрыми и простыми.

Модели линейной регрессии используются для определения взаимосвязи между непрерывной зависимой переменной и одной или несколькими независимыми переменными. Когда есть только одна независимая переменная и одна зависимая переменная, это называется простой линейной регрессией, но по мере увеличения количества независимых переменных ее называют множественной линейной регрессией. Для каждого типа линейной регрессии строится линия наилучшего соответствия через набор точек данных, который обычно рассчитывается с использованием метода наименьших квадратов.

Подобно линейной регрессии, логистическая регрессия также используется для оценки взаимосвязи между зависимой переменной и одной или несколькими независимыми переменными, но она используется для прогнозирования категориальной переменной по сравнению с непрерывной. Категориальная переменная может быть истинной или ложной, ответами «да» или «нет», 1 или 0 и т. д. Единица измерения также отличается от линейной регрессии, поскольку она дает вероятность, но логарифмическая функция преобразует S-образную кривую в прямую линию.

Хотя обе модели используются в регрессионном анализе для прогнозирования будущих результатов, линейную регрессию обычно легче понять. Кроме того, линейная регрессия не требует большого размера выборки, в то время как логистическая регрессия требует адекватной выборки для представления значений по всем категориям ответов. Без более крупной и репрезентативной выборки модель может не обладать достаточной статистической мощностью для обнаружения значительного эффекта.

Существует три типа моделей логистической регрессии, которые определяются на основе категориального ответа.

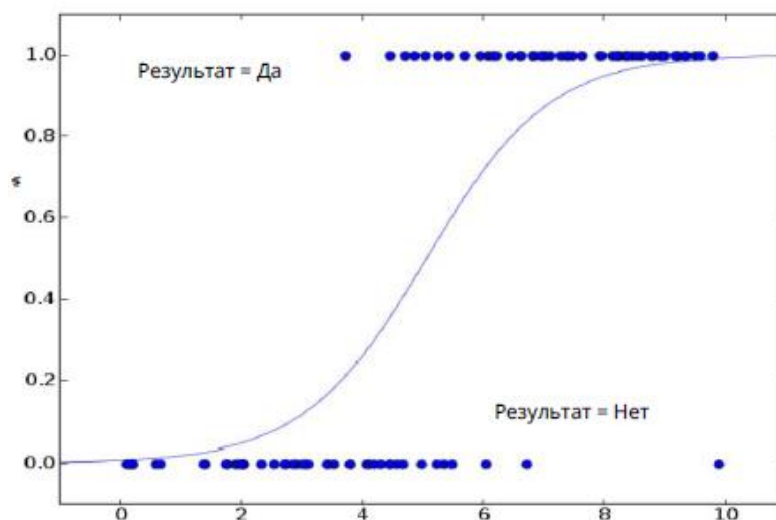
1. *Бинарная логистическая регрессия.* В этой модели ответ, или зависимая переменная, являются дихотомическими по своей природе, т. е. имеют только два возможных результата (например, 0 или 1). Некоторые популярные примеры ее использования: прогнозирование того, является ли опухоль злокачественной или нет; есть ли у человека депрессия или нет и т. д. В логистической регрессии это наиболее часто используемый подход и, в более общем смысле, один из наиболее распространенных классификаторов для бинарной классификации.

2. *Полиномиальная логистическая регрессия.* В модели логистической регрессии этого типа зависимая переменная имеет три или более возможных результата, однако эти значения не имеют определенного порядка. С помощью данной модели исследователи могут изучать влияние различных факторов окружающей среды на выживаемость организмов. Например, они могут рассмотреть такие факторы, как температура, влажность и освещенность, чтобы определить, какие комбинации этих факторов приводят к наибольшей выживаемости организма.

3. *Порядковая логистическая регрессия.* Этот тип модели логистической регрессии используется, когда переменная ответа имеет три или более возможных результата, но в этом случае эти значения обладают определенным порядком. Примеры порядковых ответов включают оценочные шкалы (скажем, от А до F) или рейтинговые шкалы (от 1 до 6).

Как видно на графике ниже, линия линейной регрессии между ответами «да» и «нет» не имеет смысла в качестве модели прогнозирования. Вместо этого S-образная кривая логистической регрессии имеет минимум 0 и максимум 1. Можно видеть, что некоторые значения предикторов перекрываются между «да» и «нет». Например, значение

прогноза, равное 5, даст равную 50 % вероятность положительного или отрицательного результата:



Таким образом, пороговые значения рассчитываются, чтобы определить, будет ли значение данного предиктора классифицировано как результат «да» или оно будет классифицировано как результаты «нет».

Условия для бинарной логистической регрессии:

а) зависимая переменная должна быть бинарной, т. е., к примеру, «да» или «нет», мужской или женский пол, хороший или плохой;

б) одна или несколько независимых переменных-предсказателей могут быть непрерывными или категориальными;

в) должна быть связь между любыми непрерывными независимыми переменными и логит-преобразованием (естественным логарифмом вероятности того, что результат соответствует одной из категорий) зависимой переменной.

Рассмотрим метрики логистической регрессии.

AIC (информационные критерии Акаике) и *BIC (байесовские информационные критерии)* являются мерами соответствия модели: лучшая модель будет иметь самые низкие значения AIC и BIC.

Четыре псевдозначения r^2 , рассчитанные в JASP, McFadden, Nagelkerke, Tjur и Cox & Snell, аналогичны r^2 в линейной регрессии, и все дают разные значения. То, что представляет собой хорошее значение r^2 , варьируется, однако оно полезно при сравнении разных моделей для одних и тех же данных. Наилучшей считается модель с наибольшей статистикой r^2 .

Тест Вальда используется для определения статистической значимости каждой из независимых переменных.

Матрица путаницы представляет собой таблицу, показывающую фактические и прогнозируемые результаты, и может использоваться для определения точности модели.

Чувствительность – это процент случаев, в которых наблюдаемый результат был правильно предсказан моделью (т. е. истинно положительные результаты).

Специфичность – это процент наблюдений, которые также были правильно предсказаны как не имеющие наблюдаемого результата (т. е. истинно отрицательные результаты).

Алгоритм анализа в MS Excel

1. Введем следующие данные на рабочий лист в MS Excel:

№	Уровень СРБ, мг/л	Летальный исход
1	83	нет
2	94	нет
3	117	нет
4	119	нет
5	120	нет
6	123	нет
7	132	нет
8	132	да
9	136	нет
10	139	нет
11	140	да
12	141	нет
13	145	нет
14	146	да
15	146	да
16	146	нет
17	151	нет
18	151	да
19	162	нет
20	163	да
21	165	нет
22	168	да
23	170	да
24	172	да
25	173	да
26	173	нет
27	188	да
28	199	да
29	202	да
30	204	да

В таблице представлены данные о смертности от пневмонии в зависимости от уровня С-реактивного белка (СРБ). Известно, что при пневмонии легкой степени тяжести уровень СРБ составляет 50–60 мг/л, средней степени тяжести – 90–110 мг/л, тяжелой степени – 130–150 мг/л. Неблагоприятным признаком тяжелого течения, показанием к интенсификации антибактериальной и дезинтоксикационной терапии является уровень СРБ выше 150 мг/л.

2. Поскольку в модели у нас есть одна объясняющая переменная (СРБ), создадим ячейки для одного коэффициента регрессии и одного коэффициента для точки пересечения в модели. Зададим значения коэффициентов, равные 0,001, и оптимизируем их:

	А	В
	Уровень СРБ, мг/л	Летальный исход
1		
2	83	нет
3	94	нет
4	117	нет
5	119	нет
6	120	нет
7	123	нет
8	132	нет
9	132	да
29	199	да
30	202	да
31	204	да
32		
33	0,001	
34	0,001	

3. Создадим столбец «logit», используя формулу, приведенную на рисунке:

C2			$=\$A\$33+\$A\$34*A2$
	А	В	С
	Уровень СРБ, мг/л	Летальный исход	Logit
1			
2	83	нет	0,084
3	94	нет	0,095
4	117	нет	0,118
5	119	нет	0,12
6	120	нет	0,121

Далее необходимо создать несколько новых столбцов, которые мы будем использовать для оптимизации этих коэффициентов регрессии, включая $e\text{logit}$, вероятность и логарифмическую вероятность.

4. Создадим столбец «eLogit» и рассчитаем для него значения по формуле, показанной на рисунке:

D2 X ✓ fx =EXP(C2)				
	A	B	C	D
1	Уровень СРБ, мг/л	Летальный исход	Logit	eLogit
2	83	нет	0,084	1,087629
3	94	нет	0,095	1,099659
4	117	нет	0,118	1,125244

5. Создадим столбец вероятности («Possibility») и рассчитаем для него значения по формуле, показанной на рисунке (но прежде заменим значения «да» на 1, а значения «нет» – на 0):

E2 X ✓ fx =ЕСЛИ(B2=1;D2/(1+D2);1-(D2/(1+D2)))						
	A	B	C	D	E	F
1	Уровень СРБ, мг/л	Летальный исход	Logit	eLogit	Probability	Log Likelihood
2	83	0	0,084	1,087629	0,479012	-0,736028921
3	94	0	0,095	1,099659	0,476268	-0,741774882

6. Создадим значения для логарифмической вероятности и рассчитаем для него значения по формуле, показанной на рисунке:

F2 X ✓ fx =LN(E2)						
	A	B	C	D	E	F
1	Уровень СРБ, мг/л	Летальный исход	Logit	eLogit	Probability	Log Likelihood
2	83	0	0,084	1,087629	0,479012	-0,736028921
3	94	0	0,095	1,099659	0,476268	-0,741774882

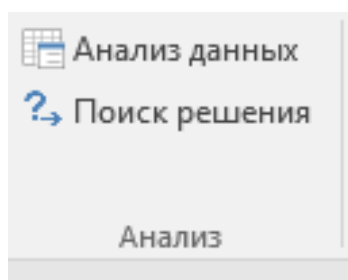
7. Найдем сумму логарифмических вероятностей по формуле, показанной на рисунке:

F32 X ✓ fx =СУММ(F2:F31)						
	A	B	C	D	E	F
28	188	1	0,189	1,208041	0,54711	-0,603105676
29	199	1	0,2	1,221403	0,549834	-0,598138869
30	202	1	0,203	1,225072	0,550576	-0,596789485
31	204	1	0,205	1,227525	0,551071	-0,595891133
32						-20,77991966
33	0,001					
34	0,001					

8. Найдем коэффициенты регрессии, используя модуль «Поиск решения».

Если ранее этот модуль не был установлен в MS Excel, необходимо выполнить следующие действия: щелкнуть «Файл», перейти в «Параметры», открыть «Настройки», нажать «Поиск решения», затем «Перейти». В новом всплывающем окне установить флажок рядом с «Поиск решения», нажать «ОК».

После установки перейти в группу «Анализ» на вкладке «Данные» и нажать «Поиск решения»:



Далее откроется следующее окно:

Введем следующую информацию:

Установим цель: выберем ячейку F32, содержащую сумму логарифмических вероятностей.

Путем изменения ячеек переменных выберем диапазон ячеек A33:A34, который содержит коэффициенты регрессии. Выставим «Минимум».

Сделаем неограниченные переменные неотрицательными, т. е. снимем этот флажок.

Выберем метод решения: поиск решений нелинейных задач методом ОПГ.

Затем нажмем «Найти решение».

9. Получим следующие коэффициенты: $b_0 = -10,2276516$ и $b_1 = 0,066770117$.

Текущие коэффициенты регрессии по умолчанию определяют вероятность отсутствия летального исхода. Однако обычно в логистической регрессии нас интересует вероятность того, что зависимая переменная = 1. Таким образом, мы можем просто поменять знаки на каждом из коэффициентов регрессии.

Теперь эти коэффициенты регрессии можно использовать для определения вероятности события в рамках этой регрессионной модели.

Алгоритм анализа в JASP

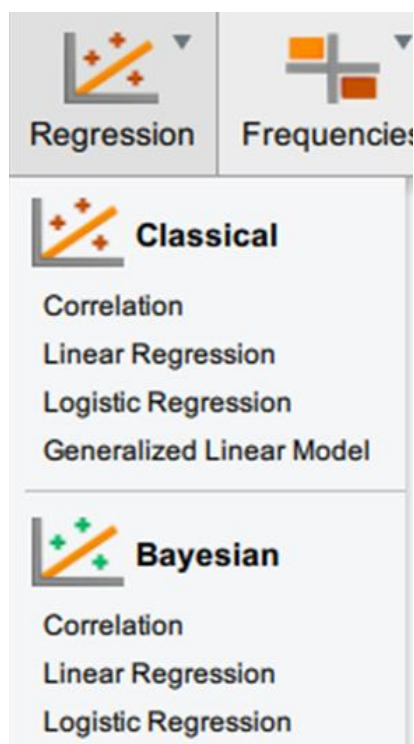
1. Загрузим представленные ниже данные в JASP. По сравнению с предыдущей задачей добавлена еще одна переменная – хронические заболевания легких:

№	Уровень СРБ, мг/л	Хронические заболевания легких	Летальный исход
1	83	нет	нет
2	94	нет	нет
3	117	нет	нет
4	119	нет	нет
5	120	нет	нет
6	123	нет	нет
7	132	нет	нет
8	132	да	да
9	136	нет	нет
10	139	нет	нет
11	140	да	да
12	141	нет	нет
13	145	нет	нет
14	146	да	да
15	146	да	да
16	146	нет	нет
17	151	да	нет
18	151	да	да
19	162	нет	нет
20	163	нет	да
21	165	да	нет

22	168	да	да
23	170	нет	да
24	172	да	да
25	173	да	да
26	173	да	нет
27	188	да	да
28	199	нет	да
29	202	нет	да
30	204	да	да

	 №	 СРБ, мг/л	 Хронические заболевания легких	 Летальный исход
1	1	83	нет	нет
2	2	94	нет	нет
3	3	117	нет	нет
4	4	119	нет	нет
5	5	120	нет	нет
6	6	123	нет	нет
7	7	132	нет	нет
8	8	132	да	да
9	9	136	нет	нет
10	10	139	нет	нет

2. Выберем «Logistic Regression»:



3. Перенесем переменные, как показано на рисунке, и оставим классический (т.е. принудительный – «Enter») ввод данных:

The screenshot shows a software interface for building a regression model. On the left is a large empty box labeled 'No' with a small icon. To its right are three buttons with right-pointing arrows. On the right side, there are three sections: 'Dependent Variable' with 'Летальный исход' (Fatal outcome) selected; 'Method' with 'Enter' selected in a dropdown menu; and 'Covariates' with 'СРБ, мг/л' (CRP, mg/l) selected. Below these is a 'Factors' section with 'Хронические заболевания легких' (Chronic lung diseases) selected. Each section has a small icon of a person and a bar chart.

Если предикторы некоррелированы, их порядок входа мало влияет на модель. В большинстве случаев переменные-предикторы в некоторой степени коррелируют, и поэтому порядок ввода предикторов может иметь значение. Различные методы ввода данных являются предметом многочисленных дискуссий в этой области.

Принудительный ввод («Enter»), или ввод по умолчанию: все предикторы принудительно вводятся в модель в том порядке, в котором они появляются в поле «Covariates». Этот метод считается лучшим.

Блочный ввод («Hierarchical Enter»): исследователь, обычно на основе предварительных знаний и предыдущих исследований, решает, в каком порядке сначала вводятся известные предикторы, в зависимости от их важности для прогнозирования результата. Дополнительные предикторы добавляются на дальнейших этапах.

Пошаговый обратный ввод («Backward»): все предикторы изначально вводятся в модель, а затем рассчитывается вклад каждого. Предикторы с уровнем вклада менее заданного ($p < 0,1$) удаляются. Этот процесс повторяется до тех пор, пока все предикторы не станут статистически значимыми.

Пошаговый прямой ввод («Forward»): первым вводится предиктор с наибольшей простой корреляцией с выходной переменной. Последующие предикторы выбираются на основе размера их частичной

корреляции с исходной переменной. Это повторяется до тех пор, пока в модель не будут включены все предикторы, которые вносят в нее значительную уникальную дисперсию.

Пошаговый ввод («Stepwise»): то же, что и метод «Enter», за исключением того, что каждый раз, когда в модель добавляется предиктор, выполняется проверка на удаление наименее полезного предиктора. Модель постоянно переоценивается, чтобы увидеть, можно ли удалить какие-либо избыточные предикторы. Однако сообщается о многих недостатках использования пошагового метода ввода данных.

Методы входа могут быть полезны для изучения ранее неиспользованных предикторов или для точной настройки модели для выбора лучших предикторов из доступных вариантов.

4. Настроим следующие параметры для статистического анализа логистической регрессии:

5. Получим следующие результаты:

Model Summary - Летальный исход

Model	Deviance	AIC	BIC	df	X ²	p	McFadden R ²	Nagelkerke R ²	Tjur R ²	Cox & Snell R ²
H ₀	41.455	43.455	44.857	29						
H ₁	24.072	30.072	34.276	27	17.383	< .001	0.419	0.587	0.484	0.440

Как видно из этой таблицы, H₁ (с самыми низкими показателями AIC и BIC) предполагает значительную связь ($X^2(27) = 17,383$, $p < 0,001$) между результатом (летальный исход) и предикторскими переменными (уровень СРБ и хронические заболевания легких). Это также подтверждает и коэффициент Макфаддена (McFadden) R₂, равный 0,419. Предполагается, что диапазон от 0,2 до 0,4 указывает на хорошее соответствие модели.

И уровень СРБ в крови, и сопутствующие хронические заболевания легких являются значимыми прогностическими переменными ($p = 0,022$ и $0,047$ соответственно). Наиболее важными значениями в таблице коэффициентов являются отношения шансов (OR):

Coefficients							
	Estimate	Standard Error	Odds Ratio	z	Wald Test		
					Wald Statistic	df	p
(Intercept)	-8.227	4.079	2.674×10^{-4}	-2.017	4.068	1	0.044
СРБ, мг/л	0.060	0.026	1.062	2.282	5.208	1	0.022
Хронические заболевания легких (нет)	-2.039	1.027	0.130	-1.984	3.936	1	0.047

Note. Летальный исход level 'да' coded as class 1.

Для непрерывного предиктора отношение шансов больше 1 предполагает положительную связь, тогда как меньше 1 подразумевает отрицательную связь. Это говорит о том, что высокий уровень СРБ в значительной степени связан с повышенным летальным исходом. Отсутствие хронических заболеваний легких связано со значительно сниженной вероятностью летального исхода. Отношение шансов 0,13 можно интерпретировать как вероятность летального исхода только в 13 % случаев отсутствия хронических заболеваний легких:

Factor Descriptives

Хронические заболевания легких	N
да	13
нет	17

Performance metrics

	Value
Sensitivity	0.786
Specificity	0.813

Матрица путаницы показывает, что модель предсказала 13 истинных отрицательных и истинных положительных случаев, в то время как ошибки (ложноотрицательные и ложноположительные результаты) были обнаружены в трех случаях:

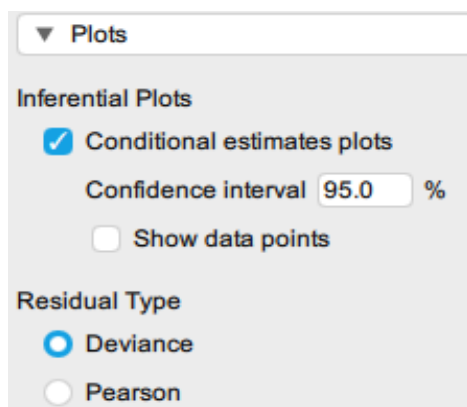
Confusion matrix

Observed	Predicted		% Correct
	нет	да	
нет	13	3	81.250
да	3	11	78.571
Overall % Correct			80.000

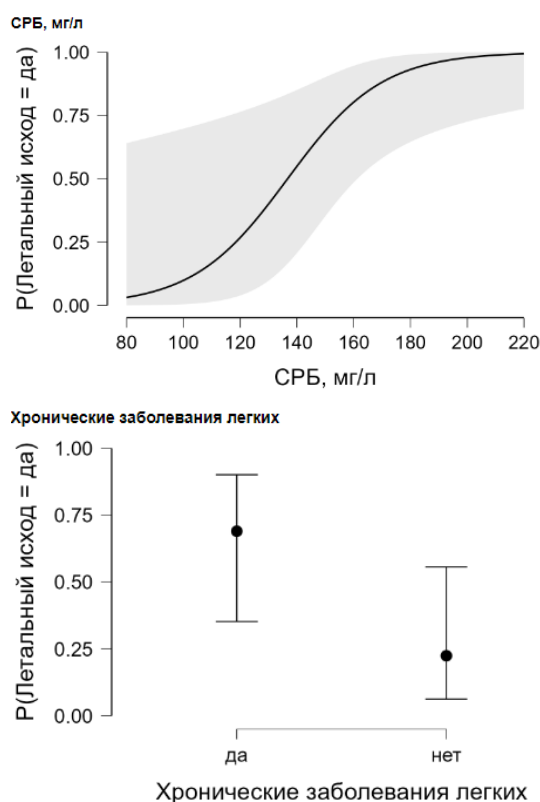
Note. The cut-off value is set to 0.5

Это подтверждается показателями эффективности, где чувствительность (% случаев, для которых результат был правильно предсказан) и специфичность (% случаев, правильно предсказанных как не имеющие исхода, т. е. истинно отрицательные результаты) составляют 78 и 80 % соответственно.

6. Представим полученные данные графически:



7. Получим следующие графики:



По мере увеличения уровня СРБ увеличивается вероятность летального исхода.

Наличие хронических заболеваний легких увеличивает вероятность летального исхода.

Контрольные вопросы

1. Что такое логистическая регрессия?
2. В чём отличие логистической регрессии от линейной регрессии?
3. Какие задачи решает логистическая регрессия?
4. Как работает логистическая регрессия?
5. Каковы преимущества использования логистической регрессии?
6. Какие есть ограничения у логистической регрессии?
7. Как оценить качество модели логистической регрессии?
8. Как интерпретировать коэффициенты логистической регрессии?
9. Как определить количество предикторов при построении модели логистической регрессии?

2.8. Анализ качественных признаков

Краткие сведения из теории

В предыдущих разделах мы анализировали количественные признаки. Эти признаки измеримы, что позволяет производить над ними арифметические действия. Кроме того, их можно упорядочить, расположив в порядке возрастания или убывания. Однако есть признаки, которые невозможно измерить. Это качественные признаки. Они не связаны между собой никакими арифметическими соотношениями, упорядочить их также нельзя.

Качественные признаки в биостатистике – это признаки, которые описывают свойства или характеристики объекта исследования, но не могут быть измерены количественно. Они представляют собой нечисловые данные, которые могут быть представлены в виде категорий или номинальных переменных.

Примеры качественных признаков: пол, возрастная группа, раса, образование, профессия, наличие или отсутствие определенного заболевания и т. д. Эти признаки не имеют числового значения и не могут быть упорядочены по величине.

Категории качественных признаков могут быть взаимоисключающими (например, пол – мужской или женский) или не взаимоисключающими (например, цвет глаз, раса и т. д.).

Качественные признаки используются в биостатистике для описания и классификации объектов исследования. Они позволяют выделить группы объектов с общими характеристиками и провести анализ их распределения. Например, можно исследовать распределение пола или возраста в группе пациентов с определенным заболеванием.

Качественные признаки также могут быть использованы для проведения анализа ассоциаций. Например, можно исследовать связь между полом и риском развития определенного заболевания.

Важно отметить, что качественные признаки не могут быть подвергнуты арифметическим операциям, таким как сложение или вычитание. Они могут быть только классифицированы и сопоставлены.

Качественные признаки можно описать, подсчитывая количество объектов, имеющих одинаковое свойство, а также определяя долю от общего числа объектов, приходящуюся на каждое свойство.

Если в группе исследуемых объектов одна часть обладает одним признаком, а другая часть – другим, то можно вычислить долю (p) или процент от общего числа объектов в группе, которые составляют объекты с тем или иным признаком. Например, если в группе из 100 человек 30 женщин и 70 мужчин, то доля (процент) женщин в группе составит $30/100 = 0,3$ или 30 %, а доля мужчин – 0,7 или 70 %. Важно отметить, что группы могут состоять не только из двух подгрупп. Но для описания совокупности, состоящей из двух подгрупп, достаточно указать численность одной из них. Если доля одной подгруппы во всей совокупности равна p , то доля другой подгруппы равна $1 - p$.

Или если известно общее число членов группы (N) и число членов группы с интересующим признаком (M), то долю (p) этих членов можно выразить формулой

$$p = \frac{M}{N},$$

или в процентном соотношении:

$$p = \frac{M}{N} \cdot 100.$$

Доля p аналогична среднему μ по совокупности.

Довольно часто необходимо сравнить между собой две группы, характеризующиеся определенным качественным признаком. Для проведения этой процедуры используется критерий z .

Критерий z аналогичен критерию Стьюдента t :

$$z = \frac{\text{Разность выборочных долей}}{\text{Стандартная ошибка разности выборочных долей}},$$

или

$$z = \frac{p_1 - p_2}{s_{p_1} - s_{p_2}} = \frac{p_1 - p_2}{\sqrt{s_{p_1}^2 + s_{p_2}^2}},$$

где p_1 и p_2 – доля исследуемого признака в первой и во второй группах соответственно.

Если хотя бы для одной выборки условие np и $n(1-p)$ больше 5 (где n – объем выборки) не выполняется, то критерий z не может быть применен и следует применить точный критерий Фишера:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

где n_1 и n_2 – объемы двух выборок.

О статистически значимом различии долей можно говорить, если значение z окажется «большим».

Критическое значение z -критерия определяется по таблице критических значений в справочных материалах по биостатистике в зависимости от количества членов выборки (вычисляется число степеней свободы) и выбранного уровня значимости. Однако, если $z_{кр}$ сравнивать с критическим значением критерия Стьюдента, то $t_{кр}$ подчиняется распределению Стьюдента, а $z_{кр}$ – стандартному нормальному распределению. При увеличении числа степеней свободы распределение Стьюдента приближается к нормальному, и критические значения z можно найти в последней строке таблицы критических значений для критерия Стьюдента.

Для компенсации излишнего «оптимизма» критерия z была введена поправка Йетса, также известная как поправка на непрерывность. С учетом этой поправки выражение для z принимает следующий вид:

$$z = \frac{|p_1 - p_2| - \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

Таблицы сопряженности – это инструмент, используемый в статистике для представления и анализа данных, когда значения двух переменных приведены в табличном виде.

Таблица сопряженности представляет собой таблицу $m \times n$, где m – значения первой переменной, n – значения второй переменной. В таблице ниже $m = 2$ (антибиотик: да или нет), $n = 2$ (рецидив: да или нет). Такие таблицы называются таблицами сопряженности 2×2 .

В таблице 2×2 каждая переменная имеет два возможных значения, и эти значения можно представить в виде четырех ячеек таблицы. В каждой ячейке указывается количество наблюдений, в которых обе переменные принимают определенные значения.

Пример таблицы сопряженности:

Антибиотик	Рецидив	
	да	нет
да	2	12
нет	15	4

Для анализа таблиц сопряженности используется критерий χ^2 (читается «хи-квадрат»). Этот критерий является статистическим тестом и позволяет проверить гипотезу о независимости двух переменных, т. е. о том, что одна переменная не влияет на другую. Если значение критерия χ^2 оказывается значимым, то мы можем отвергнуть нулевую гипотезу о независимости переменных и сделать вывод о наличии связи между ними.

Однако важно помнить, что критерий χ^2 имеет ряд ограничений и предположений, которые должны быть выполнены для его корректного использования. Например, он предполагает, что данные в таблице распределены случайным образом и что каждая ячейка таблицы имеет достаточное количество наблюдений.

Для понимания смысла анализа таблиц сопряженности необходимо учитывать, что если переменные между собой не связаны, то значения в таблице сопряженности (ожидаемые значения), при равной численности в группах, будут примерно одинаковыми. Иными словами, исходя из вышеприведенного примера, количество заболевших не будет зависеть от наличия или отсутствия прививки.

Однако в нашей таблице значения в ячейках (наблюдаемые значения) существенно отличаются, что наводит на мысль о возможной статистической значимости этих различий. На сравнении ожидаемых и наблюдаемых значений и основан критерий χ^2 .

Критерий χ^2 не требует никаких предположений относительно параметров совокупности, из которой извлечены выборки, – это непараметрический критерий.

Определяется критерий χ^2 следующим образом:

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

где O – наблюдаемое число в ячейке таблицы сопряженности, E – ожидаемое число в той же ячейке. Суммирование проводится по всем ячейкам таблицы.

Нулевая гипотеза в подобных задачах звучит следующим образом: переменные не связаны между собой, т. е. являются независимыми, видимые различия в ячейках таблицы сопряженности случайны и статистически незначимы.

Полученное значение критерия χ^2 характеризует значимость этих различий. Расчетное значение критерия сравнивается с критическим значением критерия χ^2 (значение находится по таблице для заданного уровня значимости) и если полученное значение критерия χ^2 больше критического, то нулевая гипотеза об отсутствии связи между переменными отклоняется.

Критерий χ^2 может быть применен, если ожидаемое число в любой из ячеек таблицы сопряженности больше или равно 5. В противном случае, необходимо использовать точный критерий Фишера. Критическое значение χ^2 зависит от размеров таблицы сопряженности, т. е. от числа строк и числа столбцов. Размер таблицы выражается числом степеней свободы v :

$$v = (r - 1) \cdot (c - 1),$$

где r – число строк, а c – число столбцов.

Формула для χ^2 в случае таблицы 2×2 (т. е. при 1 степени свободы) дает несколько завышенные значения, что может привести к слишком частому отвержению нулевой гипотезы. Это связано с тем, что теоретическое распределение χ^2 непрерывно, в то время как набор вычисленных значений χ^2 дискретен. Чтобы компенсировать этот эффект, используют поправку Йетса:

$$\chi^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E}.$$

Для таблиц большего размера критерий χ^2 применим, если все ожидаемые числа не меньше 1 и доля ячеек с ожидаемыми числами меньше 5 не превышает 20 %. Если эти условия не выполняются, критерий χ^2 может дать ложные результаты. В таких случаях можно собрать дополнительные данные или объединить несколько строк или столбцов таблицы.

Порядок применения критерия χ^2 :

1. Построить таблицу сопряженности, используя полученные данные.
2. Подсчитать количество объектов в каждой строке и в каждом столбце, определить долю этих величин от общего числа объектов.
3. Используя эти доли, вычислить ожидаемые числа, т. е. количество объектов, которые попали бы в каждую ячейку таблицы, если бы связь между строками и столбцами отсутствовала.
4. Определить величину, характеризующую различия между наблюдаемыми и ожидаемыми значениями. Если таблица сопряженности имеет размер 2×2 , применить поправку Йетса.
5. Вычислить число степеней свободы и выбрать уровень значимости.
6. Определить значение χ^2 . Сравнить полученное значение с критическим значением.
7. Если полученное значение критерия χ^2 больше критического, то нулевая гипотеза об отсутствии связи между переменными отклоняется, в противном случае мы остаемся в рамках нулевой гипотезы.

Точный критерий Фишера является статистическим методом, который также используется для проверки гипотез о независимости двух переменных. Он часто применяется в анализе таблиц сопряженности, где каждая строка и столбец представляют собой категории одной или двух переменных.

Критерий Фишера основан на предположении, что две переменные являются независимыми. Если это предположение неверно, то частоты в таблице сопряженности будут отличаться от тех, которые ожидаются при независимости.

Для применения критерия Фишера необходимо выполнить следующие шаги:

1. Рассчитать ожидаемые частоты для каждой ячейки таблицы сопряженности, исходя из предположения о независимости переменных.
2. Рассчитать разность между наблюдаемой и ожидаемой частотой для каждой ячейки.
3. Возвести разности в квадрат и разделить на ожидаемые частоты.
4. Сложить все полученные значения.
5. Сравнить полученную сумму с критическим значением χ^2 , которое зависит от уровня значимости и числа степеней свободы.

Если сумма больше критического значения, то гипотеза о независимости переменных отвергается. Если сумма меньше критического значения, то гипотеза не отвергается.

Обозначения, используемые в точном критерии Фишера:

Сумма по строкам			
Сумма по столбцам	O_{11}	O_{12}	R_1
	O_{21}	O_{22}	R_2
	C_1	C_2	N

Точный критерий Фишера рассчитывается по формуле

$$p = \frac{R_1! R_2! C_1! C_2!}{N! \cdot O_{11}! O_{12}! O_{21}! O_{22}!},$$

где R_1 и R_2 – суммы по строкам (например, число больных, лечившихся тем или иным способом), C_1 и C_2 – суммы по столбцам (например, число больных с тем или иным исходом заболевания). O_{11} , O_{12} , O_{21} и O_{22} – числа в ячейках таблицы, N – общее число наблюдений. Восклицательный знак обозначает факториал. (Факториал числа есть произведение всех целых чисел от этого числа до единицы: $n! = n \cdot (n-1) \cdot (n-2) \cdot 2 \cdot 1$. Например, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. Факториал нуля равен единице.)

Построив все остальные варианты заполнения таблицы, возможные при данных суммах по строкам и столбцам, по этой же формуле рассчитывают их вероятность. Вероятности, которые не превосходят вероятность исходной таблицы (включая саму эту вероятность), суммируют. Полученная сумма – это величина p для двустороннего варианта точного критерия Фишера.

В отличие от критерия χ^2 , существуют одно- и двусторонний варианты точного критерия Фишера.

Задача

Исследователи сравнивают эффективность нового антибиотика в лечении инфекции мочевыводящих путей у мужчин в возрасте 20 до 30 лет. После короткого курса приема антибиотика делали посевы мочи. При выявлении бактериурии констатировали наличие инфекции. Были получены следующие результаты:

№	Фамилия	Пол	Год рождения	Антибиотик	Рецидив	Дата исследования
1	Антонов	муж.	1995	да	нет	10.03.2022
2	Аданов	жен.	1996	нет	да	10.03.2022
3	Алманов	жен.	1996	да	нет	12.03.2022
4	Аревыев	муж.	1997	да	да	15.03.2022

5	Бирунов	жен.	1995	нет	да	15.03.2022
6	Болданов	муж.	1991	нет	да	16.03.2022
7	Болиев	муж.	1997	да	нет	14.03.2022
8	Воронов	муж.	1995	нет	да	13.03.2022
9	Геранов	жен.	1994	да	да	19.03.2022
10	Гулянов	муж.	1996	да	нет	15.03.2022
11	Исаев	муж.	1991	нет	да	14.03.2022
12	Пирунов	муж.	1997	нет	да	22.03.2022
13	Сидоловнов	муж.	1994	нет	да	25.03.2022
14	Икашкин	муж.	1995	нет	да	26.03.2022
15	Болданов	муж.	1997	нет	да	20.03.2022
16	Семюк	муж.	1991	нет	да	17.03.2022
17	Инаков	муж.	1997	нет	да	27.03.2022
18	Пинаков	муж.	1995	нет	да	23.03.2022
19	Сидоров	муж.	1995	нет	да	10.03.2022
20	Ивашкин	муж.	1997	нет	да	12.03.2022
21	Богданов	муж.	1991	нет	да	14.03.2022
22	Семаков	муж.	1997	нет	да	10.03.2022
23	Колнов	муж.	1995	нет	да	12.03.2022
24	Лосен	муж.	1995	нет	да	10.03.2022
25	Лысенков	муж.	1997	нет	да	15.03.2022
26	Лысев	муж.	1991	да	да	10.03.2022
27	Суровый	муж.	1997	да	да	11.03.2022
28	Чирмаков	муж.	1995	да	да	15.03.2022
29	Командов	муж.	1995	да	да	16.03.2022
30	Ломаков	муж.	1997	нет	нет	17.03.2022
31	Ларсен	муж.	1991	нет	нет	13.03.2022
32	Лаптьев	муж.	1997	нет	нет	14.03.2022
33	Сорматив	муж.	1995	нет	нет	17.03.2022
34	Чимаков	муж.	1995	нет	нет	18.03.2022
35	Перельтьев	муж.	1997	нет	нет	19.03.2022
36	Носатов	муж.	1997	нет	нет	20.03.2022
37	Сиволап	муж.	1995	да	нет	21.03.2022
38	Степнов	муж.	1995	да	нет	21.03.2022
39	Степкин	муж.	1997	да	нет	24.03.2022
40	Пуля	муж.	1991	да	нет	23.03.2022
41	Нансен	муж.	1997	да	нет	29.03.2022
42	Султан	муж.	1995	да	нет	28.03.2022
43	Сульянов	муж.	1995	да	нет	26.03.2022
44	Симикин	муж.	1997	да	нет	26.03.2022

Необходимо построить таблицу сопряжения и провести анализ с помощью χ^2 .

Алгоритм анализа в MS Excel

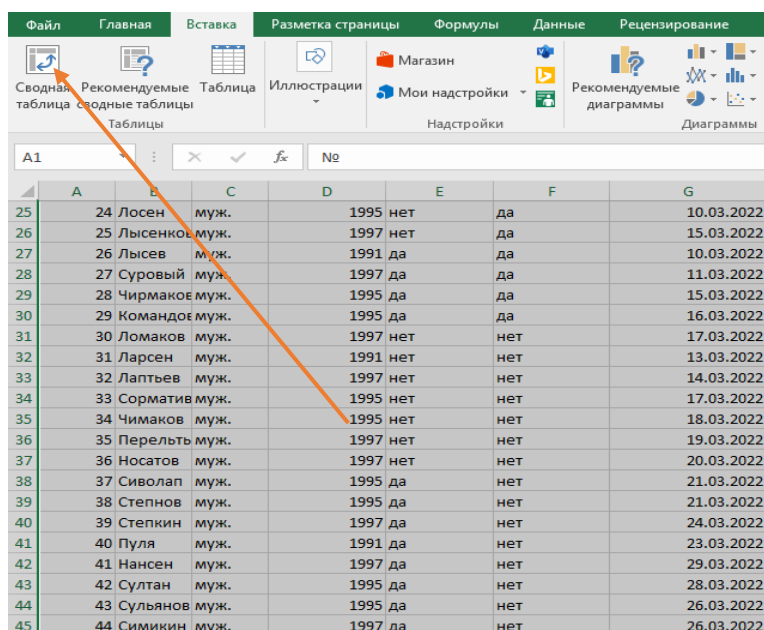
Сформулируем нулевую гипотезу: препарат не влияет на уровень рецидива. И примем уровень значимости равным 0,05.

Для решения поставленной задачи необходимо создать таблицу сопряженности, которая представляет собой двумерную матрицу, где строки и столбцы соответствуют двум переменным. В нашем случае это будет таблица 2x2. Для ее формирования можно использовать возможности сводной таблицы в MS Excel.

1. Скопируем исходную таблицу в программу MS Excel и выделим всю таблицу:

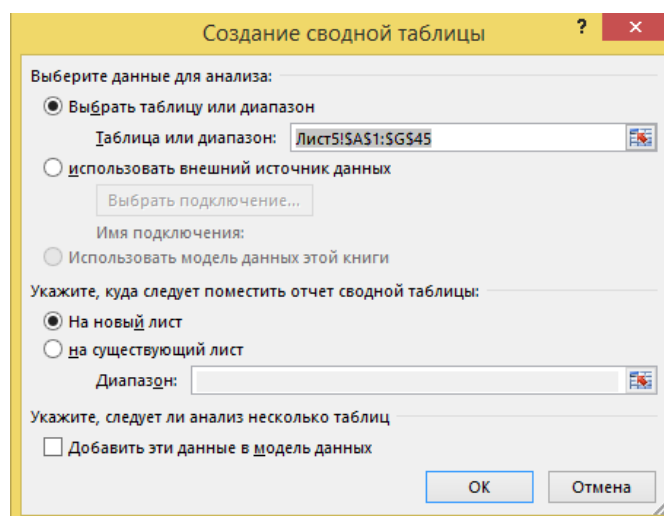
	A	B	C	D	E	F	G
1	№	Фамилия	Пол	Год рождения	Антибиотик	Рецидив	Дата исследования
2	1	Антонов	муж.	1995	да	нет	10.03.2022
3	2	Аданов	жен.	1996	нет	да	10.03.2022
4	3	Алманов	жен.	1996	да	нет	12.03.2022
5	4	Аревыев	муж.	1997	да	да	15.03.2022
6	5	Бирунов	жен.	1995	нет	да	15.03.2022
7	6	Болданов	муж.	1991	нет	да	16.03.2022
8	7	Болиев	муж.	1997	да	нет	14.03.2022
9	8	Воронов	муж.	1995	нет	да	13.03.2022
10	9	Геранов	жен.	1994	да	да	19.03.2022
11	10	Гулянов	муж.	1996	да	нет	15.03.2022
12	11	Исаев	муж.	1991	нет	да	14.03.2022
13	12	Пирунов	муж.	1997	нет	да	22.03.2022
14	13	Сидоловнов	муж.	1994	нет	да	25.03.2022
15	14	Икашкин	муж.	1995	нет	да	26.03.2022
16	15	Болданов	муж.	1997	нет	да	20.03.2022
17	16	Семюк	муж.	1991	нет	да	17.03.2022
18	17	Инаков	муж.	1997	нет	да	27.03.2022
19	18	Пинаков	муж.	1995	нет	да	23.03.2022
20	19	Сидоров	муж.	1995	нет	да	10.03.2022
21	20	Ивашкин	муж.	1997	нет	да	12.03.2022
22	21	Богданов	муж.	1991	нет	да	14.03.2022
23	22	Семаков	муж.	1997	нет	да	10.03.2022

2. Откроем вкладку «Вставка» – «Сводная таблица»:

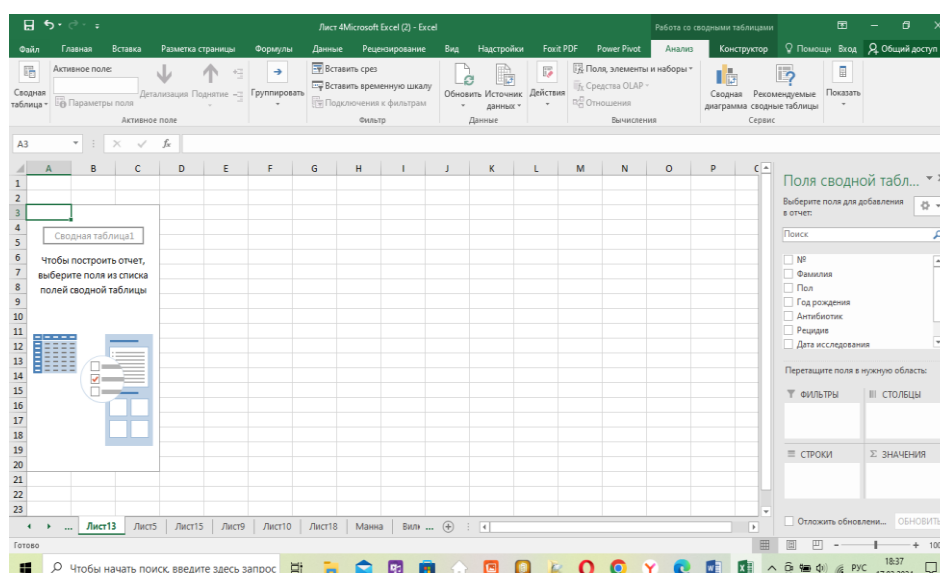


25	24	Лосен	муж.	1995	нет	да	10.03.2022
26	25	Лысенков	муж.	1997	нет	да	15.03.2022
27	26	Лысев	муж.	1991	да	да	10.03.2022
28	27	Суровый	муж.	1997	да	да	11.03.2022
29	28	Чирмаков	муж.	1995	да	да	15.03.2022
30	29	Командо	муж.	1995	да	да	16.03.2022
31	30	Ломаков	муж.	1997	нет	нет	17.03.2022
32	31	Ларсен	муж.	1991	нет	нет	13.03.2022
33	32	Лаптьев	муж.	1997	нет	нет	14.03.2022
34	33	Сорматив	муж.	1995	нет	нет	17.03.2022
35	34	Чимаков	муж.	1995	нет	нет	18.03.2022
36	35	Перельт	муж.	1997	нет	нет	19.03.2022
37	36	Носатов	муж.	1997	нет	нет	20.03.2022
38	37	Сиволап	муж.	1995	да	нет	21.03.2022
39	38	Степнов	муж.	1995	да	нет	21.03.2022
40	39	Степкин	муж.	1997	да	нет	24.03.2022
41	40	Пуля	муж.	1991	да	нет	23.03.2022
42	41	Нансен	муж.	1997	да	нет	29.03.2022
43	42	Султан	муж.	1995	да	нет	28.03.2022
44	43	Сульянов	муж.	1995	да	нет	26.03.2022
45	44	Симикин	муж.	1997	да	нет	26.03.2022

3. Выберем позиции, как показано на рисунке:



4. В появившемся окне сводной таблице будут отражены названия полей, столбцов и значений так, как показано на рисунке:



Сводная таблица в MS Excel позволяет создавать таблицы сопряженности, которые могут быть полезны для анализа данных. Внешний вид и содержимое сводной таблицы зависят от того, что мы хотим отобразить или проследить в своих данных. Мы можем настроить сводную таблицу, чтобы она отображала только те данные, которые нам нужны, и представляла их в удобном для нас формате:

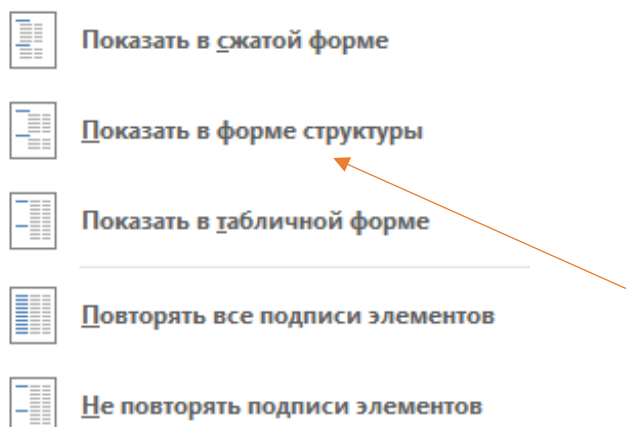
Количество по полю		Фамилия		Названия столбцов	
Названия строк					
да	нет	Общий итог			
да	нет	6	12	18	
нет	нет	19	7	26	
Общий итог		25	19	44	

В нашем случае необходимо отобразить количество людей с рецидивом, принимающих антибиотик. Исследуемый объект имеет два признака: рецидив есть или нет (рецидив: да/нет) и применялся ли антибиотик или нет (антибиотик: да/нет):

5. Для улучшения визуализации сводной таблицы в MS Excel мы можем изменить ее макет на «Показать в форме структуры». Это делается следующим образом:

- а) кликнуть по сводной таблице, чтобы открыть вкладку «Конструктор»;
- б) в группе «Параметры» выбрать «Макет отчета»;
- в) из выпадающего списка выбрать «Показать в форме структуры».

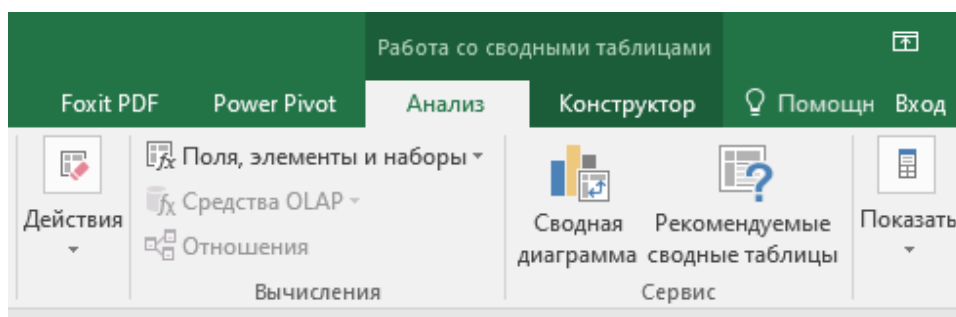
Это позволит представить нашу сводную таблицу в более структурированном и понятном виде, что облегчит восприятие данных:



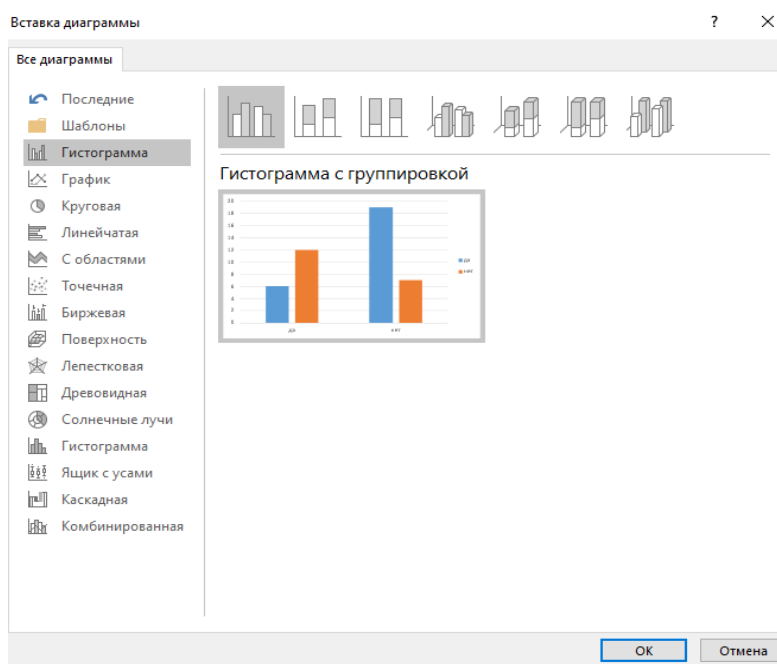
Количество по полю Фамилия		Рецидив		
Антибиотик		да	нет	Общий итог
да		6	12	18
нет		19	7	26
Общий итог		25	19	44

Использование сводной таблицы в MS Excel позволяет значительно упростить процесс создания таблицы сопряженности, особенно когда данных много. Полученную таблицу 2x2 можно было бы также создать вручную, но использование сводной таблицы позволяет автоматизировать этот процесс и сэкономить время.

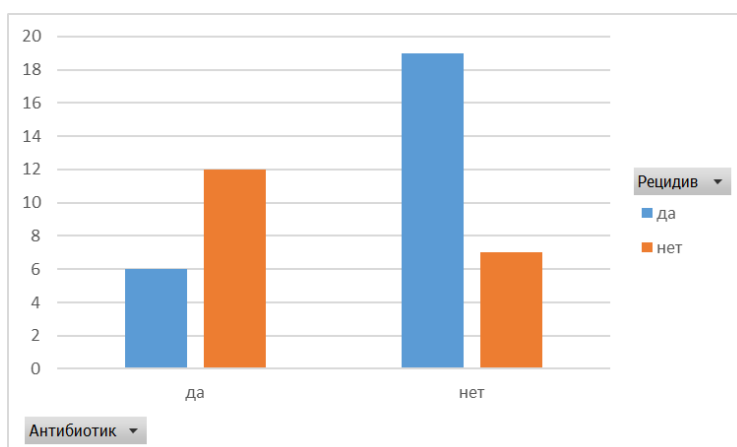
6. Предварительно выделим таблицу сопряженности и построим три типа сводных диаграмм, используя опции «Анализ» – «Сводная диаграмма»:



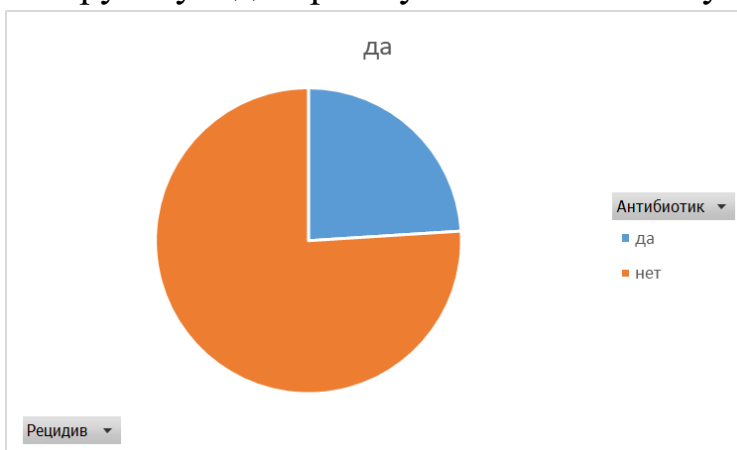
7. Выберем «Гистограмма с группировкой»:



8. Получим следующий результат:



9. Построим круговую диаграмму по аналогичному алгоритму:

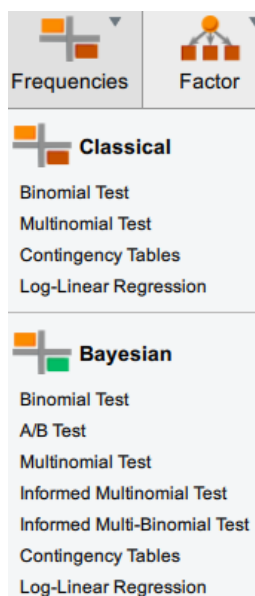


Алгоритм анализа в JASP

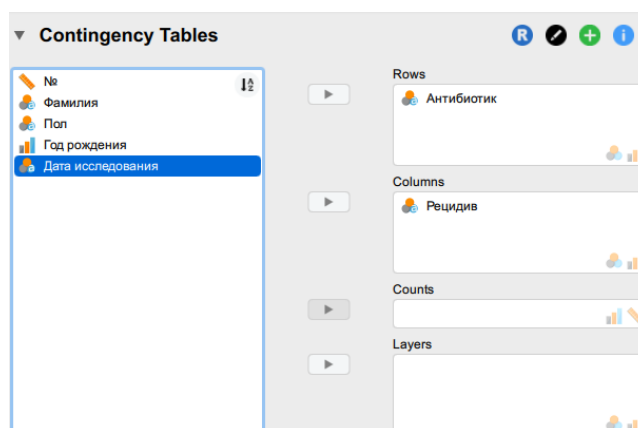
1. Откроем JASP и загрузим первичные данные для анализа:

	№	Фамилия	Пол	Год рождения	Антибиотик	Рецидив	Дата исследования
1	1	Антонов	муж.	1995	да	нет	10.03.2022
2	2	Аданов	жен.	1996	нет	да	10.03.2022
3	3	Алманов	жен.	1996	да	нет	12.03.2022
4	4	Ареьев	муж.	1997	да	да	15.03.2022
5	5	Бирунов	жен.	1995	нет	да	15.03.2022
6	6	Богданов	муж.	1991	нет	да	16.03.2022
7	7	Болиев	муж.	1997	да	нет	14.03.2022
8	8	Воронов	муж.	1995	нет	да	13.03.2022
9	9	Геранов	жен.	1994	да	да	19.03.2022
10	10	Гулянов	муж.	1996	да	нет	15.03.2022
11	11	Исаев	муж.	1991	нет	да	14.03.2022
12	12	Пирунов	муж.	1997	нет	да	22.03.2022
13	13	Сидоповнов	муж.	1994	нет	да	25.03.2022
14	14	Икашкин	муж.	1995	нет	да	26.03.2022
15	15	Богданов	муж.	1997	нет	да	20.03.2022
16	16	Семюк	муж.	1991	нет	да	17.03.2022
17	17	Инаков	муж.	1997	нет	да	27.03.2022
18	18	Пинаков	муж.	1995	нет	да	23.03.2022
19	19	Сидоров	муж.	1995	нет	да	10.03.2022

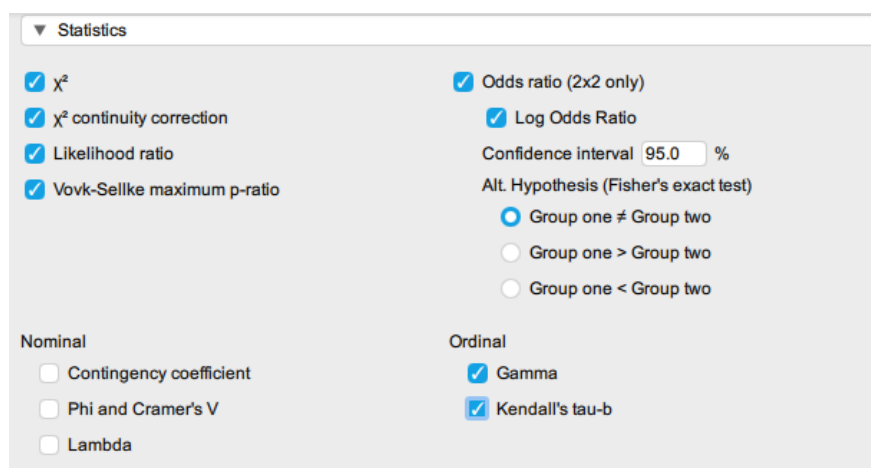
2. Откроем модуль «Freguenciens». Выберем из выпадающего списка «Contingency Tables». Введем переменные и параметры, как показано на рисунке:



3. Введем переменные и параметры, как показано на рисунке:



4. Зададим следующие настройки в разделе «Statistics»:



5. Зададим следующие параметры в меню в разделе «Cells»:

Cells

Counts

☐ Expected

Residuals

☐ Unstandardized
☐ Pearson
☐ Standardized (adjusted Pearson)

Percentages

☒ Row
☒ Column
☒ Total

6. Получим следующие результаты:

Contingency Tables ▼

Антибиотик		Рецидив		Total
		да	нет	
да	Count	6.000	12.000	18.000
	% within row	33.333 %	66.667 %	100.000 %
	% within column	24.000 %	63.158 %	40.909 %
	% of total	13.636 %	27.273 %	40.909 %
нет	Count	19.000	7.000	26.000
	% within row	73.077 %	26.923 %	100.000 %
	% within column	76.000 %	36.842 %	59.091 %
	% of total	43.182 %	15.909 %	59.091 %
Total	Count	25.000	19.000	44.000
	% within row	56.818 %	43.182 %	100.000 %
	% within column	100.000 %	100.000 %	100.000 %
	% of total	56.818 %	43.182 %	100.000 %

Будем помнить, что тесты с использованием хи-квadrat действительны только в том случае, если у нас разумный размер выборки, т. е. менее 20 % ячеек имеют ожидаемое количество менее 5 и ни одна из них не имеет ожидаемого количества менее 1:

Chi-Squared Tests

	Value	df	p	VS-MPR*
X ²	6.848	1	0.009	8.773
X ² continuity correction	5.324	1	0.021	4.528
Likelihood ratio	6.972	1	0.008	9.269
N	44			

* Vovk-Sellke Maximum p -Ratio: Based the p -value, the maximum possible odds in favor of H_1 over H_0 equals $1/(-e \cdot p \log(p))$ for $p \leq .37$ (Sellke, Bayarri, & Berger, 2001).

Статистика ($\chi^2 = 6,848$, $p < 0,05$) позволяет предположить, что существует значительная связь между приемом антибиотиков и отсутствием рецидива. χ^2 -поправку на непрерывность можно использовать для предотвращения завышения статистической значимости для малых наборов данных. В основном это используется, когда хотя бы одна ячейка таблицы имеет ожидаемое количество меньше 5.

JASP также предоставляет отношение шансов (OR), которое используется для сравнения относительных шансов наступления интересующего результата (в данном случае – рецидив) с учетом воздействия интересующей переменной (в данном случае – антибиотика):

Log Odds Ratio				
	Log Odds Ratio	95% Confidence Intervals		p
		Lower	Upper	
Odds ratio	-1.692	-3.000	-0.383	
Fisher's exact test	-1.648	-3.207	-0.217	0.014

В статистике коэффициент ранговой корреляции Кендалла, обычно называемый коэффициентом Кендалла τ (греческая буква «тау»), представляет собой статистику, используемую для измерения порядковой связи между двумя измеряемыми величинами. Тест τ – это непараметрический тест гипотезы на статистическую зависимость, основанный на коэффициенте τ . Это мера ранговой корреляции – сходство порядка данных при ранжировании по каждой из величин:

Ordinal Gamma ▼			
Gamma	Standard Error	95% Confidence Intervals	
		Lower	Upper
-0.689	0.175	-1.000	-0.345

Kendall's Tau			
Kendall's Tau-b	Z	p	VS-MPR*
-0.394	-2.587	0.010	8.191

Контрольные вопросы

1. Что такое таблицы сопряженности?
2. Как интерпретировать значения в таблице сопряженности?
3. Что такое хи-квадрат тест?
4. В каких случаях используется хи-квадрат тест?

2.9. Непараметрические критерии для анализа количественных признаков

Краткие сведения из теории

Для сравнения групп обычно применяются параметрические критерии, такие как критерий Фишера при дисперсионном анализе или критерий Стьюдента при сравнении двух групп. Эти критерии предполагают, что наблюдаемый признак в обеих группах распределен нормально. Кроме того, важно, чтобы дисперсии в обеих группах были приблизительно равны. Различия между средними значениями используются для оценки различий между совокупностями. Однако при использовании параметрических критериев необходимо быть уверенным, что условия их применения выполняются хотя бы приблизительно, чтобы избежать ошибочных выводов.

При несоблюдении требований для параметрических методов необходимо использовать непараметрические.

Непараметрические методы в биологической статистике используются для анализа данных, которые не соответствуют требованиям параметрических методов. Они не требуют предположений о распределении данных и могут быть применены к любому типу данных, включая категориальные и порядковые переменные. Непараметрические методы заменяют реальные значения признака рангами. Каждому значению признака в группе присваивается свой ранг в зависимости от величины значения признака, например:

Вариационный ряд	10	20	30	40	50	60	70	80
Ранги	1	2	3	4	5	6	7	8

Ранги – это порядковые числа, присвоенные каждому наблюдению в выборке. Например, если у нас есть выборка из 5 чисел: 1, 2, 3, 4, 5, то ранги будут 1, 2, 3, 4, 5.

Сравнение двух выборок: критерий Манна–Уитни (независимые группы)

Критерий Манна–Уитни (U) – это непараметрический метод для сравнения двух независимых выборок. Он используется, когда данные не соответствуют требованиям параметрических методов, например, когда данные не имеют нормального распределения или когда данные являются категориальными или порядковыми переменными.

Критерий Манна–Уитни основан на рангах данных, он сравнивает суммы рангов в каждой выборке. Если суммы рангов в обеих выборках примерно равны, то можно сделать вывод, что различия между средними значениями в выборках не значимы. Если же сумма рангов в одной выборке значительно больше, чем в другой, то можно сделать вывод, что различия между средними значениями в выборках значимы.

Критерий Манна–Уитни позволяет оценить значимость различий между средними значениями в выборках без предположений о распределении данных. Он также устойчив к выбросам и может быть применен к любому типу данных.

Однако критерий Манна–Уитни может быть менее мощным методом, чем параметрические методы, и может давать менее точные результаты при больших выборках. Поэтому выбор метода зависит от типа данных и задач, которые необходимо решить.

Порядок вычисления критерия Манна–Уитни (U):

1. Объединенные данные из обеих групп упорядочиваются по возрастанию. Наименьшему значению присваивается ранг 1, следующему – ранг 2 и т. д. Если значения совпадают, им присваивается средний ранг.
2. Для группы с меньшим количеством членов вычисляется сумма рангов ее членов. Если группы имеют одинаковую численность, сумма рангов может быть вычислена для любой из них.
3. Полученное значение сравнивается с критическими значениями. Если значение меньше или равно первому критическому значению или больше или равно второму критическому значению, нулевая гипотеза отвергается, что означает статистическую значимость различий.

Сравнение наблюдений «до» и «после»: критерий Уилкоксона (зависимые группы)

Критерий Уилкоксона (W) – это непараметрический метод для сравнения двух зависимых выборок. Он используется, когда данные не соответствуют требованиям параметрических методов, например, когда

данные не имеют нормального распределения или когда данные являются категориальными или порядковыми переменными.

Критерий Уилкоксона основан на разнице между рангами парных наблюдений. Рангам приписывают знак изменения и суммируют эти «знаковые ранги» – в результате получается значение критерия Уилкоксона.

Критерий Уилкоксона сравнивает сумму разностей рангов в каждой паре наблюдений. Если сумма разностей рангов в большинстве пар примерно равна нулю, то можно сделать вывод, что различия между средними значениями в выборках не значимы. Если же сумма разностей рангов в большинстве пар значительно отличается от нуля, то можно сделать вывод, что различия между средними значениями в выборках значимы.

Критерий Уилкоксона позволяет оценить значимость различий между средними значениями в выборках без предположений о распределении данных. Он также устойчив к выбросам и может быть применен к любому типу данных.

Однако критерий Уилкоксона может быть менее мощным методом, чем параметрические методы, и может давать менее точные результаты при больших выборках. Поэтому выбор метода зависит от типа данных и задач, которые необходимо решить.

Порядок вычисления критерия Уилкоксона (W):

1. Вычисляются величины изменений наблюдаемого признака для каждой пары наблюдений. Отбрасываются пары наблюдений, для которых изменение равно нулю.

2. Упорядочиваются изменения по возрастанию их абсолютной величины и им присваиваются соответствующие ранги. Рангами одинаковых величин назначаются средние тех мест, которые эти величины делят в упорядоченном ряду.

3. Присваивается каждому рангу знак в соответствии с направлением изменения: если значение увеличилось – «+», если уменьшилось – «-».

4. Вычисляется сумма знаковых рангов W .

5. Сравнивается полученная величина W с критическим значением. Если она больше критического значения, изменение показателя статистически значимо.

Сравнение нескольких групп: критерий Крускала–Уоллиса (независимые группы)

Критерий Крускала–Уоллиса (H) – это непараметрический метод для сравнения трех или более независимых выборок. Он используется, когда данные не соответствуют требованиям параметрических методов,

например, когда данные не имеют нормального распределения или когда данные являются категориальными или порядковыми переменными.

Данный критерий (аналог дисперсионного анализа) представляет собой обобщение критерия Манна–Уитни. Сначала все значения, независимо от того, какой выборке они принадлежат, упорядочивают по возрастанию. Каждому значению присваивается ранг – номер его места в упорядоченном ряду. Совпадающим значениям присваивают общий ранг, равный среднему тех мест, которые эти величины делят между собой в общем упорядоченном ряду. Затем вычисляют суммы рангов, относящихся к каждой группе, и для каждой группы определяют средний ранг. При отсутствии межгрупповых различий средние ранги групп должны оказаться близки. Напротив, если существует значительное расхождение средних рангов, то гипотезу об отсутствии межгрупповых различий следует отвергнуть. Значение критерия Крускала–Уоллиса и является мерой такого расхождения средних рангов.

Порядок вычисления критерия Крускала–Уоллиса (H):

1. Объединив все наблюдения, их упорядочивают по возрастанию. Совпадающим значениям ранги присваиваются как среднее тех мест, которые эти значения делят между собой.

2. Вычисляют критерий Крускала–Уоллиса.

3. Сравнивают вычисленное значение H с критическим значением χ^2 для числа степеней свободы, на единицу меньшего числа групп. Если вычисленное значение H окажется больше критического, различия групп статистически значимы.

Повторные измерения:

критерий Фридмана (зависимые группы)

Если, к примеру, одна и та же группа больных последовательно подвергается нескольким методам лечения или просто наблюдается в разные моменты времени, применяют дисперсионный анализ повторных измерений. Однако, чтобы использование дисперсионного анализа было правомерно, данные должны подчиняться нормальному распределению. Если в этом нет уверенности, лучше воспользоваться *критерием Фридмана* – непараметрическим аналогом дисперсионного анализа повторных измерений.

Этот критерий основан на рангах данных, он сравнивает сумму рангов в каждой выборке. Если сумма рангов в большинстве выборок примерно равна, то можно сделать вывод, что различия между средними значениями в выборках не значимы. Если же сумма рангов в одной выборке значительно больше, чем в других, то следует вывод, что различия между средними значениями в выборках значимы.

Критерий Фридмана позволяет оценить значимость различий между средними значениями в выборках без предположений о распределении данных. Он также устойчив к выбросам и может быть применен к любому типу данных.

Однако критерий Фридмана может быть менее мощным методом, чем параметрические методы, и может давать менее точные результаты при больших выборках. Поэтому выбор метода зависит от типа данных и задач, которые необходимо решить.

Порядок расчета критерия Фридмана:

1. Значения признака располагают по возрастанию, каждому значению присваивают ранг.
2. Для каждого из методов лечения (группы) подсчитывают сумму присвоенных ему рангов.
3. Вычисляют значение χ^2 .
4. Если (как пример) число методов лечения и число больных присутствует в таблице критических значений, определяют критическое значение χ^2 по этой таблице. Если число методов лечения и число больных достаточно велико (отсутствует в таблице), используют критическое значение χ^2 с числом степеней свободы $\nu = k - 1$.
5. Если рассчитанное значение χ^2 превышает критическое – различия статистически значимы.

Алгоритм анализа в JASP

*Сравнение двух выборок: критерий Манна–Уитни
(независимые группы)*

Задача

При заболеваниях сетчатки повышается проницаемость ее сосудов. Дж. Фишмен и соавторы измерили проницаемость сосудов

сетчатки у здоровых и у больных с ее поражением. Полученные результаты приведены в таблице:

Проницаемость сосудов сетчатки	
Здоровые	Больные
0,5	1,2
0,7	1,4
0,7	1,6
1,0	1,7
1,0	1,7
1,2	1,8
1,4	2,2
1,4	2,3
1,6	2,4
1,6	6,4
1,7	19,0
2,2	23,6

Необходимо проверить, подтверждают ли эти данные гипотезу о различии в проницаемости сосудов сетчатки.

1. Для начала проведем описательную статистику, чтобы определить, какое распределение имеем – нормальное либо ненормальное, а также чтобы оценить гомогенность дисперсий в двух группах. Используем для этой задачи сначала MS Excel.

2. Скопируем таблицу на лист в MS Excel и опишем данные по двум группам:

	А	В
1	Проницаемость сосудов сетчатки	
2	Здоровые	Больные
3	0,5	1,2
4	0,7	1,4
5	0,7	1,6
6	1	1,7
7	1	1,7
8	1,2	1,8
9	1,4	2,2
10	1,4	2,3
11	1,6	2,4
12	1,6	6,4
13	1,7	19
14	2,2	23,6

3. Затем опишем полученные данные, используя модуль «Описательная статистика». Полученные результаты указывают на ненормальность распределения (высокие значения асимметрии и эксцесса), кроме того, сильно различаются дисперсии:

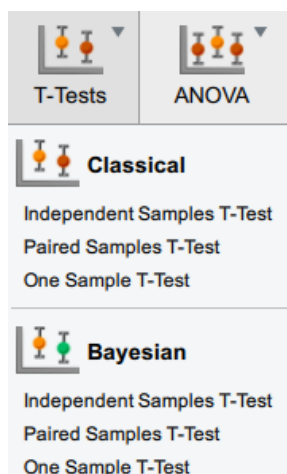
<i>Здоровые</i>		<i>Больные</i>	
Среднее	1,25	Среднее	5,441666667
Стандартная ошибка	0,142754292	Стандартная ошибка	2,192357882
Медиана	1,3	Медиана	2
Мода	0,7	Мода	1,7
Стандартное отклонение	0,494515373	Стандартное отклонение	7,594550478
Дисперсия выборки	0,244545455	Дисперсия выборки	57,67719697
Эксцесс	-0,360824201	Эксцесс	2,796218259
Асимметричность	0,230032579	Асимметричность	2,001156149
Интервал	1,7	Интервал	22,4
Минимум	0,5	Минимум	1,2
Максимум	2,2	Максимум	23,6
Сумма	15	Сумма	65,3
Счет	12	Счет	12

Далее используем критерий Шапиро–Уилка для определения нормальности, а для оценки гомогенности дисперсий – критерий Брауна–Форсайта. Так как выборки независимые, можно использовать непараметрический критерий Манна–Уитни.

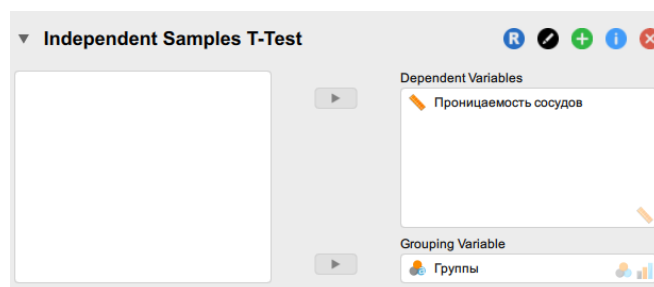
4. Откроем исходные данные в JASP. Но предварительно изменим их представление в таблице MS Excel, а именно добавим предиктор для каждой группы:

▲	А	В	Группы	Проницаемость сосудов	
1	Группы	Данные	1	a	0.5
2	a	0.5	2	a	0.7
3	a	0.7	3	a	0.7
4	a	0.7	4	a	1
5	a	1	5	a	1
6	a	1	6	a	1.2
7	a	1.2	7	a	1.4
8	a	1.4	8	a	1.4
9	a	1.4	9	a	1.6
10	a	1.6	10	a	1.6
11	a	1.6	11	a	1.7
12	a	1.7	12	a	2.2
13	a	2.2	13	b	1.2
14	b	1.2	14	b	1.4
15	b	1.4	15	b	1.6
16	b	1.6	16	b	1.7
17	b	1.7	17	b	1.7
18	b	1.7	18	b	1.8
			19	b	2.2
			20	b	2.3

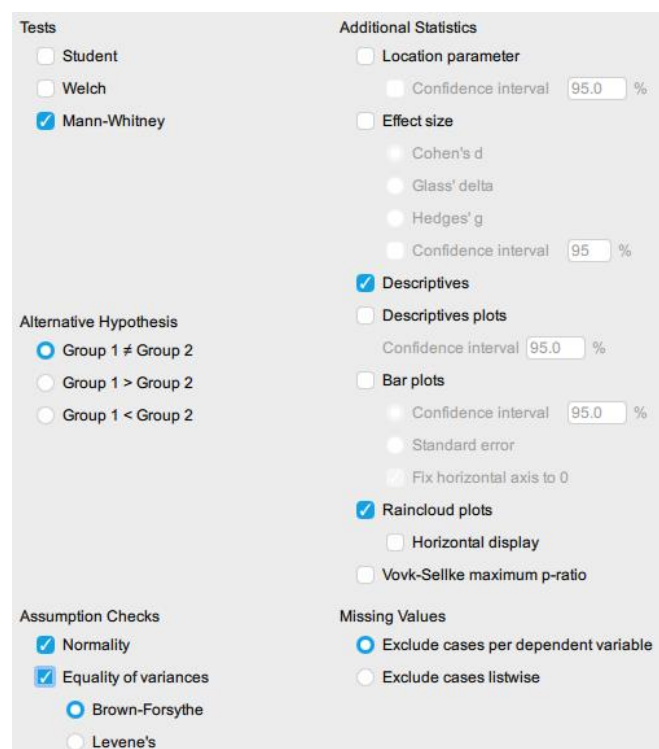
5. Перейдем во вкладку «Analyses» и выберем из списка модуля «T-Tests» раздел «Independent Samples T-Test»:



6. Перенесем переменные, как показано на рисунке:



7. Выберем следующие настройки:



8. Получим следующие результаты:

Independent Samples T-Test ▼

Independent Samples T-Test

	W	df	p
Проницаемость сосудов	19.000		0.002

Note. Mann-Whitney U test.

Assumption Checks ▼

Test of Normality (Shapiro-Wilk)

	W	p
Проницаемость сосудов a	0.967	0.874
b	0.594	< .001

Note. Significant results suggest a deviation from normality.

Test of Equality of Variances (Brown-Forsythe) ▼

	F	df ₁	df ₂	p
Проницаемость сосудов	2.669	1	22	0.117

Группы между собой статистически достоверно различаются ($p < 0,05$). Так как распределение ненормальное (тест Шапиро–Уилка, $p < 0,05$), то используется непараметрические критерий Манна–Уитни. Тест Брауна–Форсайта показывает, что дисперсии между группами равны ($p > 0,05$).

Описательную статистику по группам можно посмотреть на рисунке:

Descriptives

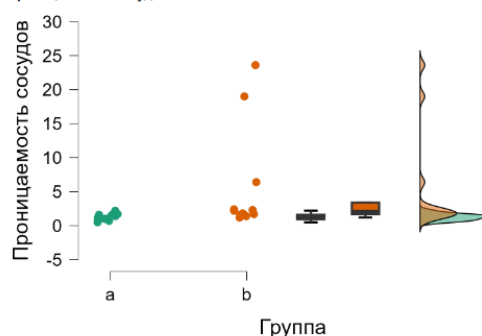
Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
Проницаемость сосудов	a	12	1.250	0.495	0.143	0.396
	b	12	5.442	7.595	2.192	1.396

А кроме того, можно визуализировать полученные данные:

Raincloud Plots ▼

Проницаемость сосудов



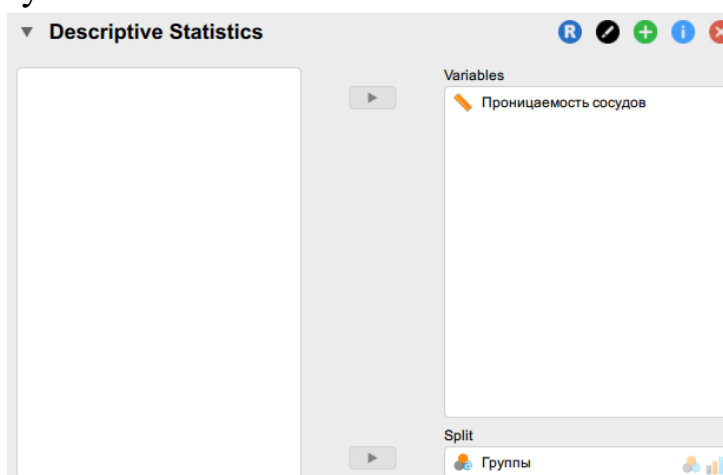
9. Заметим, что если мы проведем параметрический тест t-Стьюдента, то не получим статистически значимых различий между группами, что неверно:

Independent Samples T-Test

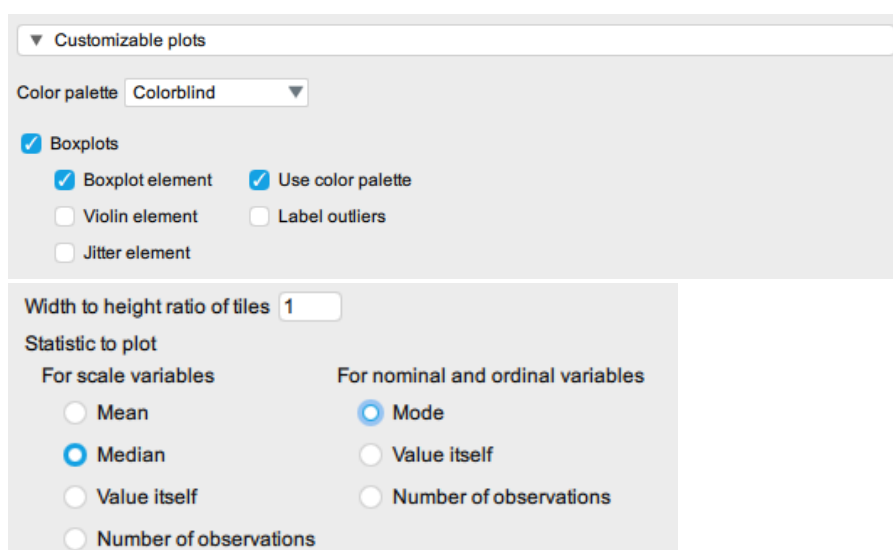
Independent Samples T-Test			
	t	df	p
Проницаемость сосудов	-1.908	22	0.070

Note. Student's t-test.

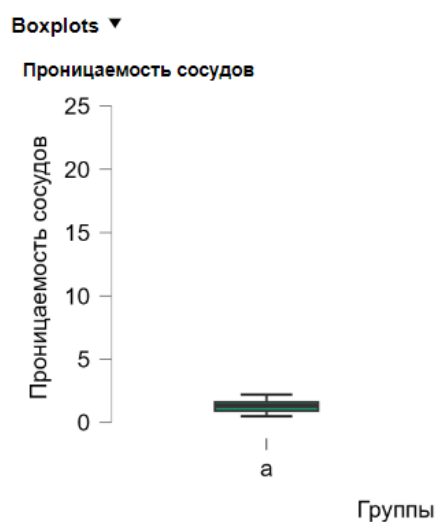
10. Теперь необходимо построить отдельно график «Boxplot». Для этого перейдем в модуль «Descriptives», перенесем переменные, как показано на рисунке:



11. Перейдем в раздел «Customizable plots» и установим следующие настройки:



12. Получим следующий график:



13. Теперь постоим «Boxplot» в MS Excel. Для этого обведем таблицу, затем перейдем во вкладку «Вставка», выберем «Гистограмма», далее «Ящик с усами»:

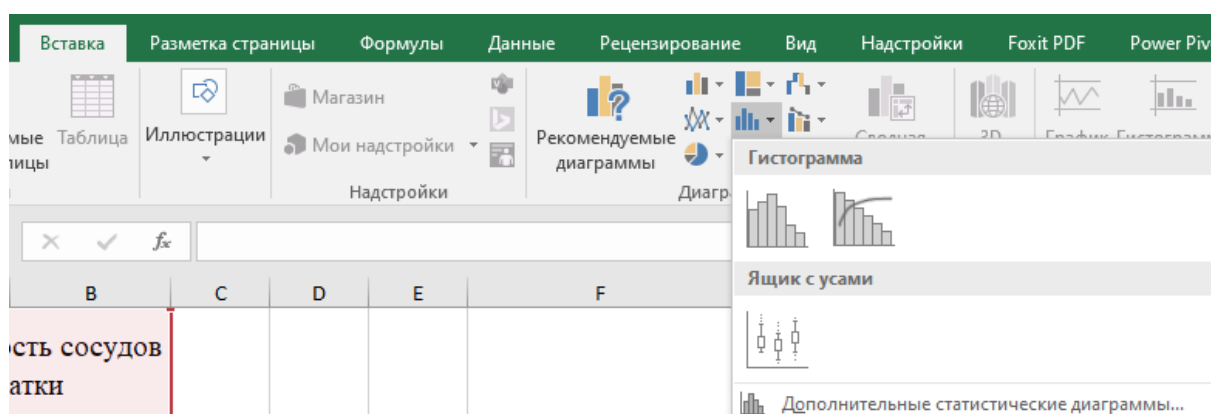
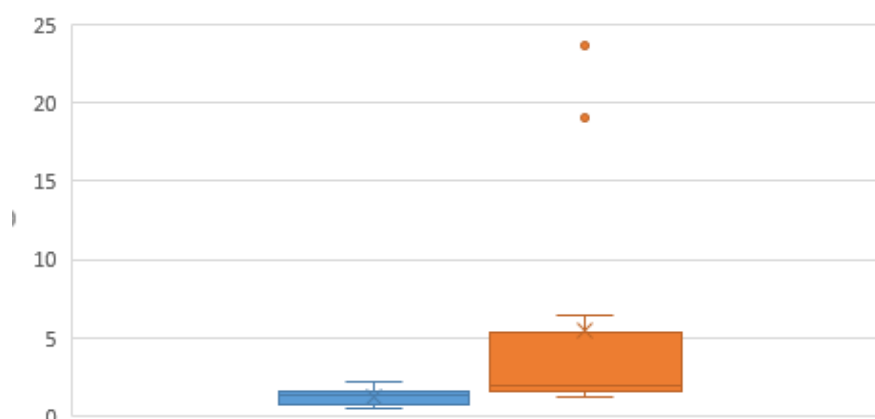


График «Boxplot»:



*Сравнение наблюдений «до» и «после»:
критерий Уилкоксона (зависимые группы)*

Задача

Исследователи провели испытание нового препарата, влияющего на агрегацию тромбоцитов. Были получены следующие данные:

№ участника	Агрегация тромбоцитов, %	
	до приема препарата, (контрольная группа)	после приема препарата, (экспериментальная группа)
1	26	23
2	16	30
3	28	38
4	45	57
5	11	27
6	68	83
7	54	58
8	54	81
9	53	110
10	98	90
11	19	24

Необходимо определить, влияет ли данный препарат на агрегацию тромбоцитов.

1. Для начала проведем описательную статистику, чтобы определить, какое распределение имеем – нормальное либо ненормальное, а также чтобы оценить гомогенность дисперсий в двух группах. Используем для этой задачи сначала MS Excel.

2. Скопируем таблицу на лист в MS Excel и опишем данные по двум группам:

	А	В	С
1	№ участника	До приема препарата, % (контрольная группа)	После приема препарата, % (экспериментальная группа)
2	1	26	23
3	2	16	30
4	3	28	38
5	4	45	57
6	5	11	27
7	6	68	83
8	7	54	58
9	8	54	81
10	9	53	110
11	10	98	90
12	11	19	24

3. Затем опишем полученные данные, используя модуль «Описательная статистика». Полученные результаты указывают на ненормальность распределения (высокие значения асимметрии и эксцесса), а кроме того, сильно различаются дисперсии:

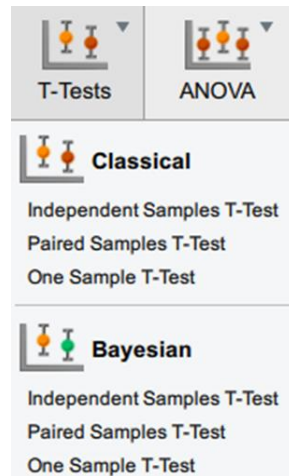
До приема препарата, % (контрольная группа)		После приема препарата, % (экспериментальная группа)	
Среднее	42,90909	Среднее	56,45454545
Стандартная ошибка	7,884916	Стандартная ошибка	9,235665686
Медиана	45	Медиана	57
Мода	54	Мода	#Н/Д
Стандартное отклонение	26,15131	Стандартное отклонение	30,63123777
Дисперсия выборки	683,8909	Дисперсия выборки	938,2727273
Эксцесс	0,406694	Эксцесс	-1,237290982
Асимметричность	0,772043	Асимметричность	0,437115452
Интервал	87	Интервал	87
Минимум	11	Минимум	23
Максимум	98	Максимум	110
Сумма	472	Сумма	621
Счет	11	Счет	11

Далее используем критерий Шапиро–Уилка для определения нормальности, а для оценки гомогенности дисперсий – критерий Брауна–Форсайта. Так как выборки зависимые, можно использовать непараметрический критерий Уилкоксона.

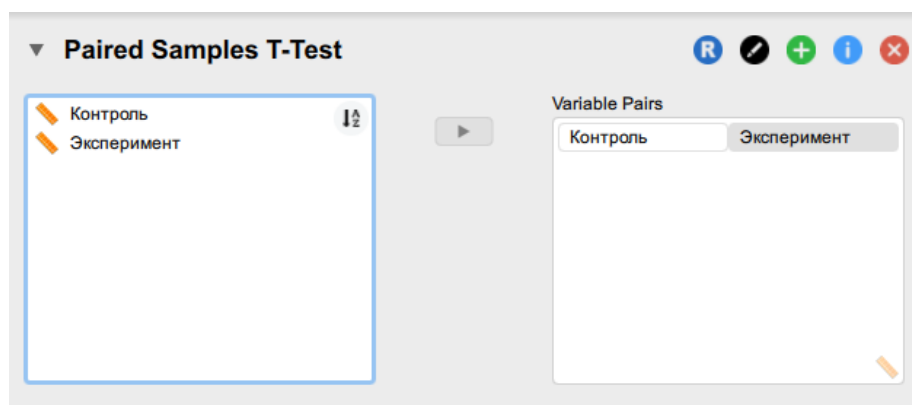
4. Откроем исходные данные в JASP. В данном случае нам не понадобится устанавливать предиктор для каждой группы:

	 Контроль	 Эксперимент
1	26	23
2	16	30
3	28	38
4	45	57
5	11	27
6	68	83
7	54	58
8	54	81
9	53	110
10	98	90
11	19	24

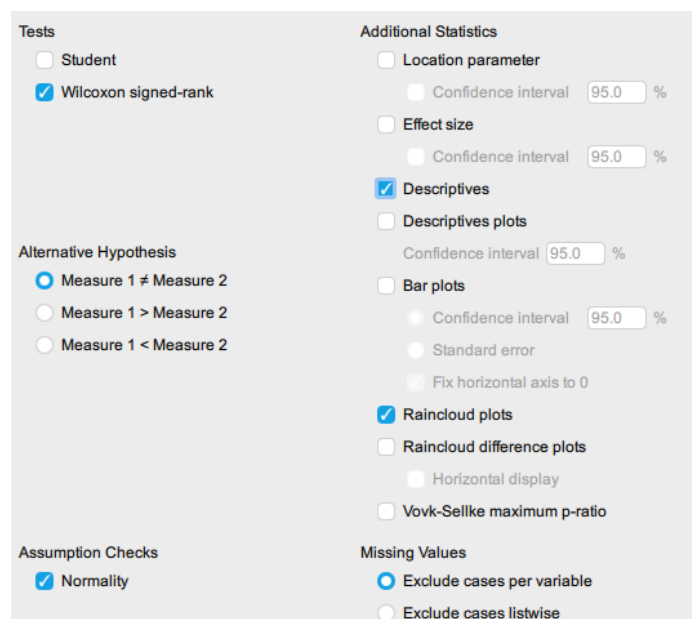
5. Перейдем во вкладку «Analyses» и выберем из списка модуля «T-Tests» раздел «Paired Samples T-Test»:



6. Перенесем переменные, как показано на рисунке:



7. Выберем следующие настройки:



8. Получим следующие результаты:

Results ▼

Paired Samples T-Test ▼

Paired Samples T-Test

Measure 1	Measure 2	W	z	df	p
Контроль	- Эксперимент	5.000	-2.490		0.010

Note. Wilcoxon signed-rank test.

Assumption Checks ▼

Test of Normality (Shapiro-Wilk) ▼

	W	p
Контроль - Эксперимент	0.861	0.059

Note. Significant results suggest a deviation from normality.

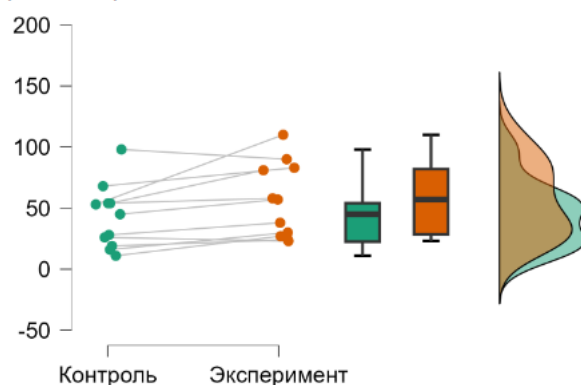
Descriptives

Descriptives

	N	Mean	SD	SE	Coefficient of variation
Контроль	11	42.909	26.151	7.885	0.609
Эксперимент	11	56.455	30.631	9.236	0.543

Raincloud Plots ▼

Контроль - Эксперимент ▼



9. Стоит обратить внимание, что проведение теста t-Стьюдента здесь тоже дает статистически значимый результат, но с большей долей вероятности ошибки по сравнению с тестом Уилкоксона. Хотя тест Шапиро–Уилка и пройден, но значение p находится на грани нормального значения, поэтому целесообразно в качестве теста использовать непараметрический метод оценки статистической значимости:

Paired Samples T-Test

Paired Samples T-Test

Measure 1	Measure 2	t	df	p
Контроль	- Эксперимент	-2.596	10	0.027

Note. Student's t-test.

10. Теперь необходимо построить отдельно график «Boxplot». Для этого перейдем в модуль «Descriptives», перенесем переменные, как было показано ранее. Но предварительно необходимо им дать группирующий предиктор ($\alpha \dots \alpha_k$, $b \dots b_k$), чтобы отразить их на одном графике:

	Группа	Контроль
1	a	26
2	a	16
3	a	28
4	a	45
5	a	11
6	a	68
7	a	54
8	a	54
9	a	53
10	a	98
11	a	19
12	b	23
13	b	30
14	b	38

11. Перейдем в раздел «Customizable plots» и установим следующие настройки:

Customizable plots

Color palette
Colorblind

☒ Boxplots

☒ Boxplot element
☒ Use color palette
☐ Violin element
☐ Label outliers
☐ Jitter element

Width to height ratio of tiles
1

Statistic to plot

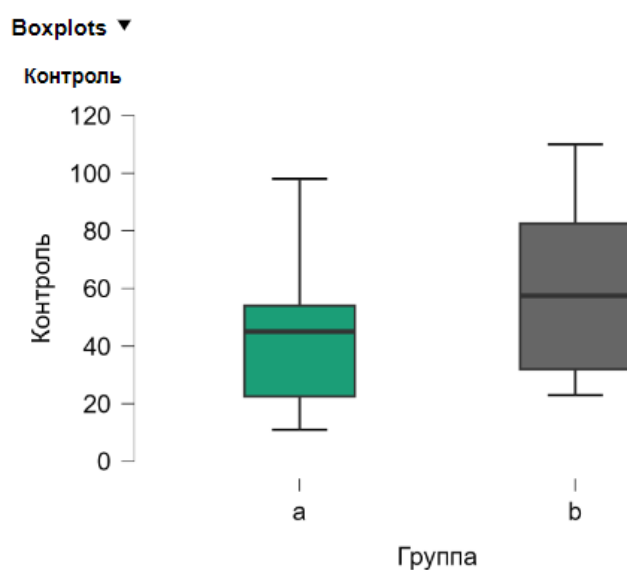
For scale variables

☐ Mean
☒ Median
☐ Value itself
☐ Number of observations

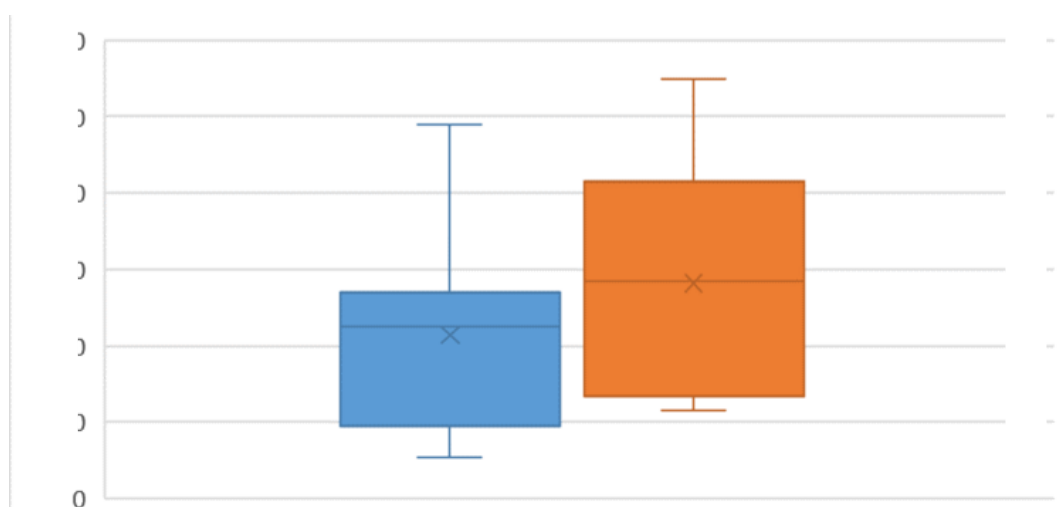
For nominal and ordinal variables

☒ Mode
☐ Value itself
☐ Number of observations

12. Получим следующий график:



13. Теперь постоим «Boxplot» в MS Excel. Для этого обведем таблицу, затем перейдем во вкладку «Вставка», выберем «Гистограмма», далее «Ящик с усами». График «Boxplot»:



*Сравнение нескольких групп: критерий Крускала – Уоллиса
(независимые группы)*

Задача

Кофеин содержится в ряде лекарственных средств и пищевых продуктов, таких как кофе, чай и прохладительные напитки. Беременным женщинам не рекомендуется употреблять крепкий кофе, так как кофеин может негативно повлиять на плод, а его выведение у беременных замедлено. Предполагается, что замедленное выведение

кофеина связано с высоким уровнем половых гормонов во время беременности. Р. Патвардан и соавторы решили проверить это предположение, измерив скорость выведения кофеина. Они определили, что скорость выведения кофеина прямо пропорциональна его концентрации в плазме. Вместо измерения скорости выведения в миллиграммах в минуту, они использовали период полувыведения ($T_{1/2}$) – время, за которое концентрация вещества уменьшается вдвое. Они измерили $T_{1/2}$ у женщин, принимающих и не принимающих пероральные контрацептивы (при приеме пероральных контрацептивов уровень эстрогенов и прогестагенов в крови повышается – то же самое происходит и при беременности), а также у мужчин. Каждый участник эксперимента принимал 250 мг кофеина, что соответствует примерно 3 чашкам кофе, после чего дважды определяли концентрацию кофеина в крови и рассчитывали $T_{1/2}$. Результаты представлены в таблице:

Период полувыведения $T_{1/2}$, ч		
Мужчины	Женщины	
	не принимающие пероральных контрацептивов	принимающие пероральные контрацептивы
2,04	5,30	10,36
5,16	7,28	13,28
6,11	8,98	11,81
5,82	6,59	4,54
5,41	4,59	11,04
3,51	5,17	10,08
3,18	7,25	14,47
4,57	3,47	9,43
4,83	7,6	13,41
11,34	–	–
3,79	–	–
9,03	–	–
7,21	–	–

Необходимо провести статистический анализ с помощью критерия Крускала–Уоллиса.

1. Для начала проведем описательную статистику, чтобы определить, какое распределение имеем – нормальное либо ненормальное, а также чтобы оценить гомогенность дисперсий в двух группах. Используем для этой задачи сначала MS Excel.

2. Скопируем таблицу на лист в MS Excel и опишем данные по трем группам:

	Мужчины	Не принимающие пероральные контрацептивов	Принимающие пероральные контрацептивы
2			
5	$T_{1/2}$, ч	$T_{1/2}$, ч	$T_{1/2}$, ч
6	2,04	5,3	10,36
7	5,16	7,28	13,28
8	6,11	8,98	11,81
9	5,82	6,59	4,54
10	5,41	4,59	11,04
11	3,51	5,17	10,08
12	3,18	7,25	14,47
13	4,57	3,47	9,43
14	4,83	7,6	13,41
15	11,34		
16	3,79		
17	9,03		
18	7,21		

3. Затем опишем полученные данные, используя модуль «Описательная статистика». Полученные результаты указывают на ненормальность распределения (высокие значения асимметрии и эксцесса), а кроме того, различаются дисперсии:

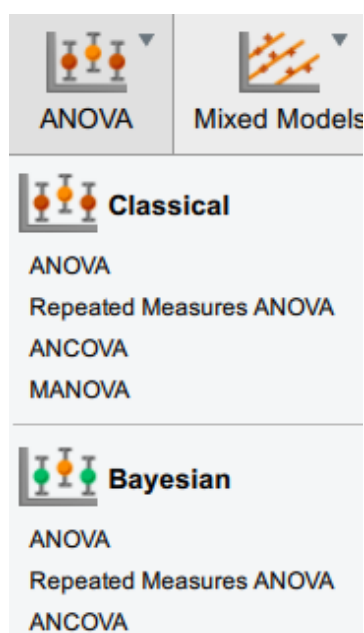
F	G	H	I	J	K
Среднее	5,83	Среднее	6,36625	Среднее	11,0075
Стандартная ошибка	0,687344	Стандартная ошибка	0,640789	Стандартная ошибка	1,107986833
Медиана	5,285	Медиана	6,92	Медиана	11,425
Мода	#Н/Д	Мода	#Н/Д	Мода	#Н/Д
Стандартное отклонение	2,381031	Стандартное отклонение	1,812425	Стандартное отклонение	3,133860011
Дисперсия выборки	5,669309	Дисперсия выборки	3,284884	Дисперсия выборки	9,821078571
Эксцесс	1,504599	Эксцесс	-0,68473	Эксцесс	2,069552287
Асимметричность	1,290947	Асимметричность	-0,33311	Асимметричность	-1,284894357
Интервал	8,16	Интервал	5,51	Интервал	9,93
Минимум	3,18	Минимум	3,47	Минимум	4,54
Максимум	11,34	Максимум	8,98	Максимум	14,47
Сумма	69,96	Сумма	50,93	Сумма	88,06
Счет	12	Счет	8	Счет	8

Далее используем критерий Шапиро–Уилка для определения нормальности, а для оценки гомогенности дисперсий – критерий Брауна–Форсайта. Так как выборки независимые, можно использовать непараметрический критерий Крускала–Уоллиса.

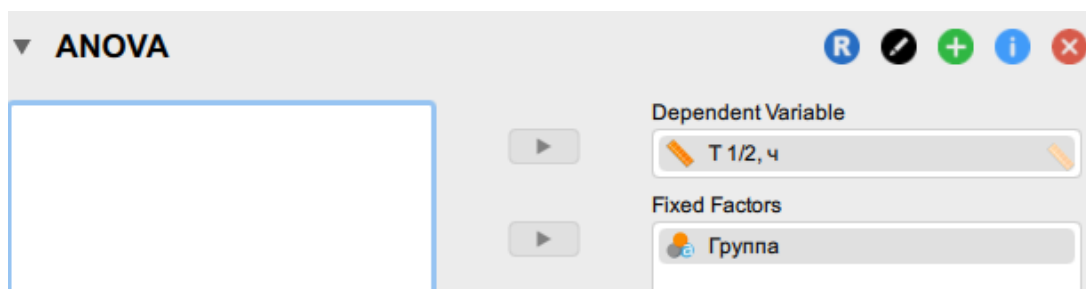
4. Откроем исходные данные в JASP. В данном случае нам необходимо устанавливать предиктор для каждой группы:

	A	B		Группа	T 1/2, ч
1	a	2,04	4	a	5.82
2	a	5,16	5	a	5.41
3	a	6,11	6	a	3.51
4	a	5,82	7	a	3.18
5	a	5,41	8	a	4.57
6	a	3,51	9	a	4.83
7	a	3,18	10	a	11.34
8	a	4,57	11	a	3.79
9	a	4,83	12	a	9.03
10	a	11,34	13	a	7.21
11	a	3,79	14	б	5.3
12	a	9,03	15	б	7.28
13	a	7,21	16	б	8.98
14	б	5,3	17	б	6.59
15	б	7,28	18	б	4.59
16	б	8,98	19	б	5.17
17	б	6,59	20	б	7.25
18	б	4,59	21	б	3.47
			22	б	7.6
			23	в	10.36

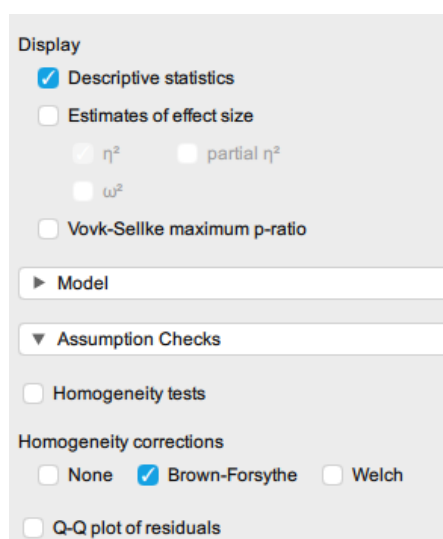
5. Перейдем во вкладку «Analyses» и выберем из списка модуля «ANOVA» раздел «ANOVA»:



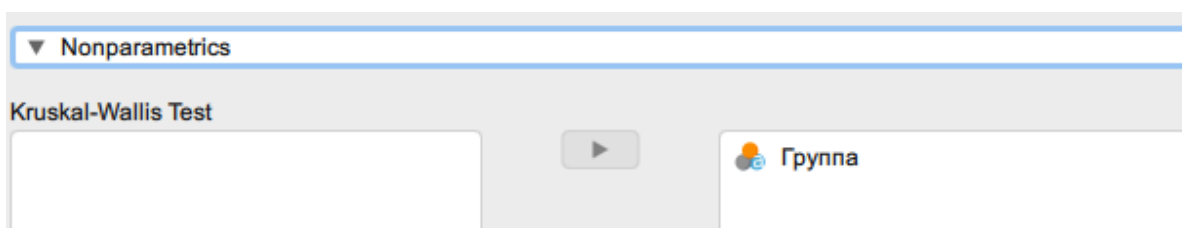
6. Перенесем переменные, как показано на рисунке:



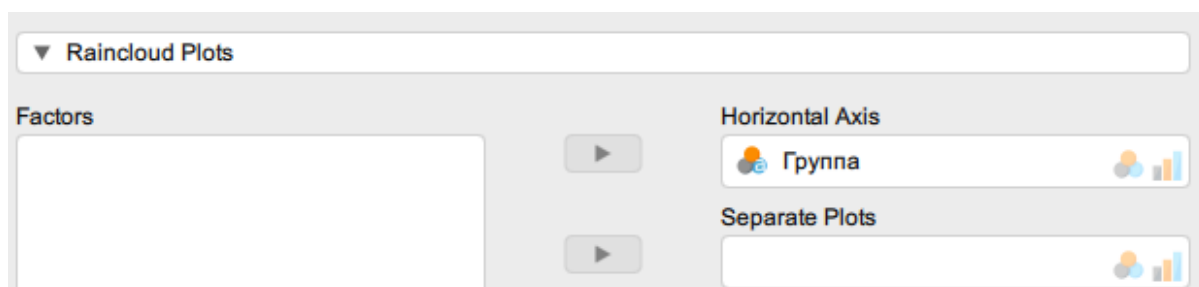
7. Выберем следующие настройки:



8. Проводим непараметрический тест Крускала–Уоллиса:



9. Построим график «Boxplot»:



10. Получим следующие результаты:

Results ▾

ANOVA ▾

ANOVA - T 1/2, ч

Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p
Brown-Forsythe	Группа	169.255	2.000	84.628	14.199	< .001
	Residuals	168.747	22.264	7.579		

Note. Type III Sum of Squares

Descriptives ▾

Descriptives - T 1/2, ч ▾

Группа	N	Mean	SD	SE	Coefficient of variation
b	9	6.248	1.732	0.577	0.277
c	9	10.936	2.939	0.980	0.269
a	13	5.538	2.510	0.696	0.453

Kruskal-Wallis Test

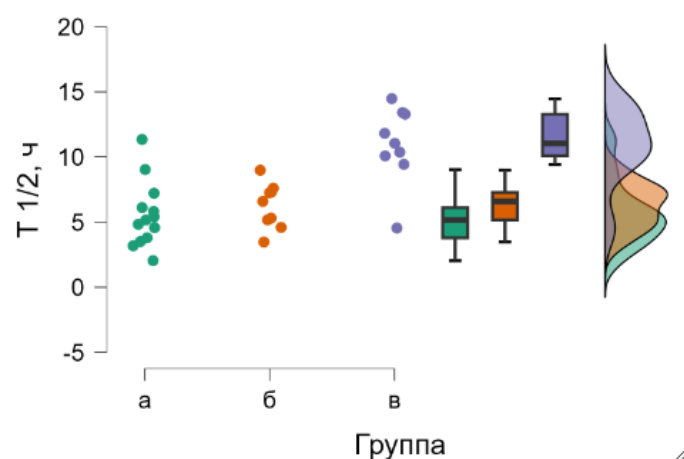
Kruskal-Wallis Test

Factor	Statistic	df	p
Группа	12.098	2	0.002

Descriptives ▾

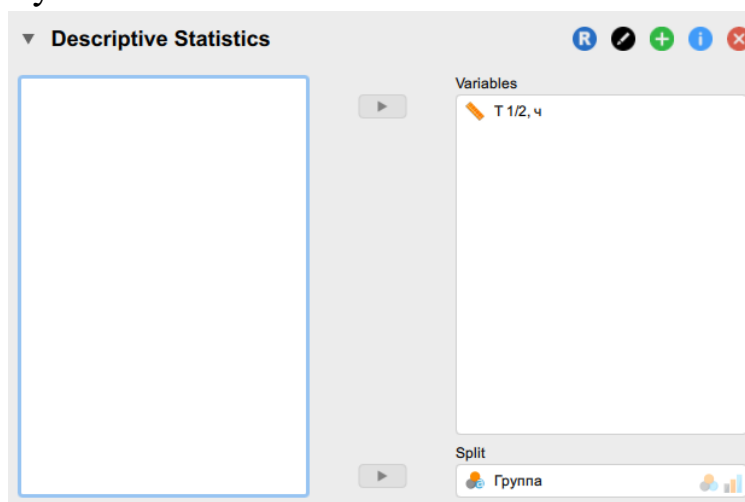
Raincloud plots ▾

T 1/2, ч ▾

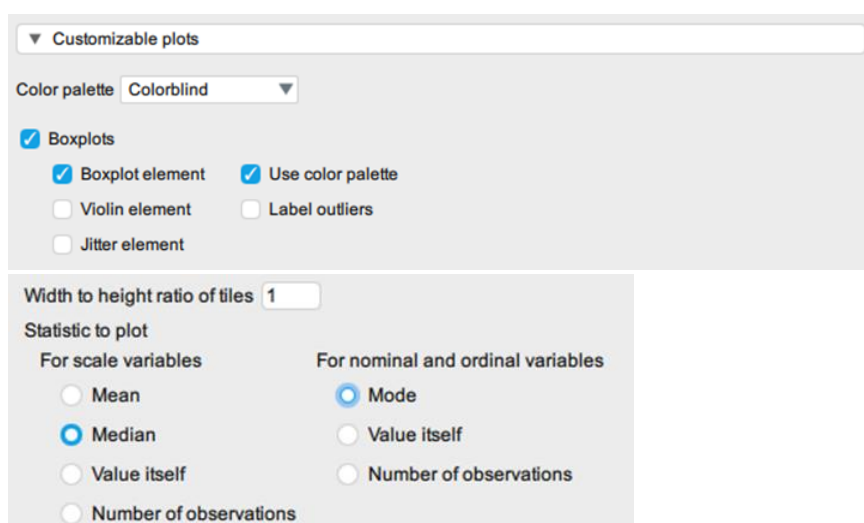


11. Стоит обратить внимание, что проведение теста F-Фишера здесь тоже дает статистически значимый результат. Однако, так как распределение ненормальное, необходимо использовать непараметрический критерий для независимых выборок.

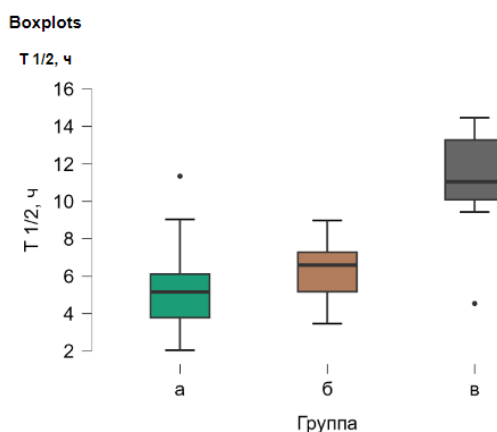
12. Теперь необходимо построить отдельно график «Boxplot». Для этого перейдем в модуль «Descriptives», перенесем переменные, как показано на рисунке:



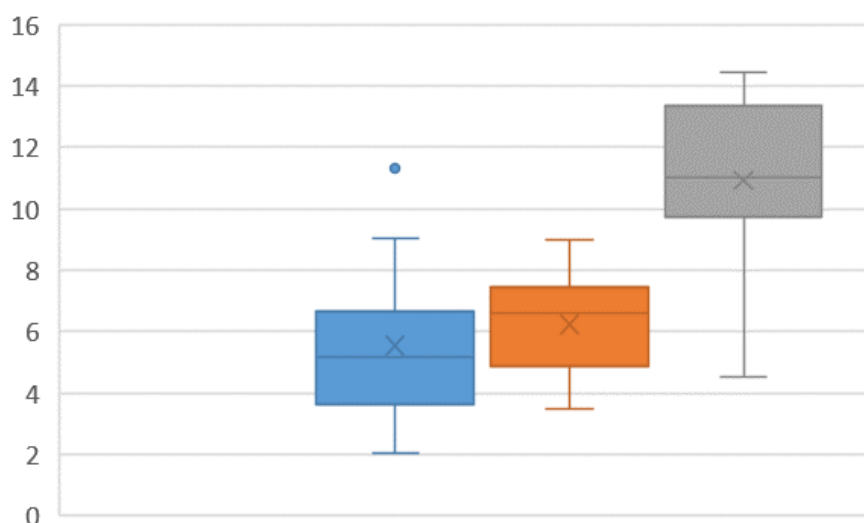
13. Перейдем в раздел «Customizable plots» и установим следующие настройки:



14. Получим следующий график:



15. Теперь постоим «Boxplot» в MS Excel. Для этого обведем таблицу, затем перейдем во вкладку «Вставка», выберем «Гистограмма», далее «Ящик с усами». График «Boxplot»:



*Повторные измерения: критерий Фридмана
(зависимые группы)*

Задача

Исследователи хотят сравнить эффективность нового препарата на легочное сосудистое сопротивление у больных, страдающих легочной гипертензией. Лёгочная гипертензия (ЛГ) – это группа заболеваний, характеризующихся прогрессивным повышением лёгочного сосудистого сопротивления, что ведёт к правожелудочковой недостаточности и преждевременной смерти. Получены следующие результаты:

Номер больного	Легочное сосудистое сопротивление, относит. ед.		
	До лечения	Спустя 7 дней	Спустя 1 месяц
1	22,1	5,4	10,2
2	18,3	10,9	4,5
3	15,4	7,2	2,8
4	14,2	10,1	11,3
5	26,5	4,2	5,3
6	12,8	9,8	10,3
7	23,1	3,2	4,4
8	19,7	13,2	2,9
9	14,9	5,5	3,4
10	27,6	6,4	5,1

Необходимо провести статистический анализ с помощью критерия Фридмана.

1. Для начала проведем описательную статистику, чтобы определить, какое распределение имеем – нормальное либо ненормальное, а также чтобы оценить гомогенность дисперсий в двух группах. Используем для этой задачи сначала MS Excel.

2. Скопируем таблицу на лист в MS Excel и опишем данные по трем группам:

	A	B	C	D
1	Легочное сосудистое сопротивление			
2	№ больного	До лечения	Спустя 7 дней	Спустя 1 месяц
3	1	22,1	5,4	10,2
4	2	18,3	10,9	4,5
5	3	15,4	7,2	2,8
6	4	14,2	10,1	11,3
7	5	26,5	4,2	5,3
8	6	12,8	9,8	10,3
9	7	23,1	3,2	4,4
10	8	19,7	13,2	2,9
11	9	14,9	5,5	3,4
12	10	27,6	6,4	5,1

3. Затем опишем полученные данные, используя модуль «Описательная статистика». Полученные результаты указывают на ненормальность распределения (высокие значения асимметрии и эксцесса), а кроме того, сильно различаются дисперсии:

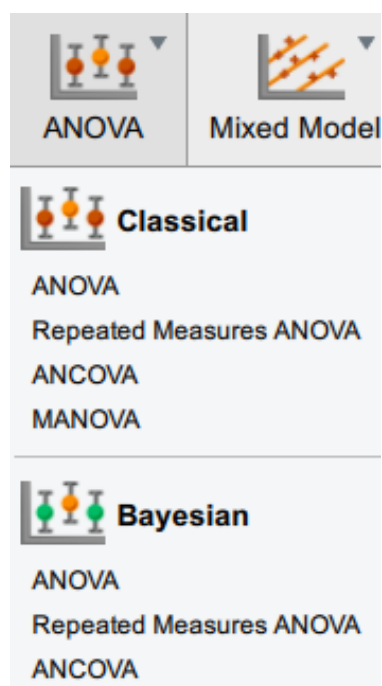
До лечения		Спустя 7 дней		Спустя 1 месяц	
Среднее	19,46	Среднее	7,59	Среднее	6,02
Стандартная ошибка	1,655778	Стандартная ошибка	1,027992	Стандартная ошибка	1,037818224
Медиана	19	Медиана	6,8	Медиана	4,8
Мода	#Н/Д	Мода	#Н/Д	Мода	#Н/Д
Стандартное отклонение	5,236029	Стандартное отклонение	3,250795	Стандартное отклонение	3,281869386
Дисперсия выборки	27,416	Дисперсия выборки	10,56767	Дисперсия выборки	10,77066667
Эксцесс	-1,28759	Эксцесс	-0,97453	Эксцесс	-1,177347632
Асимметричность	0,339277	Асимметричность	0,38202	Асимметричность	0,817185255
Интервал	14,8	Интервал	10	Интервал	8,5
Минимум	12,8	Минимум	3,2	Минимум	2,8
Максимум	27,6	Максимум	13,2	Максимум	11,3
Сумма	194,6	Сумма	75,9	Сумма	60,2
Счет	10	Счет	10	Счет	10

Далее используем критерий Шапиро–Уилка для определения нормальности, а для оценки гомогенности дисперсий – критерий Брауна–Форсайта. Так как выборки зависимые, можно использовать непараметрический критерий Фридмана.

4. Откроем исходные данные в JASP. В данном случае нам не понадобится устанавливать предиктор для каждой группы:

Т	До лечения	Спустя 7 дней	Спустя 1 месяц
1	22.1	5.4	10.2
2	18.3	10.9	4.5
3	15.4	7.2	2.8
4	14.2	10.1	11.3
5	26.5	4.2	5.3
6	12.8	9.8	10.3
7	23.1	3.2	4.4
8	19.7	13.2	2.9
9	14.9	5.5	3.4
10	27.6	6.4	5.1

5. Перейдем во вкладку «Analyses» и выберем из списка модуля «ANOVA» раздел «Repeated Measures ANOVA»:



6. Перенесем переменные, как показано на рисунке, и добавим уровень (level) 3.

Repeated Measures Factors

RM Factor 1

Level 1

Level 2

Level 3

New Level

New Factor

Repeated Measures Cells

До лечения	Level 1
Спустя 1 месяц	Level 2
Спустя 7 дней	Level 3

7. Выберем следующие настройки:

Display

☒ Descriptive statistics

☐ Estimates of effect size

☒ η^2 ☐ partial η^2 ☐ general η^2

☐ ω^2

☐ Vovk-Sellke maximum p-ratio

▼ Nonparametrics

Factors

►

RM Factor

RM Factor 1

Optional Grouping Factor

►

☒ Conover's post hoc tests

▼ Raincloud Plots

Factors

►

Horizontal Axis

RM Factor 1

Separate Plots

►

Label y-axis

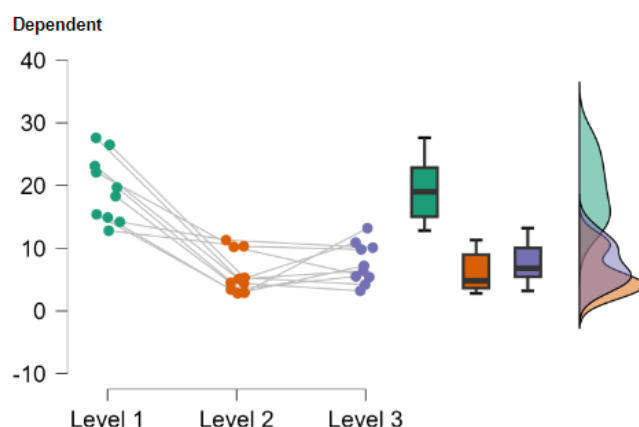
8. Получим следующие результаты:

Descriptives ▼

Descriptives

RM Factor 1	N	Mean	SD	SE	Coefficient of variation
Level 1	10	19.460	5.236	1.656	0.269
Level 2	10	6.020	3.282	1.038	0.545
Level 3	10	7.590	3.251	1.028	0.428

Raincloud plots



Nonparametrics ▼

Friedman Test ▼

Factor	Chi-Squared	df	p	Kendall's W
RM Factor 1	15.000	2	< .001	0.750

Conover Test

Conover's Post Hoc Comparisons - RM Factor 1

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
Level 1	Level 2	3.354	18	30.000	15.000	0.004	0.011	0.011
	Level 3	3.354	18	30.000	15.000	0.004	0.011	0.011
Level 2	Level 3	0.000	18	15.000	15.000	1.000	1.000	1.000

Note. Grouped by subject.

9. Стоит обратить внимание, что проведение теста F-Фишера здесь тоже дает статистически значимый результат. Однако необходимо в качестве теста использовать непараметрический метод оценки статистической значимости, так как распределение у выборок ненормальное:

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
RM Factor 1	1079.985	2	539.992	26.867	< .001
Residuals	361.775	18	20.099		

Note. Type III Sum of Squares

10. Теперь необходимо построить отдельно график «Boxplot». Для этого перейдем в модуль «Descriptives», перенесем переменные, как показано на рисунке. Но предварительно необходимо им дать группирующий предиктор, чтобы отразить их на одном графике:

	Группы	Данные
1	a	22.1
2	a	18.3
3	a	15.4
4	a	14.2
5	a	26.5
6	a	12.8
7	a	23.1
8	a	19.7
9	a	14.9
10	a	27.6
11	b	5.4
12	b	10.9
13	b	7.2
14	b	10.1
15	b	4.2
16	b	9.8
17	b	3.2
18	b	13.2
19	b	5.5
20	b	6.4

Descriptive Statistics

Variables: Данные

Split: Группы

11. Перейдем в раздел «Customizable plots» и установим следующие настройки:

Customizable plots

Color palette: Colorblind

☒ Boxplots

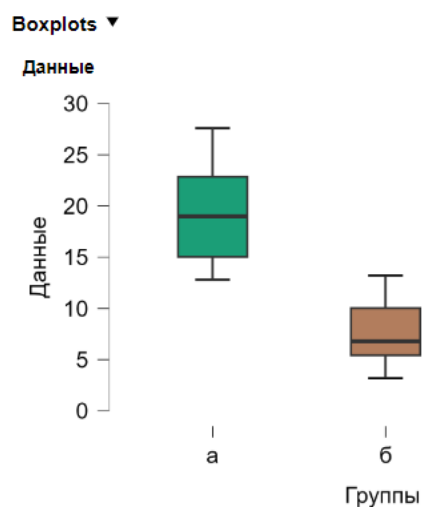
☒ Boxplot element ☒ Use color palette
☐ Violin element ☐ Label outliers
☐ Jitter element

Width to height ratio of tiles: 1

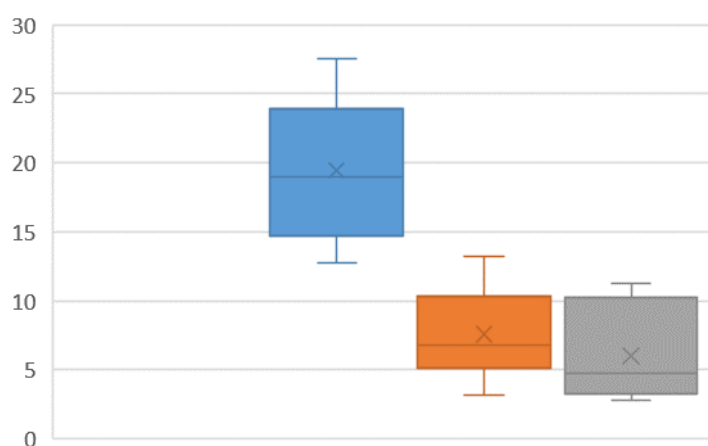
Statistic to plot

For scale variables	For nominal and ordinal variables
☐ Mean	☒ Mode
☒ Median	☐ Value itself
☐ Value itself	☐ Number of observations
☐ Number of observations	

12. Получим следующий график:



13. Теперь постоим «Boxplot» в MS Excel. Для этого обведем таблицу, затем перейдем во вкладку «Вставка», выберем «Гистограмма», далее «Ящик с усами». График «Boxplot»:



Контрольные вопросы

1. Какие статистические методы относятся к непараметрическим методам?
2. В каких случаях используются непараметрические тесты?
3. Чем отличаются параметрические тесты от непараметрических?
4. Каковы преимущества непараметрических тестов перед параметрическими?
5. Каковы недостатки непараметрических тестов?
6. Какой непараметрический метод используется для сравнения двух независимых выборок?

7. Какой непараметрический метод используется для сравнения двух связанных (зависимых) выборок?
8. Какой непараметрический метод используется для проверки нормальности распределения?
9. Какой непараметрический метод используется для определения связи между двумя переменными?
10. Какой непараметрический метод используется для проверки гипотез о медиане двух выборок?
11. Какой непараметрический метод используется для проверки однородности двух выборок?
12. Как называется непараметрический аналог t-теста?
13. Как называется непараметрический аналог F-теста?
14. Как называется непараметрический аналог дисперсионного анализа?
15. Когда используется критерий Уилкоксона?
16. Когда используется критерий Манна–Уитни?
17. Когда используется критерий Крускала–Уоллиса?
18. Когда используется критерий Фридмана?

2.10. Кластерный и дискриминантный анализы

Краткие сведения из теории

Классификация – это процесс группировки объектов исследования или наблюдения в соответствии с их общими признаками. В науке классификация обозначает разновидность деления объёма понятия по определённому основанию (признаку, критерию), при котором объём родового понятия делится на виды (классы, множества), а виды, в свою очередь, делятся на подвиды (подклассы, подмножества) и т. д. Классификация широко применяется в науке и практической деятельности, особенно в естественных науках.

Существуют два типа классификации: дискриминантный анализ, который используется при наличии обучающих выборок, и автоматическая классификация, известная как кластерный анализ, которая не требует обучения.

Обучающие выборки – это набор данных, используемый для обучения и настройки алгоритмов машинного обучения. Они представляют собой пары «объект – ответ», где объект – это входные данные, а ответ –

это ожидаемый результат или классификация. Обучающие выборки используются для обучения программ, которые затем могут классифицировать новые данные на основе полученных знаний. Важной особенностью обучающих выборок является их способность к обобщению, т. е. к адекватному отклику на данные, выходящие за пределы имеющейся обучающей выборки. Обучающие выборки могут быть подготовлены как вручную, так и автоматически, на основе анализа пользовательских кликов или других средств. Они играют важную роль в машинном обучении и позволяют алгоритмам обучаться и адаптироваться к новым данным.

Кластерный анализ – это метод классификации, который используется для разделения объектов на группы (кластеры) на основе их сходства или близости. Этот метод не требует предварительного обучения и может быть применен к любым данным, независимо от их природы.

Кластерный анализ может быть использован в различных областях, включая биологию, социологию, экономику и другие. Он позволяет выявлять скрытые структуры в данных, определять группы объектов с похожими характеристиками и анализировать их различия.

Кластер в статистике – это класс родственных элементов статистической совокупности.

Существует множество алгоритмов кластерного анализа, каждый из которых имеет свои преимущества и недостатки. Некоторые из наиболее распространенных алгоритмов включают в себя иерархический кластерный анализ, анализ К-средних, алгоритм кластеризации DBSCAN и другие.

Иерархический кластерный анализ создает иерархическую структуру кластеров, начиная с отдельных объектов и постепенно объединяя их в более крупные группы.

Алгоритм К-средних разделяет данные на К кластеров, каждый из которых имеет свой центр. Объекты затем распределяются по кластерам в зависимости от их близости к центру.

DBSCAN – это алгоритм, который определяет кластеры как области высокой плотности, окруженные областями низкой плотности.

Кластерный анализ может быть использован для решения различных задач, таких как сегментация рынка, анализ социальных сетей, анализ генома и многое другое. Он позволяет получить глубокое понимание данных и выявить скрытые закономерности.

Основные этапы кластерного анализа:

1. Определение объектов, которые будут сравниваться друг с другом.
2. Выбор характеристик, по которым будет проводиться сравнение и описание объектов.
3. Вычисление меры сходства или различия между объектами в соответствии с выбранной метрикой.
4. Группировка объектов в кластеры с использованием определенной процедуры объединения.
5. Проверка применимости полученного кластерного решения (проверка построенной модели).

Методы кластеризации можно разделить на два типа: агломеративные и итеративные дивизивные. Агломеративные методы включают последовательное объединение наиболее близких объектов в один кластер. Этот процесс можно визуализировать с помощью дендрограммы, которая представляет собой дерево объединения. Итеративные дивизивные методы, с другой стороны, включают разделение данных на группы на основе определенных критериев.

Дискриминантный анализ – это метод классификации, который используется для разделения объектов на группы на основе их характеристик. Этот метод основан на обучении с использованием обучающих выборок, где объекты классифицированы заранее. Этот вид анализа является многомерным, так как он учитывает несколько параметров объекта одновременно, таких как температура, влажность, давление, состав крови и т. д.

Дискриминантный анализ позволяет определить, какие признаки наиболее важны для классификации объектов и какие признаки наиболее точно разделяют объекты на группы. Он также позволяет оценить, насколько хорошо классификация работает на основе данных, которые не были использованы при обучении.

Существует несколько методов дискриминантного анализа, включая линейный дискриминантный анализ, квазидискриминантный анализ и другие. Каждый из этих методов имеет свои преимущества и недостатки, и выбор метода зависит от конкретной задачи и типа данных.

Например, есть n объектов с m характеристиками. После измерений каждый объект описывается вектором из m характеристик: $x_1, \dots, x_m$, где $m > 1$. Задача состоит в том, чтобы на основе этих измерений отнести каждый объект к одной из заранее определенных групп (классов): $G_1, \dots, G_k$,

где $k \geq 2$. Нам необходимо построить решающее правило, которое позволит определить, к какой группе принадлежит объект на основе результатов измерений его параметров. Число групп известно заранее, а, кроме того, мы знаем, что каждый объект принадлежит к определенной группе.

Дискриминантный анализ широко используется в различных областях, включая медицину, биологию, социологию и другие. Он позволяет проводить анализ данных, выявлять скрытые закономерности и принимать обоснованные решения на основе полученных результатов.

Алгоритм дискриминантного анализа включает следующие шаги:

1. Подготовка данных: на этом этапе необходимо собрать данные, которые будут использоваться для обучения модели. Данные должны содержать информацию о характеристиках объектов и их классификации.

2. Выделение признаков: на этом этапе необходимо выбрать признаки, которые будут использоваться для классификации объектов. Признаки могут быть как количественными, так и качественными.

3. Обучение модели: на этом этапе необходимо обучить модель на основе данных, используя выбранные признаки. Модель может быть обучена с использованием различных методов, таких как линейный дискриминантный анализ, квазидискриминантный анализ и другие.

4. Оценка модели: после обучения модели необходимо оценить ее качество. Это можно сделать, например, с помощью перекрестной проверки или оценки точности классификации.

5. Применение модели: после оценки модели ее можно использовать для классификации новых объектов на основе их характеристик.

6. Анализ результатов: после классификации новых объектов необходимо проанализировать результаты и сделать выводы на основе полученных данных.

Важно отметить, что алгоритм дискриминантного анализа может быть адаптирован для различных задач и типов данных. Например, для многомерных данных можно использовать метод главных компонент для уменьшения размерности данных перед классификацией.

Задача

Норма общего холестерина – от 3,6 до 7,8 ммоль/л, рекомендуемый уровень холестерина – менее 5 ммоль/л. Высокий уровень холестерина сигнализирует об угрозе атеросклероза. Чем выше уровень липопротеинов низкой плотности (ЛПНП), тем больше риск возникновения сердечно-сосудистых заболеваний. Зададим следующие требования для анализа:

норма ЛПНП – менее 2,5 ммоль/л, максимально допустимое значение – 4,0 ммоль/л. Нормальный уровень сахара в крови варьируется в пределах 3,3–5,5 ммоль/л. Когда нарушается обмен веществ в организме, клетки тканей перестают правильно усваивать глюкозу, и она начинает накапливаться в крови. Тревожный звонок должен срабатывать, когда уровень глюкозы в крови натошак составляет от 5,6 до 6,9 ммоль/л. В таблице представлены данные об уровне общего холестерина, ЛПНП и глюкозы у людей с разным ростом и весом:

№	Рост, см	Вес, кг	Уровень общего холестерина, ммоль/л	ЛПНП, ммоль/л	Уровень глюкозы, ммоль/л
1	153,0	98,8	12,2	7,6	8,3
2	156,5	52,2	3,5	1,2	5,6
3	154,1	78	8,2	6,2	6,8
4	157,5	60	6,5	3,1	5,9
5	155,6	89,2	10,2	5,8	7,1
6	160,6	53	4,2	2,1	5,7
7	162,0	54	5,1	2,4	5,9
8	162,8	100,1	13,4	8,2	7,6
9	163,6	72	7,2	3,6	6,3
10	166,1	58,4	5,2	2,3	5,8
11	167,3	103	13,1	8,2	8,3
12	167,6	57,7	3,5	2,7	5,6
13	169,9	58,8	2,9	2,2	5,9
14	170,7	110,3	9,2	6,7	6,8
15	170,8	63,4	5,3	2,5	5,6
16	173,7	88,5	5,9	3,2	6,1
17	174,7	130,8	8,6	6,6	7,1
18	178,3	149,2	13,2	9,2	9,1
19	178,6	72,9	3,8	2,3	6,5
20	179,1	74,3	4	2,5	5,5
21	181,1	162	11,5	9,5	12,1
22	184,0	76,8	6,5	4,5	6,9
23	189,9	87,1	4,3	2,9	5,3
24	190,6	50,5	1,1	15,4	3,4
25	192,4	44,2	1,2	18,5	2,3
26	194,7	35,6	0,8	22,2	3,5
27	194,9	230,4	15,5	14,5	14,7
28	195,6	250	16,5	13,6	13,4
29	195,9	100	4,3	3,3	6,3
30	212,4	110	2,5	2,6	5,6

Алгоритм кластерного анализа в JASP

Проведем кластерный анализ, чтобы оценить, есть ли в представленных данных закономерности, выраженные в дифференцировке на кластеры.

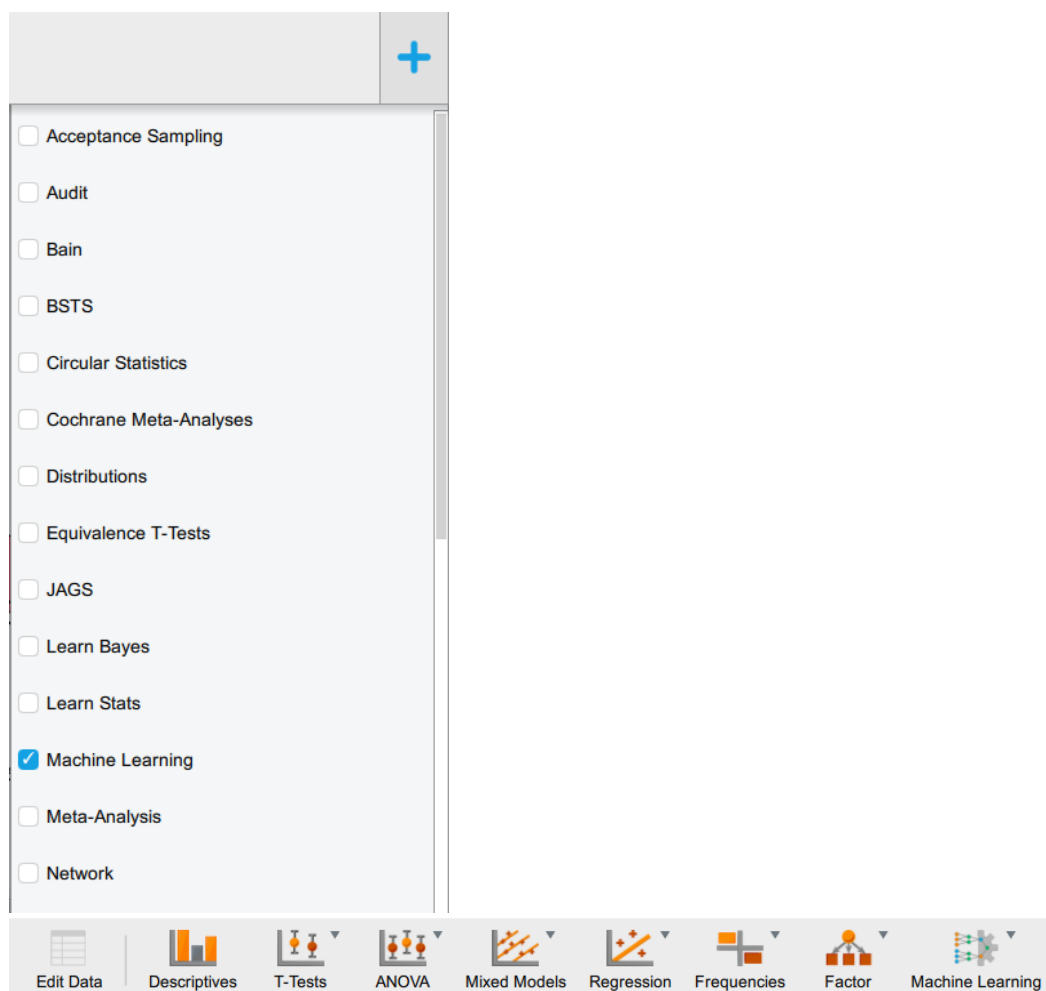
1. Скопируем данную таблицу в MS Excel:

	A	B	C	D	E	F
	№	Рост, см	Вес, кг	Уровень общего холестерина, ммоль/л	ЛПНП, ммоль/л	Уровень глюкозы, ммоль/л
1						
2	1	153,0	98,8	12,2	7,6	8,3
3	2	156,5	52,2	3,5	1,2	5,6
4	3	154,1	78	8,2	6,2	6,8
5	4	157,5	60	6,5	3,1	5,9
6	5	155,6	89,2	10,2	5,8	7,1
7	6	160,6	53	4,2	2,1	5,7
8	7	162,0	54	5,1	2,4	5,9
9	8	162,8	100,1	13,4	8,2	7,6
10	9	163,6	72	7,2	3,6	6,3
11	10	166,1	58,4	5,2	2,3	5,8
12	11	167,3	103	13,1	8,2	8,3
13	12	167,6	57,7	3,5	2,7	5,6
14	13	169,9	58,8	2,9	2,2	5,9
15	14	170,7	110,3	9,2	6,7	6,8
16	15	170,8	63,4	5,3	2,5	5,6
17	16	173,7	88,5	5,9	3,2	6,1
18	17	174,7	130,8	8,6	6,6	7,1
19	18	178,3	149,2	13,2	9,2	9,1

2. Сохраним данные в формате odf и откроем его в JASP:

	№	Рост, см	Вес, кг	Холестерин, ммоль/л	ЛПНП, ммоль/л	Глюкоза, ммоль/л
1	1	153	98.8	12.2	7.6	8.3
2	2	156.5	52.2	3.5	1.2	5.6
3	3	154.1	78	8.2	6.2	6.8
4	4	157.5	60	6.5	3.1	5.9
5	5	155.6	89.2	10.2	5.8	7.1
6	6	160.6	53	4.2	2.1	5.7
7	7	162	54	5.1	2.4	5.9
8	8	162.8	100.1	13.4	8.2	7.6
9	9	163.6	72	7.2	3.6	6.3
10	10	166.1	58.4	5.2	2.3	5.8
11	11	167.3	103	13.1	8.2	8.3
12	12	167.6	57.7	3.5	2.7	5.6
13	13	169.9	58.8	2.9	2.2	5.9
14	14	170.7	110.3	9.2	6.7	6.8
15	15	170.8	63.4	5.3	2.5	5.6
16	16	173.7	88.5	5.9	3.2	6.1
17	17	174.7	130.8	8.6	6.6	7.1
18	18	178.3	149.2	13.2	9.2	9.1
19	19	178.6	72.9	3.8	2.3	6.5

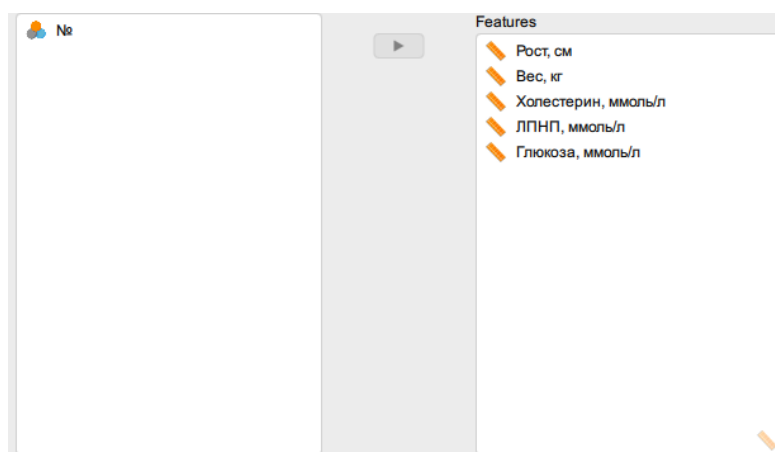
3. Перейдем во вкладку «Analyses» и подключим модуль «Machine learning». Для этого надо нажать на символ плюса в правом верхнем углу, выбрать из списка нужный пункт и поставить галочку. Модуль появится на рабочей панели для анализа данных:



4. Выберем нужный вид кластерного анализа. Для этого перейдем в модуль «Clustering» и нажмем «Neighborhood-Based»:



5. Перенесем переменные, как показано на рисунке:



6. Произведем настройку кластерного анализа:

Tables

- ☒ Cluster information
 - ☒ Within sum of squares
 - ☒ Silhouette score
 - ☒ Centers
 - ☒ Between sum of squares
 - ☒ Total sum of squares
- ☒ Model performance
- ☒ Cluster means

Export Results

- ☒ Add predictions to data

Column name

Plots

- ☒ Elbow method
- ☒ t-SNE cluster plot
 - ☒ Add data labels
- ☒ Cluster matrix plot
- ☒ Cluster means
 - ☒ Display barplot
- ☒ Group into one figure
- ☐ Cluster densities
 - ☐ Group into one figure

▼ Training Parameters

Algorithmic Settings

Center type

Algorithm

Distance

Max. iterations

Random sets

- ☒ Scale features
- ☐ Set seed

Cluster Determination

☐ Fixed

☒ Optimized according to

Clusters

Max. clusters

7. Получим следующие результаты:

K-Means Clustering

Clusters	N	R ²	AIC	BIC	Silhouette
4	30	0.788	70.730	98.750	0.450

Note. The model is optimized with respect to the BIC value.

В таблице показаны оценки соответствия модели с 4 кластерами ($K=4$) с использованием набора данных. Сообщается r^2 (т. е. соотношение сумм квадратов и общих сумм квадратов), которое также обычно указывается в моделях ANOVA/регрессии. Модель с r^2 , близким к верхней границе единицы, воспринимается как хорошо подходящая модель, тогда как r^2 близкая к нулю, указывает на то, что модель подходит плохо. Однако r^2 не проводит различие между хорошо подходящей моделью и моделью, которая переподгоняется.

Информационный критерий Акаике (AIC) и байесовский информационный критерий (BIC) также измеряют степень соответствия модели, но вдобавок к этому «наказывают» за количество свободных параметров модели. Наказывая за количество параметров, эти информационные критерии пытаются защитить от переобучения. Модели, имеющие более низкий информационный критерий, воспринимаются как модели, лучше обобщающие.

Показатель Silhouette отображает среднюю внутреннюю согласованность кластеризации, оценивая, насколько каждый случай похож на собственный кластер по сравнению с другими кластерами. Для оценок Silhouette общее правило состоит в том, что чем ближе к верхней границе (1), тем более последовательной является кластеризация, тогда как оценки Silhouette, близкие к нижней границе (–1), указывают на плохое совпадение.

В таблице об информации кластеров показаны размеры кластеров, изменчивость внутри каждого кластера с точки зрения внутренней суммы квадратов и центроиды каждого кластера в пространстве признаков:

Cluster Information

Cluster	1	2	3	4
Size	16	8	3	3
Explained proportion within-cluster heterogeneity	0.610	0.229	0.128	0.033
Within sum of squares	18.740	7.040	3.944	1.006
Silhouette score	0.430	0.400	0.446	0.732
Center Рост, см	–0.057	–0.689	1.004	1.136
Center Вес, кг	–0.421	0.299	2.422	–0.974
Center Холестерин, ммоль/л	–0.523	0.912	1.701	–1.345
Center ЛПНП, ммоль/л	–0.705	0.147	1.114	2.255
Center Глюкоза, ммоль/л	–0.351	0.304	2.483	–1.424

Note. The Between Sum of Squares of the 4 cluster model is 114.27

Note. The Total Sum of Squares of the 4 cluster model is 145

Качество модели оценивается таблицей метрик производительности:

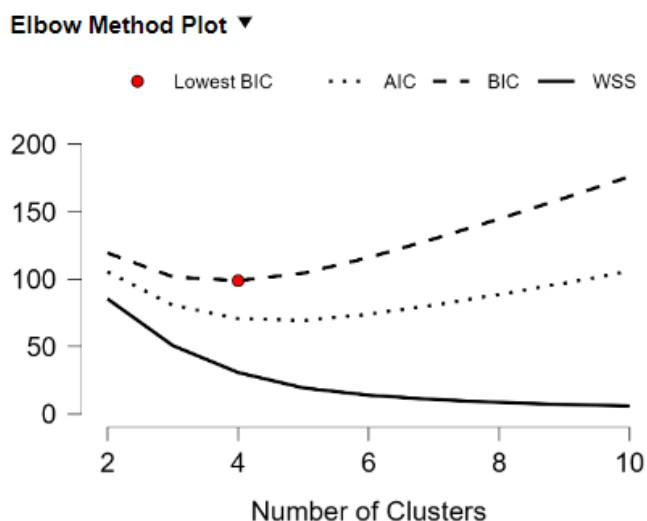
Model Performance Metrics	
	Value
Maximum diameter	3.836
Minimum separation	0.846
Pearson's γ	0.639
Dunn index	0.221
Entropy	1.148
Calinski-Harabasz index	32.227

Note. All metrics are based on the euclidean distance.

Среднее значение переменных для каждого кластера оценивается таблицей:

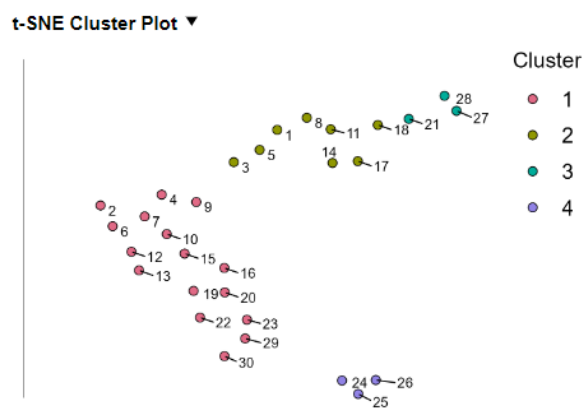
Cluster Means ▼					
	Рост, см	Вес, кг	Холестерин, ммоль/л	ЛПНП, ммоль/л	Глюкоза, ммоль/л
Cluster 1	-0.057	-0.421	-0.523	-0.705	-0.351
Cluster 2	-0.689	0.299	0.912	0.147	0.304
Cluster 3	1.004	2.422	1.701	1.114	2.483
Cluster 4	1.136	-0.974	-1.345	2.255	-1.424

Как видно из рисунка ниже, внутренние суммы квадратов r^2 монотонно уменьшаются по мере увеличения числа кластеров K и, таким образом, значения r^2 стремятся к единице по мере K -роста:

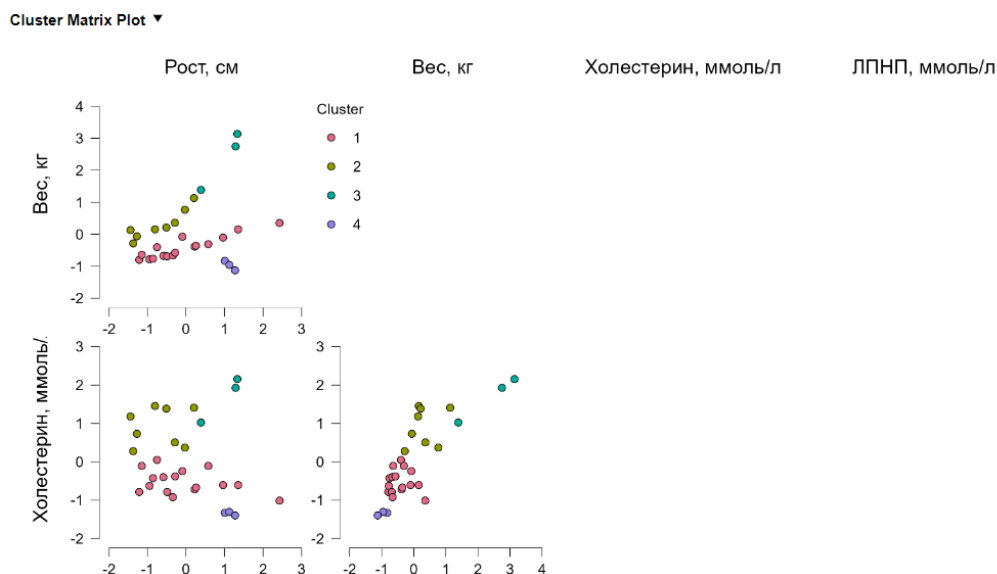


Следовательно, выбор K на основе r^2 бесполезен, поскольку тогда каждая точка данных будет находиться в отдельном кластере, что противоречит цели поиска сводки данных.

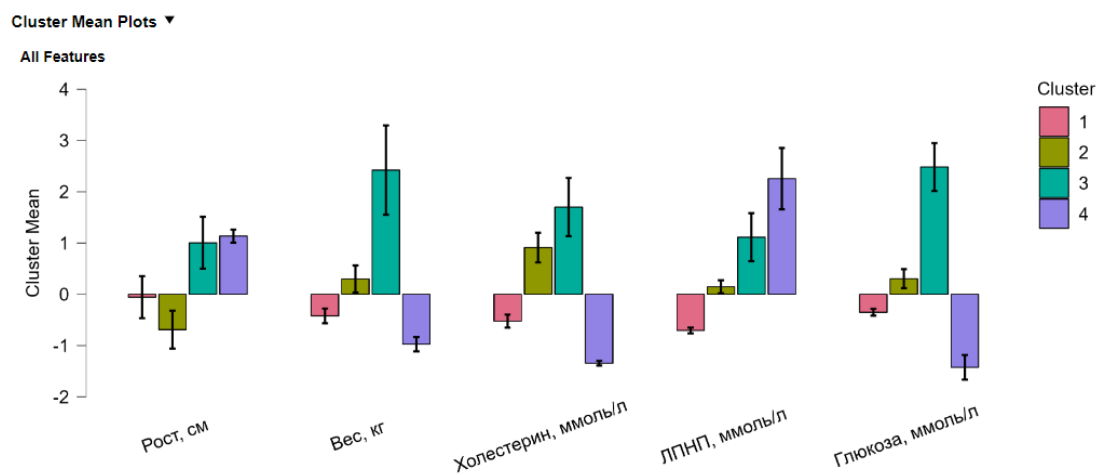
Для визуализации высокоразмерных данных используется график:



Для анализа различий и сходства между кластерами используется график:



Для визуализации средних значений переменных для каждого кластера используется график:



8. По результатам проведения анализа создадим таблицу данных по кластерам. Для этого вернемся в таблицу данных в JASP, выделим интересные столбцы, скопируем и вставим на лист в Excel. Теперь у нас есть переменная под названием «Кластер»:

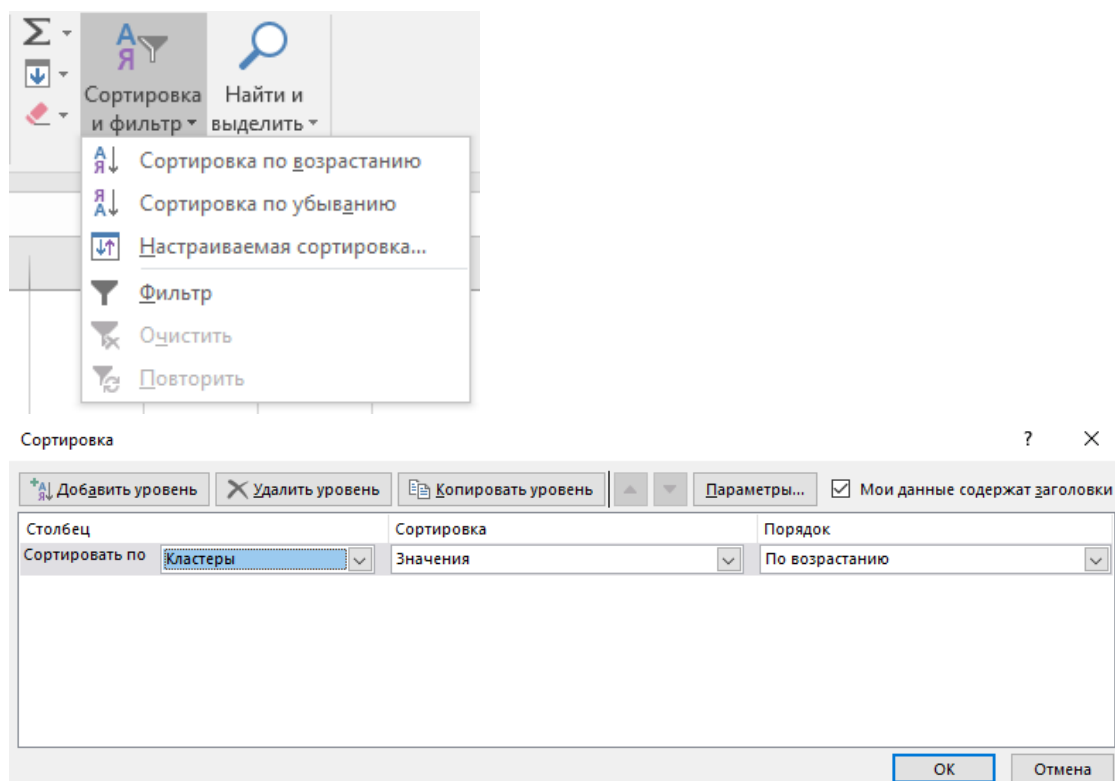
	Column 1	Рост, см	Вес, кг	Холестерин, ммоль/л	ЛПНП, ммоль/л	Глюкоза, ммоль/л	Кластер
1	1	153	98.8	12.2	7.6	8.3	2
2	2	156.5	52.2	3.5	1.2	5.6	1
3	3	154.1	78	8.2	6.2	6.8	2
4	4	157.5	60	6.5	3.1	5.9	1
5	5	155.6	89.2	10.2	5.8	7.1	2
6	6	160.6	53	4.2	2.1	5.7	1
7	7	162	54	5.1	2.4	5.9	1
8	8	162.8	100.1	13.4	8.2	7.6	2
9	9	163.6	72	7.2	3.6	6.3	1
10	10	166.1	58.4	5.2	2.3	5.8	1
11	11	167.3	103	13.1	8.2	8.3	2
12	12	167.6	57.7	3.5	2.7	5.6	1
13	13	169.9	58.8	2.9	2.2	5.9	1
14	14	170.7	110.3	9.2	6.7	6.8	2

Затем вернемся в MS Excel к нашей исходной таблице, расставим кластеры и затем отсортируем строки по кластерам:

№	А	В	С	Д	Е	F	G
	№	Рост, см	Вес, кг	Уровень общего холестерина, ммоль/л	ЛПНП, ммоль/л	Уровень глюкозы, ммоль/л	Кластеры
1							
2	1	153,0	98,8	12,2	7,6	8,3	2
3	2	156,5	52,2	3,5	1,2	5,6	1
4	3	154,1	78	8,2	6,2	6,8	2
5	4	157,5	60	6,5	3,1	5,9	1
6	5	155,6	89,2	10,2	5,8	7,1	2
7	6	160,6	53	4,2	2,1	5,7	1
8	7	162,0	54	5,1	2,4	5,9	1
9	8	162,8	100,1	13,4	8,2	7,6	2
10	9	163,6	72	7,2	3,6	6,3	1
11	10	166,1	58,4	5,2	2,3	5,8	1
12	11	167,3	103	13,1	8,2	8,3	2
13	12	167,6	57,7	3,5	2,7	5,6	1
14	13	169,9	58,8	2,9	2,2	5,9	1
15	14	170,7	110,3	9,2	6,7	6,8	2
16	15	170,8	63,4	5,3	2,5	5,6	1
17	16	173,7	88,5	5,9	3,2	6,1	1
18	17	174,7	130,8	8,6	6,6	7,1	2
19	18	178,3	149,2	13,2	9,2	9,1	2

После того как мы расставили кластеры, можно отсортировать значения по кластерам в порядке возрастания. Для этого выделим таблицу

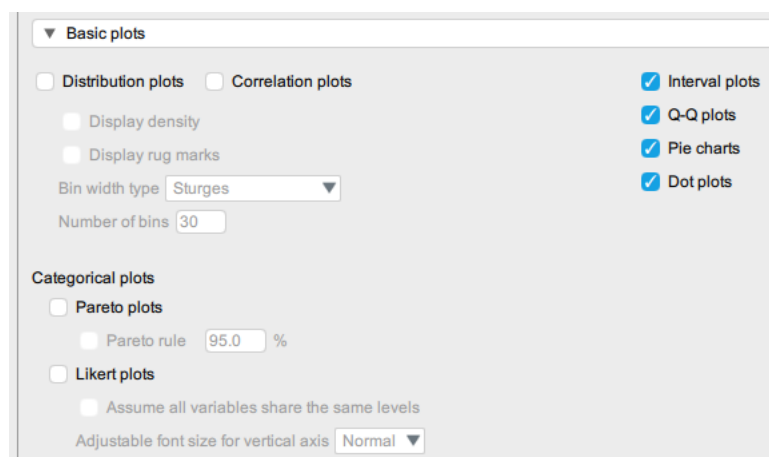
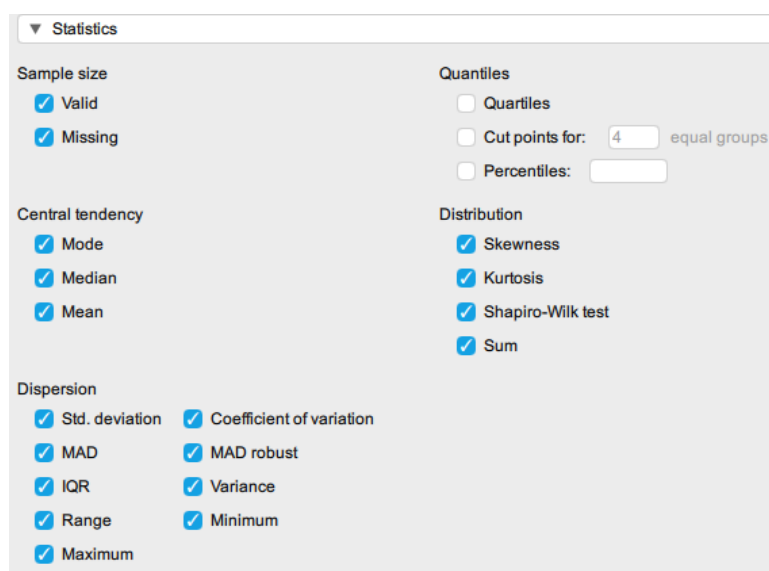
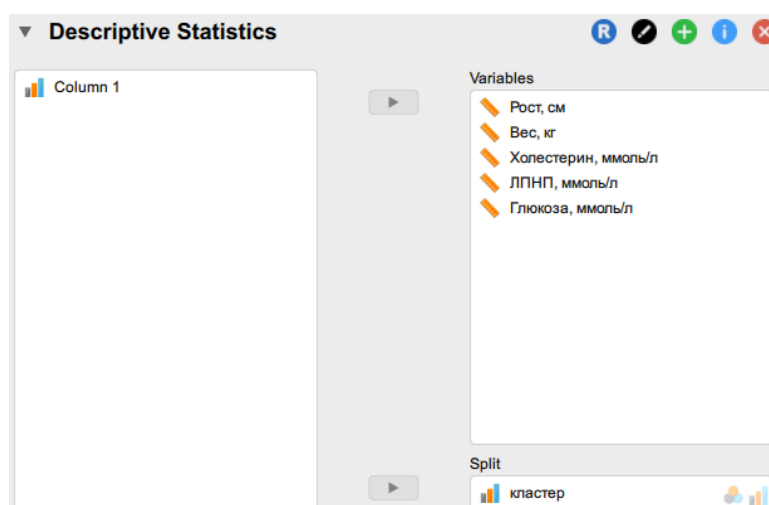
и перейдем через вкладку «Главная, затем «Сортировка и фильтры» и здесь выберем настраиваемую сортировку:

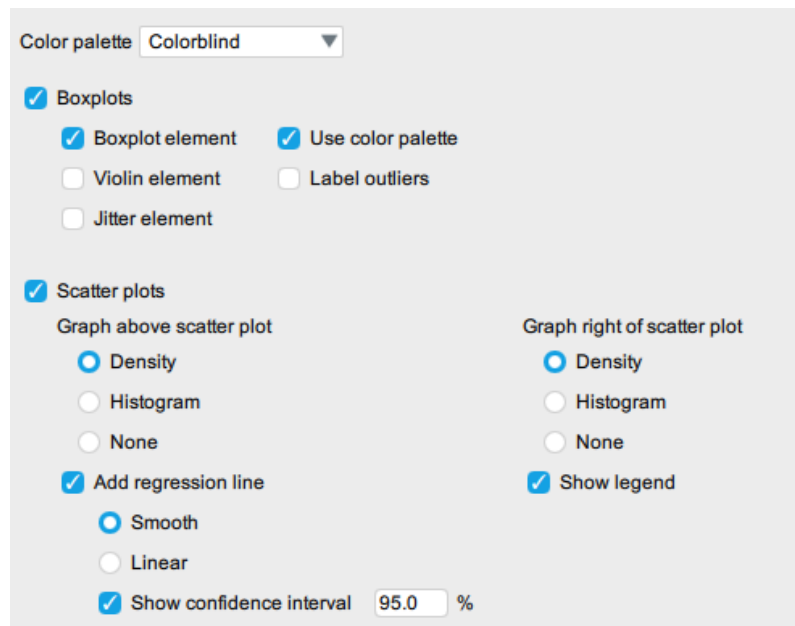


В результате получим следующую таблицу:

	A	B	C	D	E	F	G
	№	Рост, см	Вес, кг	Уровень общего холестерина, ммоль/л	ЛПНП, ммоль/л	Уровень глюкозы, ммоль/л	Кластеры
1							
2	2	156,5	52,2	3,5	1,2	5,6	1
3	4	157,5	60	6,5	3,1	5,9	1
4	6	160,6	53	4,2	2,1	5,7	1
5	7	162,0	54	5,1	2,4	5,9	1
6	9	163,6	72	7,2	3,6	6,3	1
7	10	166,1	58,4	5,2	2,3	5,8	1
8	12	167,6	57,7	3,5	2,7	5,6	1
9	13	169,9	58,8	2,9	2,2	5,9	1
10	15	170,8	63,4	5,3	2,5	5,6	1
11	16	173,7	88,5	5,9	3,2	6,1	1
12	19	178,6	72,9	3,8	2,3	6,5	1
13	20	179,1	74,3	4	2,5	5,5	1
14	22	184,0	76,8	6,5	4,5	6,9	1
15	23	189,9	87,1	4,3	2,9	5,3	1
16	29	195,9	100	4,3	3,3	6,3	1
17	30	212,4	110	2,5	2,6	5,6	1
18	1	153,0	98,8	12,2	7,6	8,3	2
19	3	154,1	78	8,2	6,2	6,8	2

9. Опишем полученные данные по кластерам. Для этого вернемся в JASP, выберем описательный модуль и проведем описательный анализ полученных данных по кластерам:

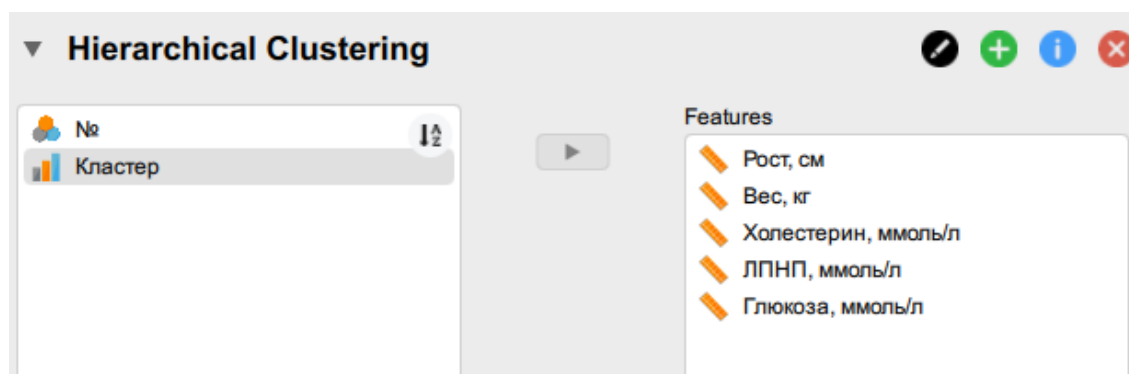




10. Построим дендрограмму. Для этого выберем в модуле «Machine learning» опцию «Clustering», а в ней – раздел «Hierarchical»:



11. Перенесем переменные, как показано на рисунке:



12. Установим следующие параметры:

Tables
☒ Cluster information
 ☒ Within sum of squares
 ☐ Silhouette score
 ☐ Between sum of squares
 ☐ Total sum of squares
 ☐ Model performance
 ☐ Cluster means

Plots
☐ Elbow method
 ☐ t-SNE cluster plot
 ☐ Add data labels
 ☐ Cluster matrix plot
 ☐ Cluster means
 ☒ Display barplot
 ☐ Group into one figure
 ☒ Cluster densities
 ☒ Group into one figure
 ☒ Dendrogram

Export Results
☒ Add predictions to data
 Column name:

Training Parameters

Algorithmic Settings
Distance:
Linkage:
☒ Scale features
☐ Set seed:

Cluster Determination
☒ Fixed
 Clusters:
☐ Optimized according to:
Max. clusters:

13. Получим следующие результаты:

Hierarchical Clustering ▾

Hierarchical Clustering ▾

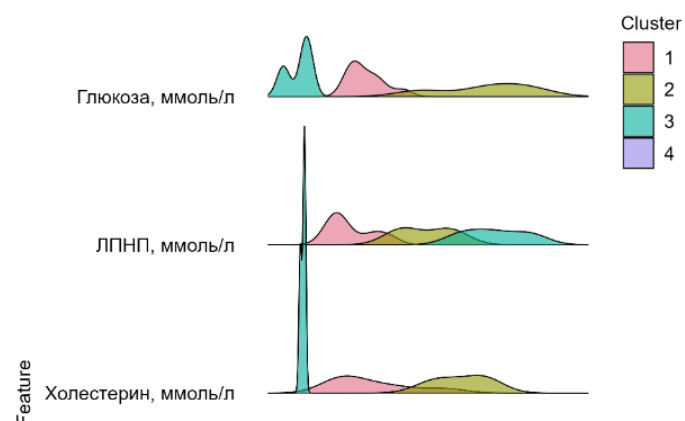
Clusters	N	R ²	AIC	BIC	Silhouette
4	30	0.713	81.600	109.630	0.450

Cluster Information

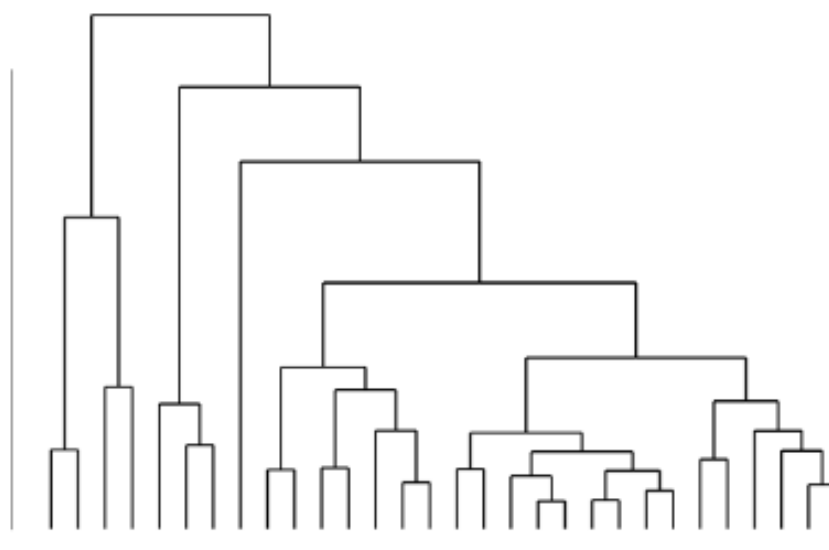
	Cluster	1	2	3	4
Size		22	4	3	1
Explained proportion within-cluster heterogeneity		0.783	0.192	0.024	0.000
Within sum of squares		32.590	8.006	1.006	0.000

Cluster Density Plots ▾

All Features ▾



Dendrogram ▼



14. Перейдем к исходным данным и сравним кластеры, полученные в этом анализе, с предыдущими:

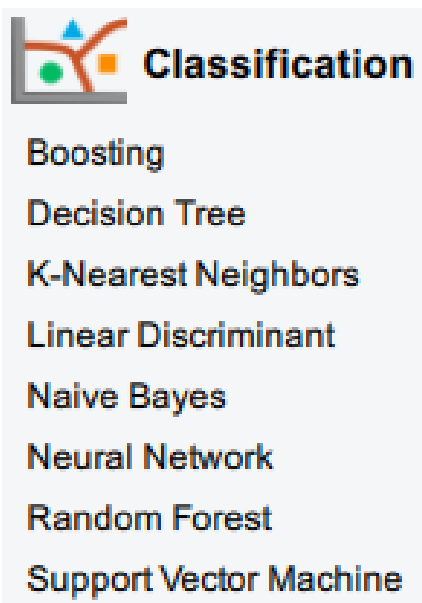
▼	Column 1	Рост, см	Вес, кг	Холестерин, ммоль/л	ЛПНП, ммоль/л	Глюкоза, ммоль/л	f _х кластер	f _х кластеры 2
1		153	98.8	12.2	7.6	8.3	2	1
2		156.5	52.2	3.5	1.2	5.6	1	1
3		154.1	78	8.2	6.2	6.8	2	1
4		157.5	60	6.5	3.1	5.9	1	1
5		155.6	89.2	10.2	5.8	7.1	2	1
6		160.6	53	4.2	2.1	5.7	1	1
7		162	54	5.1	2.4	5.9	1	1
8		162.8	100.1	13.4	8.2	7.6	2	1
9		163.6	72	7.2	3.6	6.3	1	1
10		166.1	58.4	5.2	2.3	5.8	1	1
11		167.3	103	13.1	8.2	8.3	2	1
12		167.6	57.7	3.5	2.7	5.6	1	1
13		169.9	58.8	2.9	2.2	5.9	1	1
14		170.7	110.3	9.2	6.7	6.8	2	1
15		170.8	63.4	5.3	2.5	5.6	1	1
16		173.7	88.5	5.9	3.2	6.1	1	1
17		174.7	130.8	8.6	6.6	7.1	2	1
18		178.3	149.2	13.2	9.2	9.1	2	2
19		178.6	72.9	3.8	2.3	6.5	1	1

Алгоритм дискриминантного анализа в JASP

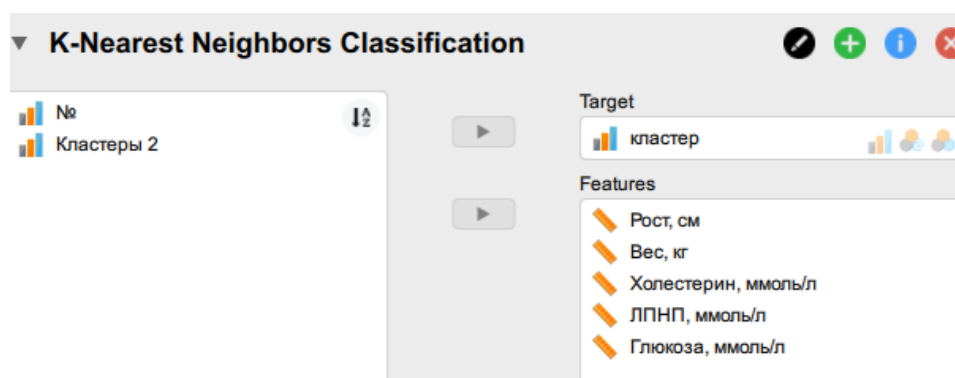
В прошлой работе нам удалось удачно провести анализ К-средних и иерархический анализ. Теперь мы можем выполнить дискриминантный анализ, так как наши переменные уже разбиты на соответствующие кластеры.

1. Откроем в модуле «Classification» раздел «K-Nearest Neighbors»:

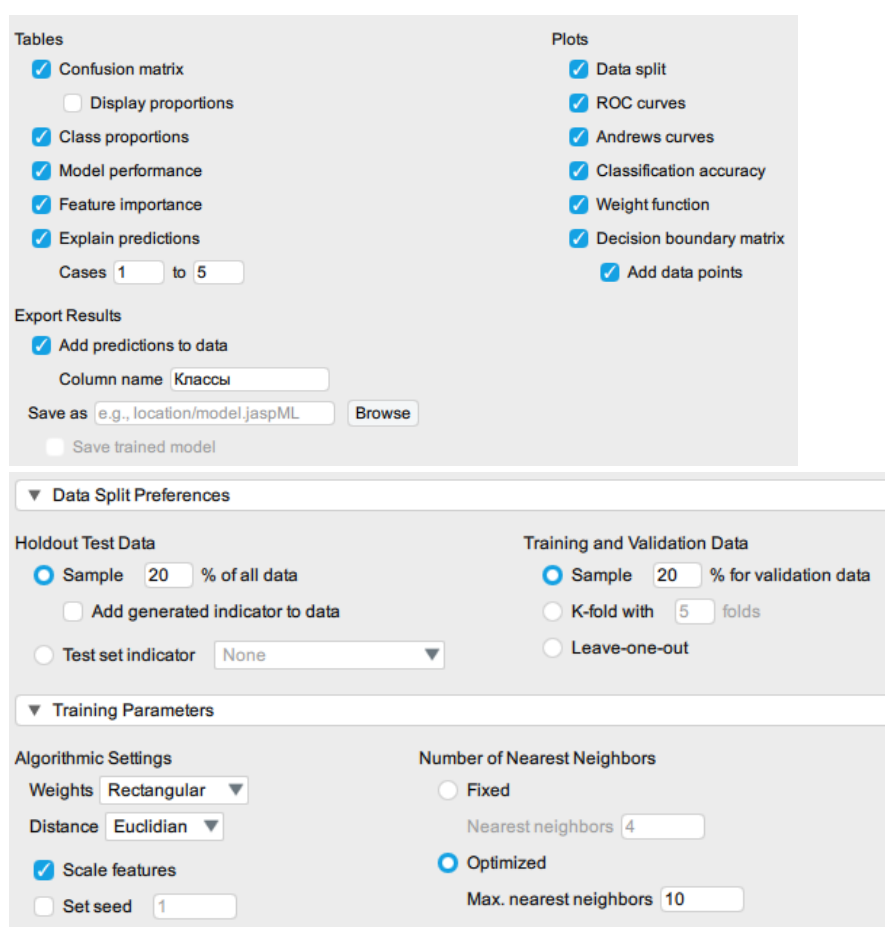
№	Рост, см	Вес, кг	Уровень общего холестерина, ммоль/л	ЛПНП, ммоль/л	Уровень глюкозы, ммоль/л
1	153,0	98,8	12,2	7,6	8,3
2	156,5	52,2	3,5	1,2	5,6
3	154,1	78	8,2	6,2	6,8
4	157,5	60	6,5	3,1	5,9
5	155,6	89,2	10,2	5,8	7,1
6	160,6	53	4,2	2,1	5,7
7	162,0	54	5,1	2,4	5,9
8	162,8	100,1	13,4	8,2	7,6
9	163,6	72	7,2	3,6	6,3
10	166,1	58,4	5,2	2,3	5,8
11	167,3	103	13,1	8,2	8,3
12	167,6	57,7	3,5	2,7	5,6
13	169,9	58,8	2,9	2,2	5,9
14	170,7	110,3	9,2	6,7	6,8
15	170,8	63,4	5,3	2,5	5,6
16	173,7	88,5	5,9	3,2	6,1
17	174,7	130,8	8,6	6,6	7,1
18	178,3	149,2	13,2	9,2	9,1
19	178,6	72,9	3,8	2,3	6,5
20	179,1	74,3	4	2,5	5,5
21	181,1	162	11,5	9,5	12,1
22	184,0	76,8	6,5	4,5	6,9
23	189,9	87,1	4,3	2,9	5,3
24	190,6	50,5	1,1	15,4	3,4
25	192,4	44,2	1,2	18,5	2,3
26	194,7	35,6	0,8	22,2	3,5
27	194,9	230,4	15,5	14,5	14,7
28	195,6	250	16,5	13,6	13,4
29	195,9	100	4,3	3,3	6,3
30	212,4	110	2,5	2,6	5,6



2. Перенесем переменные, как показано на рисунке:



3. Настроим следующие параметры:



4. Получим следующие результаты:

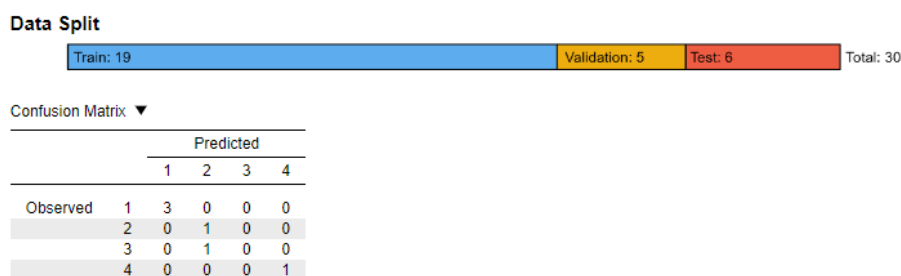
K-Nearest Neighbors Classification ▼

K-Nearest Neighbors Classification

Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
1	rectangular	Euclidean	19	5	6	0.800	0.833

Note. The model is optimized with respect to the validation set accuracy.

Из таблицы видно, что 19 наблюдений (80 % из 30) были использованы для обучения нашей К-модели, а 6 наблюдений в индикаторной переменной тестового набора использовались для получения ошибки прогнозирования. Информация о разделении данных также визуализируется с помощью панели, показанной на рисунке:



Результаты показывают, что наша К-модель достигла точности классификации тестового набора 0,833, а это означает, что мы можем правильно предсказать 83,3 % наших тестовых данных.

Матрица путаницы (или матрица ошибок) – это метод визуализации результатов алгоритма классификатора. Более конкретно, это таблица, в которой количество экземпляров основного истинного класса разбивается на количество предсказанных экземпляров класса. Матрицы неточностей – это один из нескольких показателей оценки, измеряющих эффективность модели классификации.

Процентное распределение объектов по разным классам можно оценить по таблице:

Class Proportions

	Data Set	Training Set	Validation Set	Test Set
1	0.533	0.526	0.600	0.500
2	0.267	0.316	0.200	0.167
3	0.100	0.105	0.000	0.167
4	0.100	0.053	0.200	0.167

Основные показатели качества модели можно увидеть в таблице:

Model Performance Metrics

	1	2	3	4	Average / Total
Support	3	1	1	1	6
Accuracy	1.000	0.833	0.833	1.000	0.917
Precision (Positive Predictive Value)	1.000	0.500	NaN	1.000	0.750
Recall (True Positive Rate)	1.000	1.000	0.000	1.000	0.833
False Positive Rate	0.000	0.200	0.000	0.000	0.050
False Discovery Rate	0.000	0.500	NaN	0.000	0.167
F1 Score	1.000	0.667	NaN	1.000	0.778
Matthews Correlation Coefficient	1.000	0.632	NaN	1.000	0.877
Area Under Curve (AUC)	1.000	0.900	0.500	1.000	0.850
Negative Predictive Value	1.000	1.000	0.833	1.000	0.958
True Negative Rate	1.000	0.800	1.000	1.000	0.950
False Negative Rate	0.000	0.000	1.000	0.000	0.250
False Omission Rate	0.000	0.000	0.167	0.000	0.042
Threat Score	∞	0.500	0.000	∞	∞
Statistical Parity	0.500	0.333	0.000	0.167	1.000

Note: All metrics are calculated for every class against all other classes.

Относительную важность каждого признака иллюстрирует таблица:

Feature Importance Metrics

	Mean dropout loss
ЛПНП, ммоль/л	26.894
Вес, кг	18.236
Холестерин, ммоль/л	14.737
Рост, см	12.710
Глюкоза, ммоль/л	11.237

Note. Mean dropout loss is based on 50 permutations.

Вклад каждого признака в предсказание модели для объектов показан в таблице:

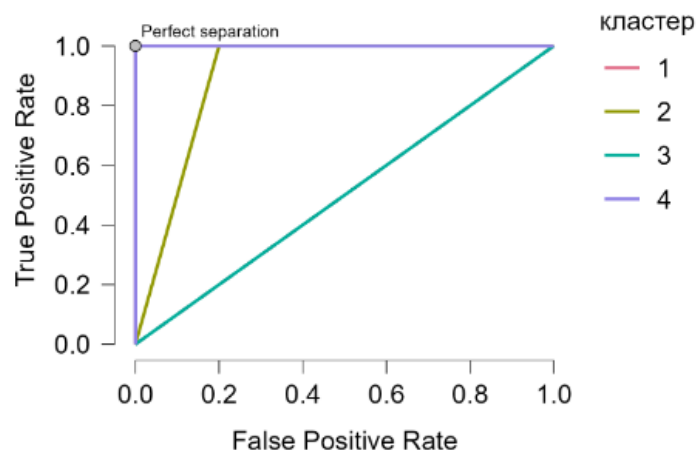
Additive Explanations for Predictions of Test Set Cases

Case	Predicted (Prob.)	Base	Рост, см	Вес, кг	Холестерин, ммоль/л	ЛПНП, ммоль/л	Глюкоза, ммоль/л
1	2 (1)	0.316	0.316	0.000	0.105	0.263	0.000
2	1 (1)	0.526	-0.158	0.211	0.263	0.158	0.000
3	1 (1)	0.526	-0.105	0.053	0.316	0.053	0.158
4	1 (1)	0.526	-0.105	0.053	0.263	0.263	0.000
5	2 (1)	0.316	-0.211	0.684	0.000	0.158	0.053

Note. Displayed values represent feature contributions to the predicted class probability without features (column 'Base') for the test set.

Кривая ROC (кривая рабочих характеристик приемника) представляет собой график, показывающий эффективность модели классификации при всех пороговых значениях классификации. Эта кривая отображает два параметра: истинную положительную скорость и ложноположительный результат:

ROC Curves Plot ▼



Визуализация многомерных данных показана на графике:

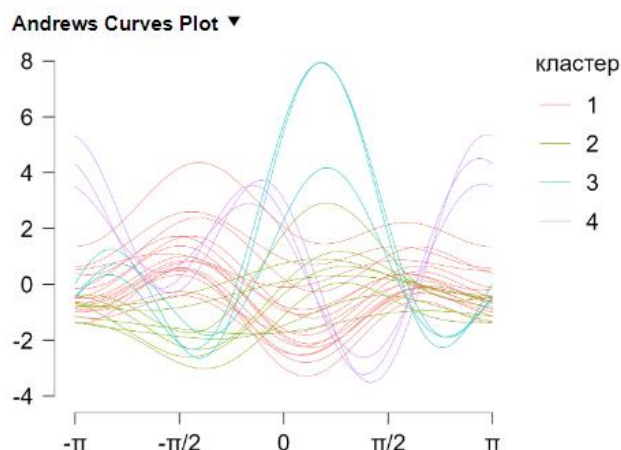
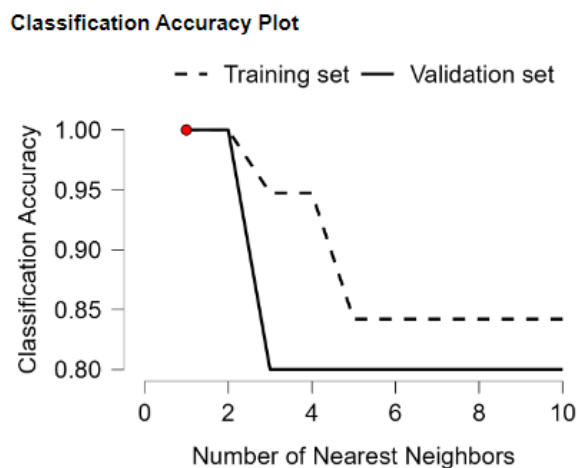
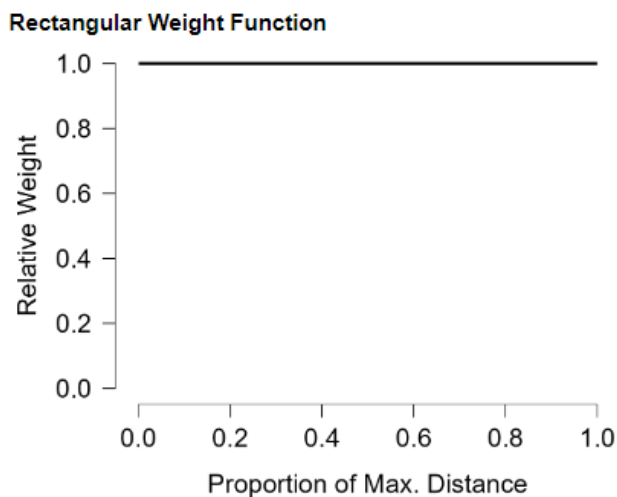


График точности классификации визуализирует производительность 10 прогонов алгоритма (на основе набора обучения и проверки) для моделей с $K=1$, $K=2$, ..., $K=10$:



Распределение весов отдельных данных показано на графике:



5. Сравним полученные классы в таблице с предыдущими классами, сделаем выводы:

№	Рост, см	Вес, кг	Холестерин, ммоль/л	ЛПНП, ммоль/л	Глюкоза, ммоль/л	Класс	Классы 2	Классы
1	153	98.8	12.2	7.6	8.3	2	1	2
2	156.5	52.2	3.5	1.2	5.6	1	1	1
3	154.1	78	8.2	6.2	6.8	2	1	2
4	157.5	60	6.5	3.1	5.9	1	1	1
5	155.6	89.2	10.2	5.8	7.1	2	1	2
6	160.6	53	4.2	2.1	5.7	1	1	1
7	162	54	5.1	2.4	5.9	1	1	1
8	162.8	100.1	13.4	8.2	7.6	2	1	2
9	163.6	72	7.2	3.6	6.3	1	1	1
10	166.1	58.4	5.2	2.3	5.8	1	1	1
11	167.3	103	13.1	8.2	8.3	2	1	2
12	167.6	57.7	3.5	2.7	5.6	1	1	1
13	169.9	58.8	2.9	2.2	5.9	1	1	1
14	170.7	110.3	9.2	6.7	6.8	2	1	2
15	170.8	63.4	5.3	2.5	5.6	1	1	1
16	173.7	88.5	5.9	3.2	6.1	1	1	1
17	174.7	130.8	8.6	6.6	7.1	2	1	2
18	178.3	149.2	13.2	9.2	9.1	2	2	2
19	178.6	72.9	3.8	2.3	6.5	1	1	1
20	179.1	74.3	4	2.5	5.5	1	1	1

6. С помощью обученной модели попробуем предсказать классы в новой таблице с данными, которые коррелируют с нынешними классами. Для этого сохраним полученную модель, указав путь:

Export Results

☒ Add predictions to data

Column name
Классы

Save as
Neighbors Classification.jspML
Browse

☒ Save trained model

7. Если не получается сохранить модель, необходимо произвести следующие действия:

- открыть MS Excel;
- оформить таблицу, полученную в результате кластерного анализа «Neighborhood-Based», следующим образом в MS Excel:

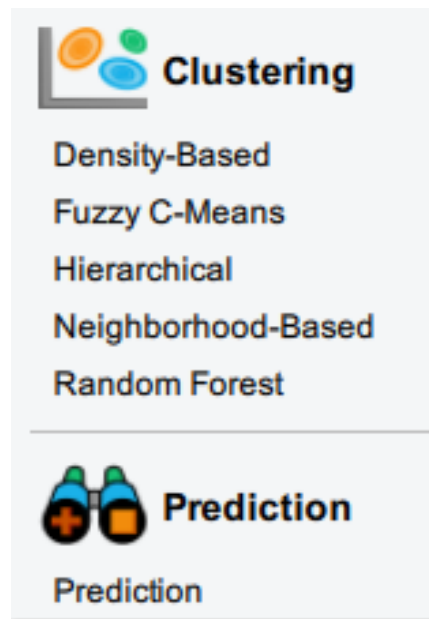
	A	B	C	D	E	F	G
1	Number	Height	Weight	Cholesterol	LDL	Glucose	Cluster
2	1	153	98.8	12.2	7.6	8.3	3
3	2	156.5	52.2	3.5	1.2	5.6	4
4	3	154.1	78	8.2	6.2	6.8	3
5	4	157.5	60	6.5	3.1	5.9	4
6	5	155.6	89.2	10.2	5.8	7.1	3
7	6	160.6	53	4.2	2.1	5.7	4
8	7	162	54	5.1	2.4	5.9	4
9	8	162.8	100.1	13.4	8.2	7.6	3
10	9	163.6	72	7.2	3.6	6.3	4
11	10	166.1	58.4	5.2	2.3	5.8	4
12	11	167.3	103	13.1	8.2	8.3	3
13	12	167.6	57.7	3.5	2.7	5.6	4
14	13	169.9	58.8	2.9	2.2	5.9	4
15	14	170.7	110.3	9.2	6.7	6.8	3
16	15	170.8	63.4	5.3	2.5	5.6	4
17	16	173.7	88.5	5.9	3.2	6.1	4
18	17	174.7	130.8	8.6	6.6	7.1	3
19	18	178.3	149.2	13.2	9.2	9.1	3
20	19	178.6	72.9	3.8	2.3	6.5	4
21	20	179.1	74.3	4	2.5	5.5	4
22	21	181.1	162	11.5	9.5	12.1	1
23	22	184	76.8	6.5	4.5	6.9	4
24	23	189.9	87.1	4.3	2.9	5.3	4
25	24	190.6	50.5	1.1	15.4	3.4	2
26	25	192.4	44.2	1.2	18.5	2.3	2
27	26	194.7	35.6	0.8	22.2	3.5	2
28	27	194.9	230.4	15.5	14.5	14.7	1
29	28	195.6	250	16.5	13.6	13.4	1
30	29	195.9	100	4.3	3.3	6.3	4
31	30	212.4	110	2.5	2.6	5.6	4

- сохранить документ в формате odf;
- открыть её в JASP;
- произвести расчеты, как было показано ранее, в «K-Nearest Neighbors»;
- сохранить модель в формате jaspML;
- проверить сохраненный документ в соответствующей папке, куда было произведено сохранение;
- закрыть текущую сессию в JASP.

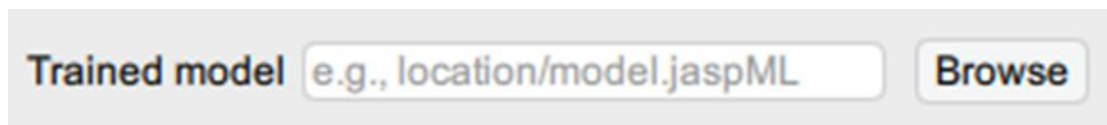
8. Закроем программу и откроем её заново. Загрузим в нее новые данные, представленные в таблице:

Number	Height	Weight	Cholesterol	LDL	Glucose
1	152,0	95,8	11,2	6,6	7,2
2	153,5	50,2	4,5	2,2	4,2
3	153,1	77	9,1	7,1	7,1
4	156,4	59	5,5	3,0	5,4
5	154,5	88,1	9,2	5,6	7,0
6	169,6	52	4,4	2,4	5,5
7	160,0	53	4,1	2,3	5,5
8	161,4	98,1	10,2	7,2	7,5
9	162,1	71	7,1	4,2	6,1
10	164,1	55,2	5,1	2,2	5,5
11	165,3	94	8,1	4,3	7,3
12	147,6	46,7	2,5	2,3	5,6
13	162,2	55,5	2,3	2,1	4,3
14	160,7	100,1	8,2	6,2	6,2
15	167,5	62,1	5,2	2,6	5,9
16	172,2	86,3	5,4	3,1	6,2
17	172,7	110,8	8,2	4,6	7,1
18	172,3	135,2	10,2	9,1	9,0
19	173,2	70,2	3,2	2,2	6,2
20	174,1	72,3	4,0	2,2	5,7
21	180,1	155	11,4	8,5	11,1
22	182,0	74,5	6,4	4,3	6,4
23	182,9	86,1	4,2	2,7	5,1
24	188,6	50,1	1,5	14,4	4,4
25	182,4	44,5	1,6	16,5	2,4
26	194,1	45,6	1,8	15,2	4,5
27	194,9	230,4	15,5	14,5	14,7
28	190,6	220	18,5	12,6	14,4
29	145,9	100	4,3	3,3	6,3
30	202,4	110	2,1	2,4	5,5

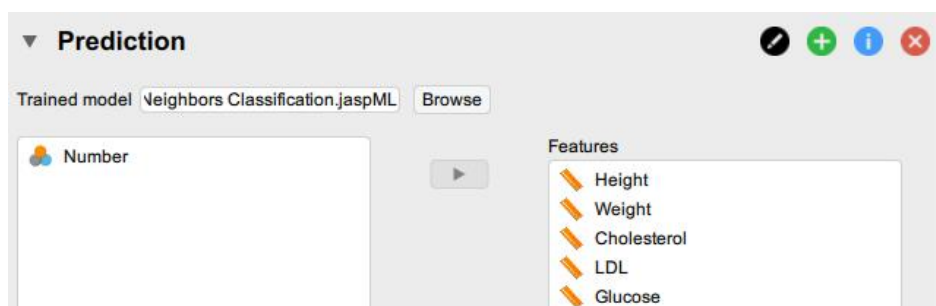
9. Откроем в модуле «Clustering» раздел «Prediction»:



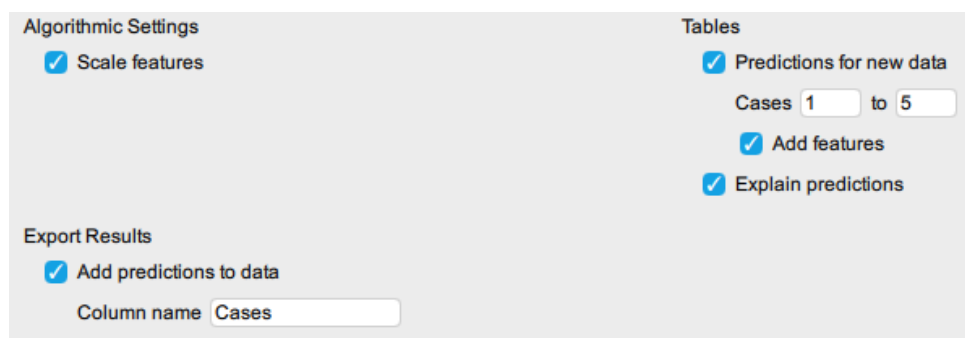
10. Загрузим нашу обученную модель, которую мы ранее сохранили, и посмотрим, как она прошла наше обучение:



11. Перенесем переменные, как показано на рисунке:



12. Зададим настройки:



13. Получим следующие результаты:

Loaded Model: Classification

Method	Nearest Neighbors	n(Train)	n(New)
K-nearest neighbors	1	19	30

Predictions for New Data

Case	Predicted	Height	Weight	Cholesterol	LDL	Glucose
1	3	-1.202	0.163	1.140	0.164	0.210
2	4	-1.101	-0.827	-0.485	-0.816	-0.920
3	3	-1.128	-0.245	0.631	0.276	0.172
4	4	-0.907	-0.636	-0.243	-0.638	-0.468
5	3	-1.034	-0.004	0.655	-0.059	0.134

Additive Explanations for Predictions of New Cases

Case	Predicted (Prob.)	Base	Height	Weight	Cholesterol	LDL	Glucose
1	3 (1)	0.211	0.368	0.053	0.368	0.000	0.000
2	4 (1)	0.632	0.000	0.053	0.105	0.105	0.105
3	3 (1)	0.211	0.632	0.000	0.158	0.000	0.000
4	4 (1)	0.632	0.000	0.053	0.158	0.105	0.053
5	3 (1)	0.211	0.316	0.105	0.158	0.211	0.000

Note. Displayed values represent feature contributions to the predicted class probability without features (column 'Base') for the test set.

14. Перейдем в раздел данных и проанализируем полученные классы. Для этого составим таблицу в MS Excel, как было описано в анализе «Neighborhood-Based»:

	Number	Height	Weight	Cholesterol	LDL	Glucose	fx Cases
1	1	152	95.8	11.2	6.6	7.2	3
2	2	153.5	50.2	4.5	2.2	4.2	4
3	3	153.1	77	9.1	7.1	7.1	3
4	4	156.4	59	5.5	3	5.4	4
5	5	154.5	88.1	9.2	5.6	7	3
6	6	169.6	52	4.4	2.4	5.5	4
7	7	160	53	4.1	2.3	5.5	4
8	8	161.4	98.1	10.2	7.2	7.5	3
9	9	162.1	71	7.1	4.2	6.1	4
10	10	164.1	55.2	5.1	2.2	5.5	4
11	11	165.3	94	8.1	4.3	7.3	3

При необходимости откорректируем полученные кластеры.

Контрольные вопросы

1. Что такое кластерный анализ?
2. Какие методы кластеризации существуют?
3. Как определить количество кластеров?
4. Каковы преимущества кластерного анализа?
5. В каких областях применяется кластерный анализ?

6. Чем кластерный анализ отличается от других методов анализа данных?
7. Какие алгоритмы используются в кластерном анализе?
8. Как выбрать подходящий метод кластеризации?
9. Как оценить качество полученных кластеров?
10. Каковы основные этапы кластерного анализа?
11. Какие существуют подходы к кластеризации?
12. Каковы основные принципы кластерного анализа?
13. Какие типы данных можно обрабатывать с помощью кластерного анализа?
14. Какие методы кластеризации основаны на методах машинного обучения?
15. Каковы перспективы развития кластерного анализа?
16. Что такое дискриминантный анализ?

2.11. Факторный анализ

Краткие сведения из теории

Факторный анализ – это статистический метод, который используется для выявления скрытых факторов, влияющих на набор переменных. Этот метод применим во многих науках, включая биологию и медицину.

Факторный анализ и кластерный анализ – это два разных метода статистического анализа данных. Они могут быть полезными в разных областях, но используются для разных целей и имеют свои особенности.

Факторный анализ используется для выявления скрытых факторов, влияющих на набор переменных. Он позволяет исследователям понять, какие факторы объясняют связь между переменными. Факторный анализ часто применяется в психологии, социологии, экономике, машинном обучении для изучения сложных явлений.

Кластерный анализ используется для группировки объектов на основе их сходства. Он позволяет исследователям выделить группы объектов, которые имеют схожие характеристики. Кластерный анализ часто используется в биологии, медицине и маркетинге для классификации объектов.

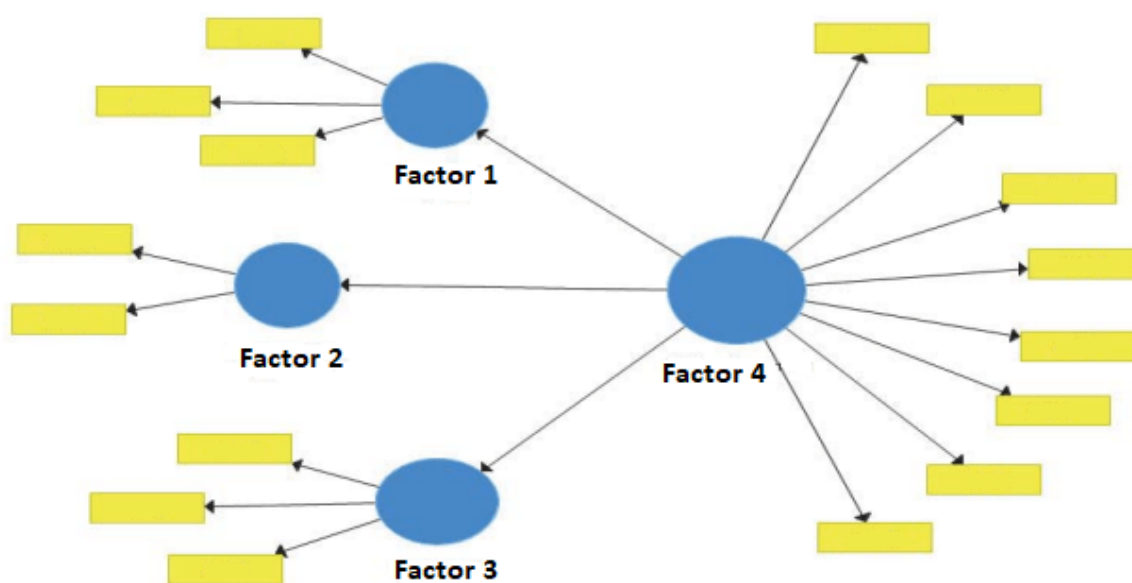
Основное различие между этими двумя методами заключается в том, что факторный анализ ищет скрытые факторы, влияющие на переменные, а кластерный анализ группирует объекты на основе их сходства.

Факторный анализ использует структуру корреляции между наблюдаемыми переменными для моделирования меньшего количества ненаблюдаемых скрытых переменных, известных как факторы. Исследователи используют этот статистический метод, когда знания в предметной области предполагают, что скрытые факторы вызывают ковариацию наблюдаемых переменных, и необходимо выявить скрытые переменные.

Аналитики часто называют наблюдаемые переменные индикаторами, поскольку они буквально указывают информацию о факторе. Факторный анализ рассматривает эти показатели как линейные комбинации факторов анализа с учетом ошибки. Процедура оценивает, какую часть дисперсии объясняет каждый фактор в показателях. Идея состоит в том, что скрытые факторы создают общность в некоторых наблюдаемых переменных.

Например, социально-экономический статус – это фактор, который невозможно измерить напрямую. Однако можно оценить род занятий, доход и уровень образования. Все эти переменные относятся к социально-экономическому статусу. Люди с определенным социально-экономическим статусом, как правило, имеют схожие значения наблюдаемых переменных. Если фактор имеет сильную связь с этими показателями, то на его долю приходится большая часть дисперсии показателей.

На рисунке показано, как четыре скрытых фактора (темные круги) влияют на измеримые значения (более светлые прямоугольники):



Цели факторного анализа

Факторный анализ упрощает сложный набор данных, беря большее количество наблюдаемых переменных и сводя их к меньшему набору ненаблюдаемых (скрытых) факторов. Каждый раз, когда мы что-то упрощаем, мы жертвуем точностью ради простоты понимания. В идеале мы получаем результат, в котором упрощение помогает лучше понять основную реальность предметной области. Однако этот процесс включает в себя несколько методологических и интерпретативных суждений. Действительно, хотя анализ выявляет факторы, исследователи должны их назвать. Следовательно, аналитики обсуждают результаты факторного анализа чаще, чем другие статистические анализы.

Хотя весь факторный анализ направлен на поиск скрытых факторов, исследователи используют его для достижения двух основных целей. Они либо хотят изучить и обнаружить структуру набора данных, либо подтвердить обоснованность существующих гипотез и инструментов измерения.

Исследовательский факторный анализ (EFA)

Ученые используют исследовательский факторный анализ (EFA), когда у них еще нет хорошего понимания факторов, присутствующих в наборе данных, содержащем множество переменных. В этом случае факторный анализ помогает им найти факторы в таком наборе данных. Рекомендуется использовать этот подход, прежде чем формировать гипотезы о закономерностях в том или ином наборе данных. При исследовательском факторном анализе исследователи, скорее всего, будут использовать статистические данные и графики, чтобы определить количество факторов, которые необходимо извлечь.

Исследовательский факторный анализ наиболее эффективен, когда с каждым фактором связано несколько переменных. В ходе EFA исследователи должны решить, как проводить анализ (например, определиться с количеством факторов, методом извлечения и типом ротации), поскольку не существует гипотез или инструментов оценки, которыми можно было бы руководствоваться. Необходимо использовать методологию, которая имеет смысл для данного исследования.

Например, исследователи могут использовать EFA для создания шкалы – набора вопросов, которыми измеряется один фактор. Исследовательский факторный анализ позволяет найти элементы опроса, которые нагружают определенные конструкции.

Подтверждающий факторный анализ (CFA)

Подтверждающий факторный анализ (CFA) является более жестким процессом, чем EFA. Используя этот метод, исследователи стремятся подтвердить существующие гипотезы, выдвинутые ими самими или другими. Этот процесс направлен на подтверждение предыдущих идей, исследований, а также инструментов измерения и оценки. Следовательно, характер того, что они хотят проверить, будет накладывать ограничения на анализ.

Перед факторным анализом исследователи должны указать свою методологию, включая метод извлечения, количество факторов и тип ротации. Они основывают свои решения на характере того, что они подтверждают. После этого исследователи определяют, соответствуют ли качество модели и характер факторных нагрузок предсказанным теорией или инструментами оценки.

В этом смысле подтверждающий факторный анализ может помочь оценить валидность конструкции. Базовые конструкции являются скрытыми факторами, а элементы инструмента оценки являются индикаторами. Аналогично, данный анализ может также оценивать достоверность систем измерения. Измеряет ли инструмент ту конструкцию, которую он якобы измеряет?

Например, исследователи могут захотеть подтвердить факторы, лежащие в основе элементов личностного опросника. Сопоставление набора данных потребует от исследователей методологического выбора, например, количества факторов.

Факторы

В этом контексте факторы – это более широкие концепции или конструкции, которые исследователи не могут измерить напрямую. Эти более глубокие факторы управляют другими наблюдаемыми переменными. Следовательно, исследователи делают выводы о свойствах ненаблюдаемых факторов, измеряя переменные, которые коррелируют с этим фактором. Таким образом, факторный анализ позволяет исследователям выявить факторы, которые они не могут оценить напрямую.

Психологи часто используют факторный анализ, потому что многие из его факторов по своей сути не наблюдаемы, поскольку они существуют внутри человеческого мозга.

Например, депрессия – это состояние внутри разума, которое исследователи не могут наблюдать напрямую. Однако они могут задавать вопросы и делать замечания о различном поведении и отношениях. Депрессия – это невидимая движущая сила, которая влияет на многие результаты,

которые мы можем измерить. Следовательно, люди с депрессией, как правило, имеют более схожие реакции на эти последствия, чем те, кто не находится в депрессии.

По тем же причинам факторный анализ в психологии часто выявляет и оценивает другие психические характеристики, такие как интеллект, настойчивость и чувство собственного достоинства. Исследователи могут увидеть, как набор измерений влияет на эти и другие факторы.

Методы извлечения факторов

Первым критерием выбора методологии факторного анализа является математический подход к извлечению факторов из того или иного набора данных. Наиболее распространенными вариантами являются максимальное правдоподобие (ML), факторинг по главной оси (PAF) и анализ главных компонентов (PCA).

Преимущественно следует использовать либо ML, либо PAF. ML используется, когда данные имеют нормальное распределение. Помимо извлечения факторных нагрузок, ML может также выполнять проверку гипотез, строить доверительные интервалы и рассчитывать статистику согласия.

PAF используется, когда данные нарушают многомерную нормальность. PAF не предполагает, что наши данные подчиняются какому-либо распределению, поэтому мы можем использовать их, когда они распределены нормально. Однако этот метод не может предоставить все статистические показатели, как ML.

PCA – это метод факторного анализа по умолчанию в некоторых пакетах статистического программного обеспечения. Это метод сокращения данных для поиска компонентов. Существуют технические различия, но, если коротко, факторный анализ направлен на выявление скрытых факторов, тогда как PCA предназначен только для сокращения данных. При расчете компонентов PCA не оценивает основные сходства, вызванные ненаблюдаемыми факторами.

PCA приобрел популярность, потому что это был более быстрый алгоритм во времена более медленных и дорогих компьютеров. Прежде чем использовать PCA для факторного анализа, желательно провести небольшое исследование, чтобы убедиться, что этот метод подходит для данного исследования.

Существуют и другие методы извлечения факторов, но в литературе по факторному анализу не было убедительных доказательств того, что какой-либо из них лучше, чем метод максимального правдоподобия (ML) или факторинг по главной оси (PAF).

Количество факторов для извлечения

Следует указать количество факторов, которые необходимо извлечь из данных, за исключением случаев использования компонентов главных компонент. Метод определения этого числа зависит от того, какой факторный анализ выполняется – исследовательский или подтверждающий.

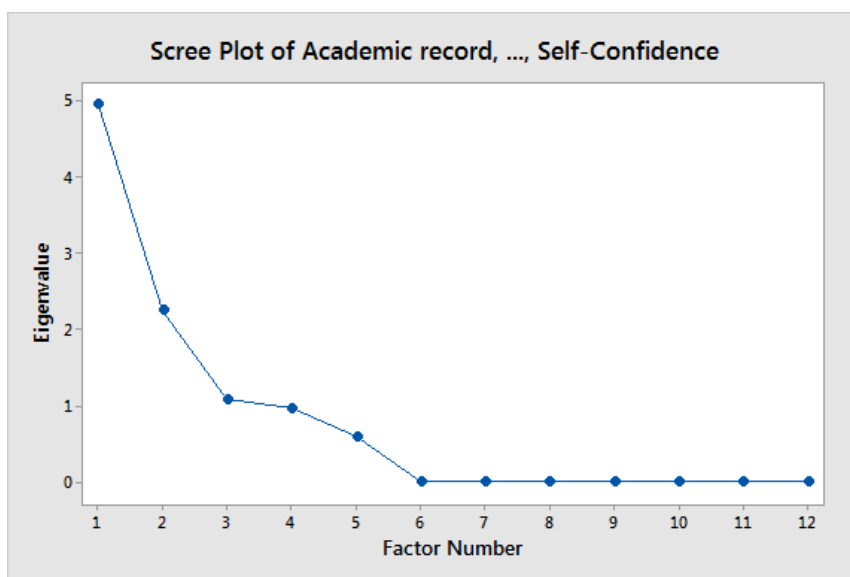
В ЕФА исследователи должны указать количество факторов, которые необходимо сохранить. Максимальное количество факторов, которые можно извлечь, равно количеству переменных в наборе данных. Однако обычно мы хотим максимально сократить количество факторов, максимизируя при этом общую величину дисперсии, которую эти факторы объясняют.

Это понятие экономной модели в статистике. При добавлении факторов происходит убывающая отдача. В какой-то момент мы обнаруживаем, что дополнительный фактор не увеличивает существенно объясненную дисперсию. В этом случае добавление факторов излишне усложняет модель. Следует использовать простейшую модель, которая объясняет большую часть дисперсии.

К счастью, такой простой статистический инструмент, как *график осыпи*, помогает справиться с этим компромиссом.

Для создания графика осыпи используется статистическое программное обеспечение. На графике нужно найти изгиб в данных, где кривая выравнивается. Количество точек до изгиба часто является корректным количеством факторов для извлечения.

Пример графика осыпи:



На графике отображаются собственные значения по количеству факторов. Собственные значения факторов относятся к величине объясненной дисперсии.

График осыпи на рисунке выше показывает изгиб кривой, происходящий при количестве факторов, равном 6. Следовательно, в данном случае нужно извлечь пять факторов. Они объясняют большую часть различий. Дополнительные факторы мало что объясняют.

Конечно, изучая данные и оценивая результаты, можно использовать теорию и знания в предметной области, чтобы корректировать количество факторов. Факторы и их интерпретации должны соответствовать контексту исследования.

В СФА исследователи определяют количество факторов, которые следует сохранить, используя существующую теорию или инструменты измерения, прежде чем проводить анализ. Например, если измерительный прибор предназначен для измерения трех показателей, то факторный анализ должен помочь извлечь три фактора и оценить, соответствуют ли результаты теории.

Факторные нагрузки

В факторном анализе нагрузки описывают отношения между факторами и наблюдаемыми переменными. Оценивая факторные нагрузки, можно понять силу связи между каждой переменной и фактором. Дополнительно можно идентифицировать наблюдаемые переменные, соответствующие конкретному фактору.

Интерпретируем нагрузки как коэффициенты корреляции. Значения варьируются от -1 до $+1$. Знак указывает направление связи (положительное или отрицательное), а абсолютное значение указывает на силу. Более сильные отношения имеют факторную нагрузку ближе к -1 и $+1$. Более слабые отношения близки к нулю.

Более сильные связи в контексте факторного анализа указывают на то, что факторы объясняют большую часть различий в наблюдаемых переменных.

Ротация факторов

В факторном анализе исходный набор нагрузок является лишь одним из бесконечного числа возможных решений, которые одинаково описывают данные. К сожалению, первоначальный ответ часто трудно интерпретировать, поскольку каждый фактор может содержать средние

нагрузки для многих показателей. Это затрудняет их маркировку. Определенные переменные сильно коррелируют с фактором, в то время как большинство других не коррелируют вообще. Поэтому резкий контраст между высокими и низкими нагрузками облегчает задачу определения количества факторов.

Ротация факторов решает эту проблему путем максимизации и минимизации всего набора факторных нагрузок. Цель состоит в том, чтобы создать ограниченное количество высоких нагрузок и множество низких нагрузок для каждого фактора.

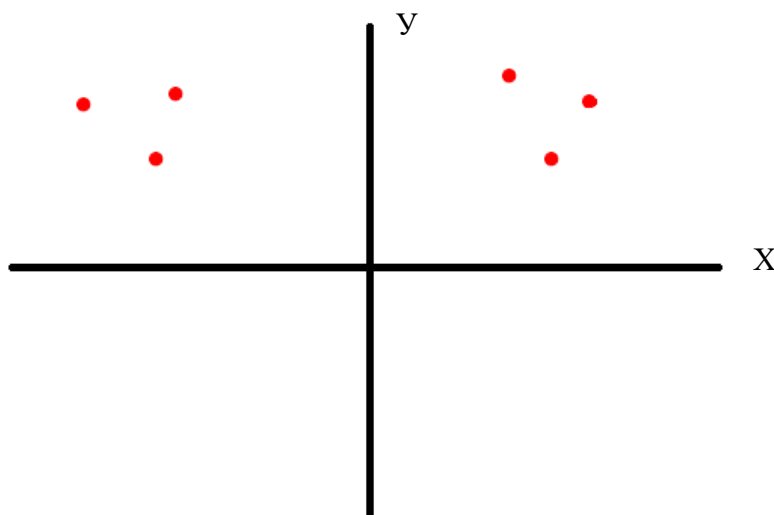
Эта комбинация позволяет выявить относительно небольшое количество показателей, которые сильно коррелируют с фактором, и большее количество переменных, которые с ним не коррелируют. Становится легче определить, что относится к фактору, а что нет. Это упрощает результаты факторного анализа и облегчают их интерпретацию.

Графическая иллюстрация вращения факторов

Покажем, как работает ротация факторов графически с помощью диаграмм рассеяния.

Факторный анализ начинается с расчета структуры факторных нагрузок. Однако он выбирает произвольный набор осей для их отчета. Вращение осей без изменения точек данных сохраняет исходную модель и структуру данных на месте, обеспечивая при этом более интерпретируемые результаты.

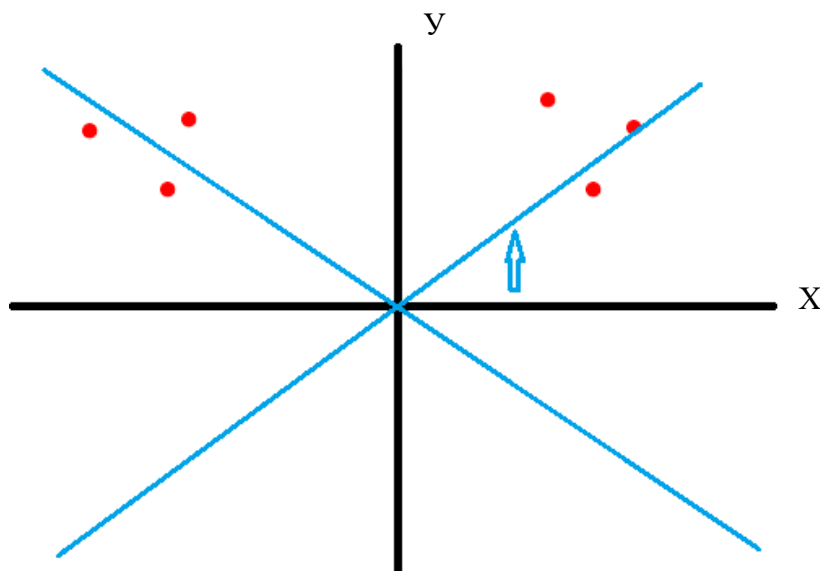
Чтобы сделать это графически в двух измерениях, будем использовать два фактора, представленные осями X и Y. На диаграмме рассеяния ниже шесть точек данных представляют наблюдаемые переменные, а координаты X и Y указывают их нагрузки для двух факторов:



В идеале точки располагаются прямо на оси, поскольку это показывает высокую нагрузку для одного фактора и нулевую нагрузку для другого.

Для исходного решения факторного анализа на диаграмме рассеяния точки содержат смесь координат X и Y и не расположены близко к оси фактора. Это затрудняет интерпретацию результатов, поскольку переменные имеют среднюю нагрузку на все факторы. Визуально они не сгруппированы возле осей, что затрудняет присвоение фактора переменным.

Вращение осей вокруг диаграммы рассеяния увеличивает или уменьшает значения X и Y , сохраняя при этом исходный шаблон точек данных. Во вращении (более тонкие линии на графике ниже) мы максимизируем одну факторную нагрузку и минимизируем другую для всех точек данных:



В результате нагрузки по одному показателю высоки, а по другому – низкие.

На графике все точки данных группируются рядом с одним из двух факторов на более тонкой повернутой оси, что позволяет легко связать наблюдаемые переменные с одним фактором.

Типы ротаций

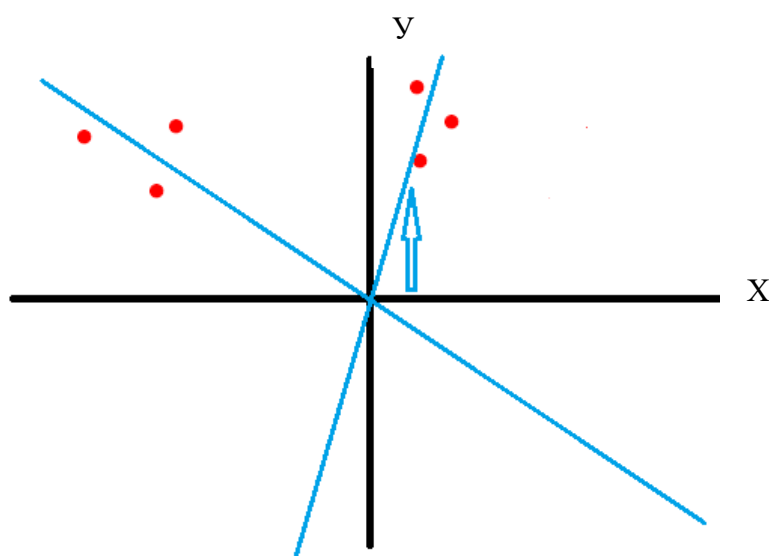
На протяжении всех этих ротаций мы работаем с одними и теми же точками данных и моделью факторного анализа. Модель соответствует данным как для повернутых нагрузок, так и для начальных нагрузок, но их легче интерпретировать. Мы используем другую систему координат, чтобы по-другому взглянуть на один и тот же набор точек.

В факторном анализе есть два основных типа вращения: наклонное и ортогональное.

Наклонные вращения допускают корреляцию между факторами, тогда как ортогональные вращения предполагают, что они совершенно некоррелированы.

Графически ортогональные повороты обеспечивают разделение между осями на 90° , как показано на предыдущем графике, где повернутые оси образуют прямые углы.

Для наклонных вращений не требуется, чтобы оси образовывали прямые углы, как показано ниже для другого набора данных:



Обратим внимание, что свобода принятия любой ориентации каждой оси позволяет им более точно соответствовать данным, чем при применении ограничения 90° . Следовательно, в некоторых случаях косые вращения могут создавать более простые структуры, чем ортогональные вращения. Однако эти результаты могут содержать коррелирующие факторы.

На практике наклонные вращения дают те же результаты, что и ортогональные вращения, когда факторы в реальном мире не коррелируют. Однако, если применить ортогональное вращение к действительно коррелирующим факторам, это может отрицательно повлиять на результаты. Несмотря на преимущества наклонного вращения, аналитики склонны чаще использовать ортогональные вращения, что в некоторых случаях может быть ошибкой.

Выбирая метод ротации в факторном анализе, необходимо убедиться, что он соответствует нашим основным предположениям и знаниям предметной области о том, коррелируют ли факторы.

Факторный анализ включает следующие этапы:

1. Сбор данных о переменных, которые мы хотим изучить. Это могут быть результаты опросов, экспериментов или наблюдений.

2. Вычисление корреляций между всеми парами переменных. Корреляция показывает, насколько сильно две переменные связаны друг с другом.

3. Выделение скрытых факторов, которые влияют на переменные. Факторы – это абстрактные понятия, которые объясняют, почему переменные связаны друг с другом.

4. Оценка значимости факторов с помощью статистических тестов, таких как тест Бартлетта или тест Кайзера–Мейера–Оливера.

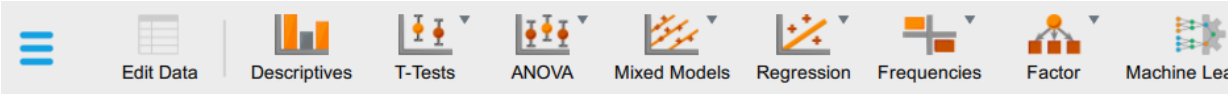
5. Интерпретация содержания факторов.

6. Проверка гипотез о влиянии факторов на переменные.

Факторный анализ позволяет исследователям лучше понять сложные явления и сделать более точные прогнозы. Однако он требует большого объема данных.

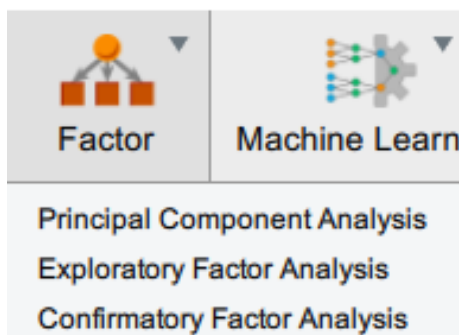
Алгоритм анализа в JASP

Для иллюстрации этапов факторного анализа возьмем таблицу данных, использованную в кластерном и дискриминантном анализе, и откроем эти данные в JASP:

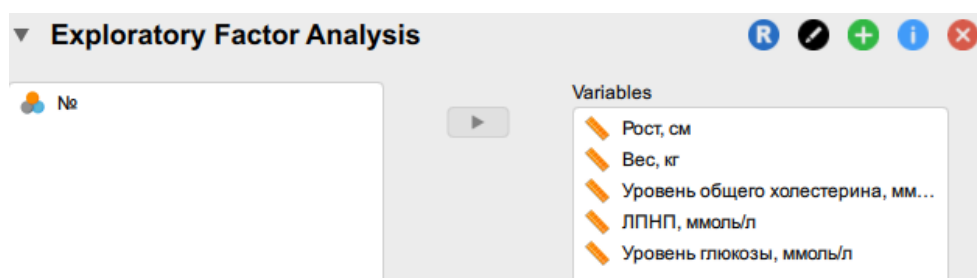


	№	Рост, см	Вес, кг	Уровень общего холестерина, ммоль/л	ЛПНП, ммоль/л	Уровень глюкозы, ммоль/л
1	1	153	98.8	12.2	7.6	8.3
2	2	156.5	52.2	3.5	1.2	5.6
3	3	154.1	78	8.2	6.2	6.8
4	4	157.5	60	6.5	3.1	5.9
5	5	155.6	89.2	10.2	5.8	7.1
6	6	160.6	53	4.2	2.1	5.7
7	7	162	54	5.1	2.4	5.9
8	8	162.8	100.1	13.4	8.2	7.6
9	9	163.6	72	7.2	3.6	6.3
10	10	166.1	58.4	5.2	2.3	5.8
11	11	167.3	103	13.1	8.2	8.3

1. Выберем нужный вид факторного анализа. Для этого перейдем в модуль «Factor» и найдем «Exploratory Factor Analysis»:

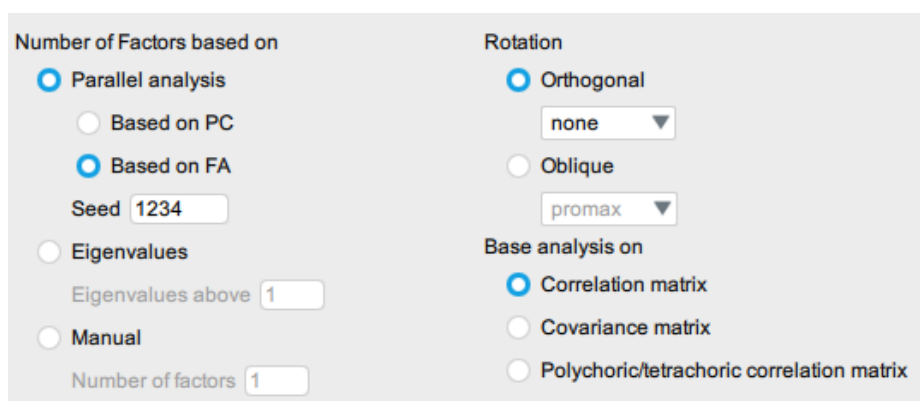


2. Перенесем переменные:



При выборе переменных для факторного анализа следует учитывать проблему мультиколлинеарности, т. е. излишнее коррелирование между переменными. Необходимо провести корреляционный анализ и убедиться, что нет переменных, тесно связанных между собой (с коэффициентом корреляции $r > 0,95$). Кроме того, не следует включать в анализ переменные, которые являются производными друг от друга или получены из одного и того же источника в разное время или при разных условиях. Это может привести к усилению совместной дисперсии и возникновению артефактов статистического анализа. Рекомендуется использовать как минимум три переменные для проведения факторного анализа.

3. Настроим факторный анализ:



Используемые пункты меню анализа:

Number of factors based on – позволяет выбрать количество выделяемых факторов. В JASP представлены три варианта:

– **Parallel analysis** – параллельный анализ (наиболее предпочтительный способ), имеет две опции: 1) Based on PC – основанный на методе выделения главных компонент; 2) Based on FA – основанный на методе эксплораторного факторного анализа;

– **Eigenvalues** – способ выделения количества факторов, при котором выделяются те факторы, которые имеют указанное собственное значение. Чем выше значение, тем более строгий отбор. Установленный по умолчанию уровень (равный 1) носит название критерия Кайзера;

– **Manual** – способ вручную указать количество выделяемых факторов. Обычно используется для проверки разных вариантов факторной структуры.

Factoring method – позволяет выбрать тот или иной метод извлечения факторов. Наиболее популярный вариант – метод максимального правдоподобия (Maximum likelihood).

Rotation – позволяет выбрать способ вращения факторов, имеет два варианта: 1) ортогональный, или прямоугольный (Orthogonal), приводящий к такому факторному решению, где факторы не могут коррелировать между собой ($r = 0$); 2) обlique, или косоугольный (Oblique), где факторы могут коррелировать.

4. Настроим «параметры вывода» (Output options):

▼ Output Options

Display loadings above

0.4

Order factor loadings by

☒ Factor size

☐ Variables

Assumption checks

☒ KMO test

☒ Bartlett's test

☒ Mardia's test

Tables

☐ Structure matrix

☐ Factor correlations

☐ Additional fit indices

☐ Residual matrix

☐ Parallel analysis

Based on PC

Based on FA

Plots

☒ Path diagram

☒ Scree plot

☒ Parallel analysis results

Missing Values

☒ Exclude cases pairwise

☐ Exclude cases listwise

Используемые пункты подменю анализа:

Display loadings above – позволяет указать пороговое значение факторных нагрузок, которые будут показаны в таблице факторных нагрузок. На рисунке выше указано значение 0,4. Соответственно, факторные нагрузки, которые меньше данного значения, не будут отображаться в таблице.

Order factor loadings by – позволяет упорядочить строки в таблице факторных нагрузок либо по величине факторных нагрузок (Factor size), либо по изначальному порядку, указанному в окне «Variables».

Path diagram – путевая диаграмма. Данный вариант позволяет представить матрицу факторных нагрузок в виде диаграммы, где переменные представлены прямоугольниками, стрелки являются связями (корреляциями и регрессиями), а факторы представлены окружностями. Чем ярче цвет стрелок, тем сильнее связь между факторами и переменными.

Scree plot – график каменистой осыпи. Данный график помогает определить число факторов.

KMO test – мера выборочной адекватности Кайзера–Мейера–Олкина (КМО), используемая для проверки гипотезы о том, что частные корреляции между переменными малы. Общее значение для всех переменных должно быть больше 0,5.

Bartlett's test – критерий сферичности Бартлетта, который проверяет гипотезу о том, что корреляционная матрица является единичной матрицей. При $p < 0,05$ принимается решение, что корреляционная матрица отличается от нулевой, и это является требованием для использования факторного анализа.

Mardia's test – критерий согласия Мардиа, который позволяет оценить допущение о соответствии распределения нормальному закону. При $p < 0,05$ принимается решение о ненормальном многомерном типе распределения.

5. Получим следующие данные:

Kaiser-Meyer-Olkin Test

	MSA
Overall MSA	0.514
Рост, см	0.187
Вес, кг	0.532
Уровень общего холестерина, ммоль/л	0.570
ЛПНП, ммоль/л	0.422
Уровень глюкозы, ммоль/л	0.728

Для использования факторного анализа есть условия, или допущения. В таблице выше представлены результаты теста КМО. Необходимо обратить внимание на значение MSA напротив позиции «Overall MSA». Выборка является адекватной для использования факторного анализа в случае, когда Overall MSA больше 0,5.

В данном случае это условие соблюдается. Следовательно, выборка является адекватной для использования факторного анализа согласно результатам теста Кайзера–Мейера–Олкина.

Результаты критерия Бартлетта представлены в данной таблице:

Bartlett's Test		
χ^2	df	p
134.366	10.000	< .001

Нам необходимо значение последнего столбика – уровня значимости p . При $p < 0,05$ принимается решение, что корреляции отличны от нуля.

Согласно результатам теста Бартлетта, корреляции переменных отличны от нуля, и это позволяет нам корректно использовать и интерпретировать результаты факторного анализа.

Результаты оценки многомерной нормальности с помощью критерия Мардиа отражены в данной таблице:

Mardia's Test of Multivariate Normality				
	Value	Statistic	df	p
Skewness	17.649	88.247	35	< .001
Small Sample Skewness	17.649	100.307	35	< .001
Kurtosis	37.484	0.813		0.416

Note. The statistic for skewness is assumed to be χ^2 distributed and the statistic for kurtosis standard normal.

Результаты данного критерия влияют на выбор метода извлечения факторов. Популярный метод максимального правдоподобия чувствителен к отклонениям от нормального распределения. В данном случае следует выбрать другой метод извлечения, менее зависимый от нормального распределения. Например, метод выделения главной оси (Principal axis method) или взвешенный метод наименьших квадратов (Weighted least squares). Однако большинство критериев чрезмерно «наказывают» за небольшие отклонения от нормального распределения. Для извлечения факторов следует сравнить разные методы для выбора наилучшего решения.

Согласно критерию Мардиа, данные оказались распределены многомерно ненормально:

Chi-squared Test

	Value	df	p
Model	6.021	1	0.014

В таблице факторных нагрузок представлены факторные нагрузки – мера связи между факторами и переменными:

Factor Loadings

	Factor 1	Factor 2	Uniqueness
Уровень общего холестерина, ммоль/л	0.975		0.119
Уровень глюкозы, ммоль/л	0.957		0.088
Вес, кг	0.872		0.025
Рост, см		1.043	-0.011
ЛПНП, ммоль/л		0.415	0.800

Note. Applied rotation method is promax.

На первом этапе в качестве способа выбора числа факторов использовался параллельный анализ, так как согласно результатам было выявлено два фактора:

Factor Characteristics ▼

	Eigenvalues	Unrotated solution			Rotated solution		
		SumSq. Loadings	Proportion var.	Cumulative	SumSq. Loadings	Proportion var.	Cumulative
Factor 1	2.829	2.731	0.546	0.546	2.637	0.527	0.527
Factor 2	1.428	1.261	0.252	0.798	1.343	0.269	0.796

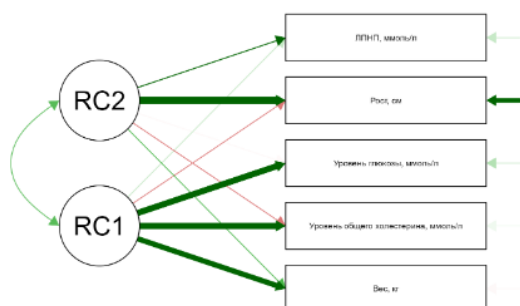
Примечание:

1. (Un) Rotated solution – факторное решение (без) с вращением.
2. SumSq. Loadings – собственное значение фактора.
3. Proportion var. – доля объясненной (общей – для анализа главных компонент) дисперсии.

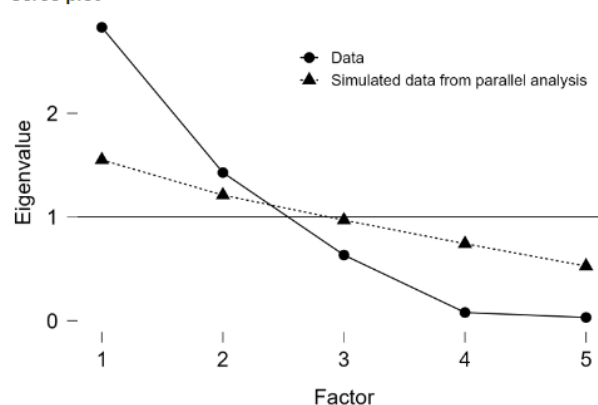
Доля объясненной дисперсии составляет 0,798. Факторы в достаточной степени объясняют дисперсию переменных. Каждый фактор имеет собственное значение, существенно больше 1, что говорит о большой роли каждого выделенного фактора. Однако стоит отметить, что второй фактор не информативный и его стоит исключить. Первый фактор наиболее информативный и его можно интерпретировать как фактор риска сердечно-сосудистых заболеваний. Вероятно, для факторного анализа лучше использовать модель с бóльшим значением Overall MSA, так как у нас это значение было на нижней границе нормы.

График собственных значений факторов представлен на рисунке:

Path Diagram



Scree plot



Дополнительно построена кривая для данных параллельного факторного анализа. Точки (кружки и треугольники) на ломанной линии обозначают выделенные факторы, а расстояния между ними приблизительно отражают степень различий между значениями факторов. Обычно выбирают такое количество факторов, чтобы они имели наибольшее собственное значение.

6. Используем более подходящие данные для факторного анализа (ниже) и загрузим их в JASP:

№	Вес, кг	Систолическое давление	Диастолическое давление	Глюкоза, ммоль/л	Гликозилированный гемоглобин, %	С-пептид, от 0,9 до 7,1 нг/мл
1	74	120	80	4,2	5	3,4
2	75	125	76	4,9	5,5	2,4
3	98	148	100	5,5	5	5
4	75	150	98	5,3	5,9	2,1
5	123	132	97	6,5	6,9	0,4
6	88	133	86	4,6	5,2	5,2
7	115	156	102	7,5	6,9	3,3
8	82	158	108	5,4	5,2	3,4
9	124	148	99	8	8,2	0,3
10	103	128	82	6,4	6	0,8
11	90	160	110	5,3	5,4	2,3
12	102	124	78	10	9,2	0,1
13	82	129	74	4,4	4,2	4,7
14	80	165	112	4,9	5,6	4,3
15	138	118	68	6,9	7,1	0,4
16	99	115	75	6,1	6	6,2
17	118	121	80	9,5	11	0,1
18	125	148	105	8,5	4,4	0,9

19	106	155	109	4,9	5,4	1,2
20	116	115	78	7,2	7,9	0,4
21	143	118	80	10,2	12	0,2
22	102	116	74	8,2	9,1	0,3
23	94	149	105	4,5	5,9	4,2
24	59	110	69	4,8	4	4,2
25	61	118	79	5,6	5,2	2,8
26	99	124	82	7,5	6,3	0,8
27	101	130	80	4,5	5,1	1,9
28	76	140	82	8	7,8	6,1
29	77	138	78	5,6	5,9	5,4
30	118	155	98	5,8	8,1	0,5

7. Выставим настройки для эксплораторного факторного анализа, указанные ранее.

8. Получим следующие данные:

Kaiser-Meyer-Olkin Test

	MSA
Overall MSA	0.662
Вес, кг	0.814
Систолическое давление	0.509
Диастолическое давление	0.489
Глюкоза, ммоль/л	0.743
Гликозилированный гемоглобин, ммоль/л	0.729
С-пептид, нг/мл	0.747

Обратим внимание на значение MSA напротив позиции «Overall MSA». Выборка является адекватной для использования факторного анализа, так как Overall MSA больше 0,5. Полученное значение выше, чем в прошлой модели.

Результаты критерия Бартлетта представлены в таблице:

Bartlett's Test ▼

χ^2	df	p
115.533	15.000	< .001

Нам необходимо значение последнего столбика – уровня значимости p . При $p < 0,05$ принимается решение, что корреляции отличны от нуля.

Результаты оценки нормального распределения данных представлены в таблице:

Mardia's Test of Multivariate Normality

	Value	Statistic	df	p
Skewness	12.926	64.630	56	0.201
Small Sample Skewness	12.926	73.115	56	0.062
Kurtosis	44.670	-0.931		0.352

Note. The statistic for skewness is assumed to be Chi^2 distributed and the statistic for kurtosis standard normal.

В таблице факторных нагрузок представлены факторные нагрузки:

Factor Loadings ▼

	Factor 1	Factor 2	Uniqueness
Глюкоза, ммоль/л	0.840		0.259
Гликозилированный гемоглобин, ммоль/л	0.833		0.278
Вес, кг	0.687	0.420	0.352
C-пептид, нг/мл	-0.650		0.509
Систолическое давление	-0.485	0.781	0.155
Диастолическое давление	-0.413	0.913	-0.003

Note. No rotation method applied.

Доля объясненной дисперсии составляет 0,742:

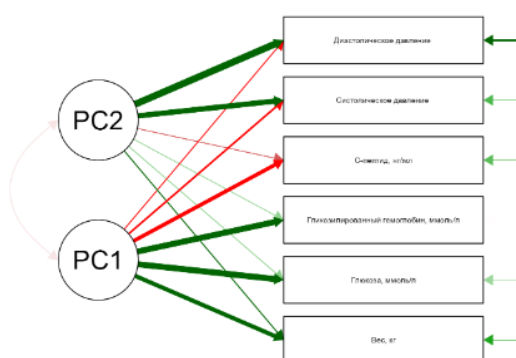
Factor Characteristics

	Eigenvalues	SumSq. Loadings	Proportion var.	Cumulative
Factor 1	3.004	2.699	0.450	0.450
Factor 2	1.869	1.754	0.292	0.742

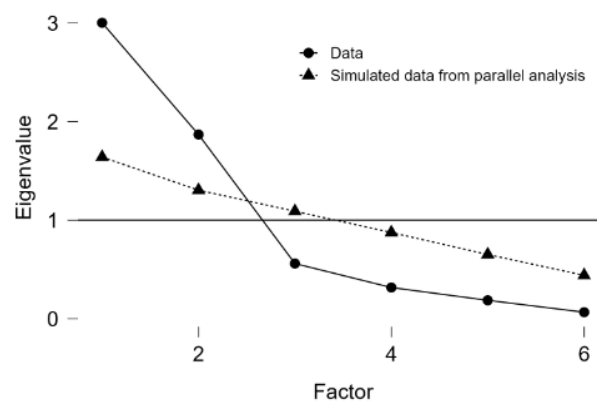
Факторы в достаточной степени объясняют дисперсию переменных. Каждый фактор имеет собственное значение существенно больше 1, что говорит о большой роли каждого выделенного фактора. Первый фактор включает все переменные, второй фактор включает три переменные.

Первый фактор наиболее информативный, и его можно интерпретировать как фактор сердечно-сосудистых патологий, а второй – как фактор диабета:

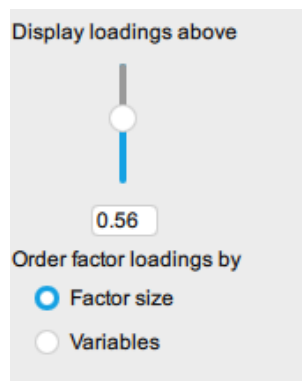
Path Diagram



Scree plot



9. Откорректируем факторные нагрузки. «Display loadings above» позволяет указать пороговое значение факторных нагрузок, которые будут показаны в таблице факторных нагрузок. На рисунке ниже указано значение 0,56:



Соответственно, факторные нагрузки, которые меньше данного значения, не будут отображаться в таблице.

10. Получим следующие факторы:

Factor Loadings				
	Factor 1	Factor 2	Uniqueness	
Глюкоза, ммоль/л	0.840		0.259	
Гликозилированный гемоглобин, ммоль/л	0.833		0.278	
Вес, кг	0.687		0.352	
С-пептид, нг/мл	-0.650		0.509	
Диастолическое давление		0.913	-0.003	
Систолическое давление		0.781	0.155	
Note. No rotation method applied.				
Factor Characteristics				
	Eigenvalues	SumSq. Loadings	Proportion var.	Cumulative
Factor 1	3.004	2.699	0.450	0.450
Factor 2	1.869	1.754	0.292	0.742

Первый фактор можно интерпретировать как сахарный диабет, второй фактор – как сердечно-сосудистые заболевания.

Контрольные вопросы

1. Что такое факторный анализ?
2. Для чего используется факторный анализ?
3. Какие существуют методы факторного анализа?
4. Что такое вращение факторов?
5. Какие существуют методы вращения факторов?
6. Что такое факторная нагрузка?

2.12. Мета-анализ

Краткие сведения из теории

Мета-анализ – это методика, которая используется для объединения результатов нескольких исследований на одну и ту же тему. Основная цель мета-анализа заключается в том, чтобы увеличить статистическую мощность и точность результатов путём объединения данных из нескольких исследований. Это позволяет уменьшить случайную ошибку выборки и улучшить обобщение результатов.

Мета-анализ может быть полезен в различных областях науки, включая медицину, биологию, психологию и другие. Он помогает получить более точные и надёжные выводы, основанные на множестве исследований, а не только на одном. Кроме того, мета-анализ может помочь выявить противоречия между различными исследованиями и определить причины этих противоречий.

Основные понятия мета-анализа:

1. *Систематический обзор* – это процесс поиска, отбора и анализа всех доступных исследований по определенной теме. Цель систематического обзора – найти все соответствующие исследования и оценить их качество.

2. *Эффективность исследуемого фактора* (например, лечения) – это мера того, насколько хорошо работает данный исследуемый фактор. Например, эффективность лечения обычно выражается через различие в исходах между группами лечения и контроля.

3. *Отношение шансов* (Odds Ratio – OR) – это статистический показатель, который используется для сравнения вероятностей двух событий. OR выше 1 указывает на повышенный риск, OR ниже 1 указывает на сниженный риск, а OR, равный 1, указывает на отсутствие разницы в риске.

4. *Доверительный интервал* (CI) – это диапазон значений, в котором с определенной степенью уверенности находится истинное значение показателя. Обычно используются 95 %-е доверительные интервалы.

5. *Гетероскедастичность* – это ситуация, когда дисперсия ошибок в модели меняется в зависимости от предиктора. Это может привести к неправильной оценке стандартных ошибок и доверительных интервалов.

6. *Публикационная предвзятость* – это тенденция публиковать только положительные или значимые результаты исследований, что может исказить представление о действительной эффективности исследуемого фактора (например, эффективности лечения).

Основные правила выполнения мета-анализа:

1. Четкая формулировка вопроса исследования и определение критериев включения (отбора) исследований.

2. Поиск всех соответствующих исследований с использованием систематического подхода.

3. Извлечение данных из каждого исследования с использованием стандартизированных форм.

4. Качественная оценка каждого исследования для определения его надежности и достоверности.

5. Статистическая обработка данных с использованием специальных методов мета-анализа.

6. Интерпретация результатов с учетом потенциальных источников систематических ошибок и ограничений исследования.

7. Документирование всего процесса и результатов мета-анализа для обеспечения прозрачности и воспроизводимости исследования.

Для проведения мета-анализа в JASP необходимо рассчитать *общий размер эффекта* (ES) и *стандартную ошибку* (SE) исследования. ES – это безразмерная оценка (без единиц измерения), которая указывает как направление, так и величину эффекта/результата. Стандартная ошибка измеряет дисперсию различных средних значений выборки, взятых из одной и той же совокупности, и может быть оценена по стандартному отклонению одной выборки.

Некоторые исследования предоставляют эту информацию, хотя многие этого не делают. Однако они дают такие результаты, как:

- центральные тенденции и дисперсия;
- t- или F-статистика;
- p-значения;
- коэффициенты корреляции;
- хи-квадрат.

Все это можно преобразовать в ES, а SE определить с помощью «Калькулятора величины эффекта мета-анализа» Дэвида Уилсона¹.

Например, исследование, сравнивающее результаты лечения с результатами в контрольной группе, может предоставить переменные средние значения, SD и n для каждой группы. На основании этого можно рассчитать стандартизированную среднюю разницу (d) и расчетную дисперсию ошибки (v):

	Mean	SD	N
Treatment	6.25	1.4	10
Control	4.95	1.7	10

Calculate Reset

$d = 0.8348$

95% C.I. = -0.0791 1.7487

$v = 0.2174$

95 %-е доверительные интервалы (CI) можно использовать для расчета SE по следующему уравнению:

$$SE = \frac{\text{верхний предел } CI - \text{нижний предел } CI}{3,92},$$

где $3,92 = 2 \cdot SD = 95 \%$.

Подставляя значения, имеем:

$$SE = \frac{1,7487 - (-0,0791)}{3,92} = 0,466.$$

Более быстрый способ – найти квадратный корень из предполагаемой дисперсии ошибки (v):

$$SE = \sqrt{0,2174} = 0,466.$$

Таким образом, для этого исследования общий $ES = 0,835$ и $SE = 0,466$.

Эти данные можно поместить в MS Excel, а затем использовать для выполнения мета-анализа в JASP. Общий ES, полученный на основе мета-анализа, рассчитывается путем объединения размеров эффекта включенных (отобранных) исследований.

Чтобы интерпретировать результаты мета-анализа, необходимо оперировать неоднородностью данных, моделью, размером эффекта и «лесным графиком».

¹По ссылке: <https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php>

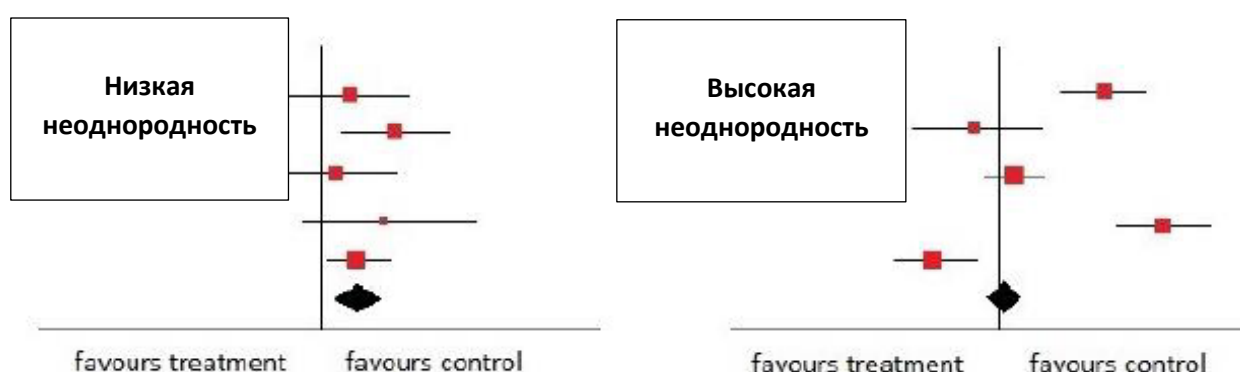
Неоднородность

Гетерогенность описывает любую вариабельность, которая может существовать между различными исследованиями (в противоположность однородности, которая означает, что данные одинаковы). Есть три типа различий:

- клинические: различия в участниках, вмешательствах или результатах;
- методологические: различия в дизайне исследования, риск систематической ошибки;
- статистические: изменение эффектов или результатов вмешательства.

Если вариабельности (различий) нет, то данные можно охарактеризовать как однородные. Мета-анализ занимается статистической неоднородностью. Статистическая неоднородность, используемая для описания изменчивости данных/исследований, и возникает, когда оценки эффекта рассматриваемого лечения, набора данных/исследований различаются между собой. Исследования с методологическими недостатками и небольшие исследования могут переоценивать эффекты лечения и способствовать статистической неоднородности.

На рисунке ниже представлены примеры «лесных участков» с низкой и высокой неоднородностью:



В случае низкой неоднородности все исследования обычно расположены справа от вертикальной оси принятия решения, и их доверительные интервалы перекрываются. При высокой неоднородности исследования распределены по обе стороны от вертикальной линии, и их доверительные интервалы перекрываются незначительно. Кроме визуального наблюдения, мета-анализ использует количественные статистические методы для измерения гетерогенности.

Тесты на гетерогенность

Тест Кокрана Q (Q) имеет низкую статистическую мощность, когда в мета-анализ включено лишь несколько исследований:

- омнибусный тест коэффициентов модели проверяет нулевую гипотезу о том, что все ES равны нулю;
- тест на остаточную гетерогенность проверяет нулевую гипотезу о том, что все размеры эффекта в исследованиях одинаковы (однородность размеров эффекта).

Тау-квадрат (τ^2) представляет собой оценку общей суммы неоднородности. Значение τ^2 интерпретируется как систематические необъяснимые различия между наблюдаемыми эффектами в отдельных исследованиях. На его значение не влияет количество исследований, но зачастую трудно интерпретировать, насколько актуальна эта ценность с практической точки зрения.

I²-тест измеряет степень неоднородности, не вызванную ошибкой выборки. Если существует высокая гетерогенность, можно провести анализ подгрупп. Если гетерогенность очень низкая, нет смысла проводить дальнейший анализ подгрупп. I²-тест не чувствителен к изменениям количества исследований и поэтому широко используется в медицинских и психологических исследованиях, тем более что существует «эмпирическое правило» для его результатов:

- от 0 до 40 %: может быть не важно;
- от 30 до 60 %: может отражать умеренную неоднородность;
- от 50 до 90 %: может отражать значительную неоднородность;
- от 75 до 100 %: значительная неоднородность.

Разница между исследованиями H^2 , определяемая путем приравнивания статистики Q к ее ожидаемому значению, имеет значение 1 в случае однородности, а при гетерогенности $H^2 > 1$.

Модели мета-анализа

В мета-анализе обычно используются две различные модели.

Модель с фиксированными эффектами предполагает, что все исследования имеют одну и ту же истинную величину эффекта, т. е. данные однородны. Это означает, что все факторы, которые могли бы повлиять на величину эффекта, одинаковы во всех исследуемых выборках и, следовательно, незначительны. Различия между исследованиями считаются случайными и поэтому не учитываются в модели. Таким образом, каждое исследование, включенное в мета-анализ, оценивает один и тот же

эффект лечения населения, который теоретически является истинным эффектом лечения населения. Каждое исследование имеет свой вес, причем больший вес придается исследованиям с большими размерами выборки, т. е. с большим объемом информации.

Эта модель отвечает на вопрос: «Какова наилучшая оценка размера популяционного эффекта?».

Модель случайных эффектов предполагает распределение эффекта лечения для некоторых групп населения, т. е. различные исследования оценивают разные, но взаимосвязанные эффекты вмешательства. Следовательно, неоднородность нельзя объяснить случайностью. Эта модель назначает более сбалансированный вес между исследованиями.

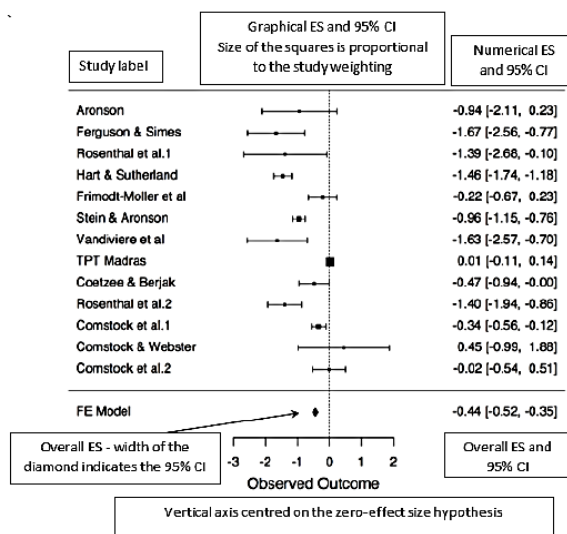
Эта модель отвечает на вопрос: «Каков средний эффект лечения?».

Поэтому важно проверить значительную гетерогенность, чтобы выбрать правильную модель для мета-анализа.

JASP предоставляет 8 оценщиков случайных эффектов для оценки 4 индексов неоднородности, показывающих распределение результатов исследований. Все они используют несколько иной подход, что приводит к разным объединенным оценкам ES и CI. На сегодняшний день наиболее часто используемым оценщиком в медицинских исследованиях является оценщик Дерсимониана–Лэрда, однако недавние исследования показали лучшие оценки дисперсии между исследованиями с использованием методов максимального правдоподобия, эмпирического Байеса и Сидака–Джонкмана.

График «Лесной участок»

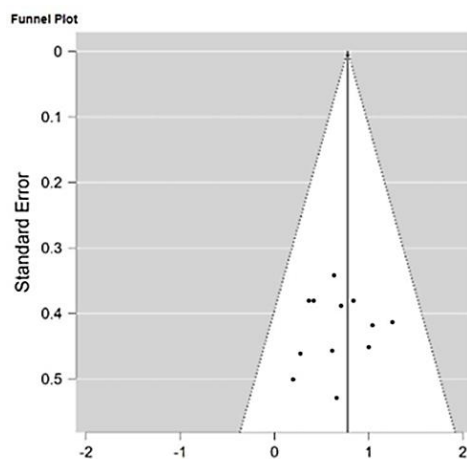
Этот график представляет собой визуальное резюме результатов мета-анализа:



На графике представлены ES и 95 %-е CI для каждого исследования, а также оценка общего ES и 95 %-е CI для всех включенных исследований. Как уже говорилось ранее, «лесной график» можно использовать для визуальной оценки уровня неоднородности.

Обычно, когда все границы погрешностей индивидуального исследования включают итоговый комбинированный стандартный показатель и его 95 %-е доверительные интервалы (ромб), это говорит о низкой гетерогенности (или о высокой однородности). Однако несколько исследований, как показано на рисунке выше, не пересекаются с этим ромбом.

Воронкообразная диаграмма представляет собой график разброса оценок эффекта вмешательства из отдельных исследований в зависимости от их размера или точности:



Согласно этой диаграмме, эффект вмешательства должен возрастать по мере увеличения размера выборки для исследования. Исследования с небольшим эффектом будут располагаться внизу, а исследования с большим приростом будут находиться ближе к вершине. В идеале точки данных должны образовывать симметричное распределение. Асимметрия воронкообразной диаграммы указывает на то, что результаты мета-анализа могут быть необъективными. Возможными причинами такой необъективности являются:

- предвзятость публикаций (тенденция авторов/издателей публиковать только исследования со значительными результатами);
- истинная гетерогенность;
- неправильный выбор меры эффекта (артефакт);
- неточности в данных (ошибки в методологической разработке, анализе данных и т. д.).

Асимметрию воронкообразного графика можно анализировать статистически с помощью теста Ранка или теста Эггерса.

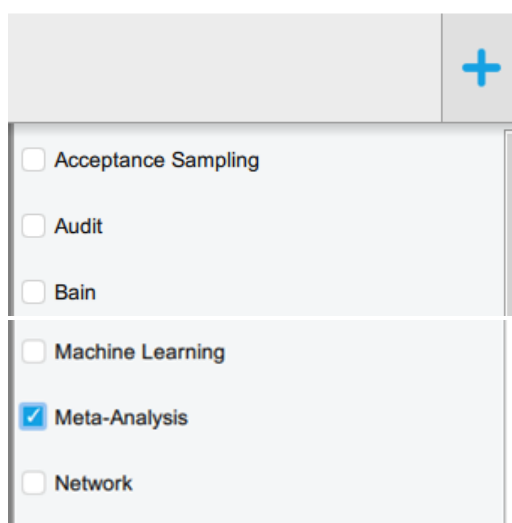
Алгоритм анализа в JASP

1. Загрузим в JASP следующие данные:

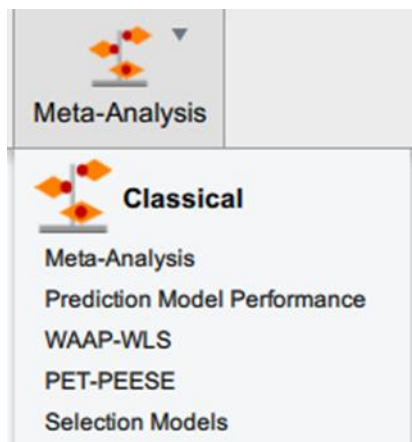
№	Авторы	Год	ES	SEM
1	Бурлаков и др.	2000	–0,2455	0,2132
2	Семенов и др.	2004	–0,4411	0,3043
3	Антонов и др.	2008	–0,5453	0,3351
4	Князева и др.	2009	–0,4764	0,2011
5	Петров и др.	2010	–0,4986	0,2418
6	Мягков и др.	2012	–0,5764	0,4002
7	Степанов и др.	2014	–0,4863	0,3023
8	Глушков и др.	2016	–0,5974	0,3987
9	Черепях и др.	2018	–0,5453	0,5974
10	Гудков и др.	2019	–0,6763	0,6754
11	Дубов и др.	2020	–0,6454	0,4543
12	Гвоздев и др.	2021	–0,7415	0,5643
14	Нянькин и др.	2022	–0,8352	0,5768

В таблице указаны размеры эффекта и стандартная ошибка среднего из 14 исследований изменений показателей мышечной боли с отсроченным началом, также известной как DOMS, в группах, не получавших терапии, по сравнению с группами, получавшими криотерапию. DOMS – это скованность и боль, которые человек чувствует через 24–48 часов после выполнения высокоинтенсивных физических упражнений, к которым его тело еще привыкло.

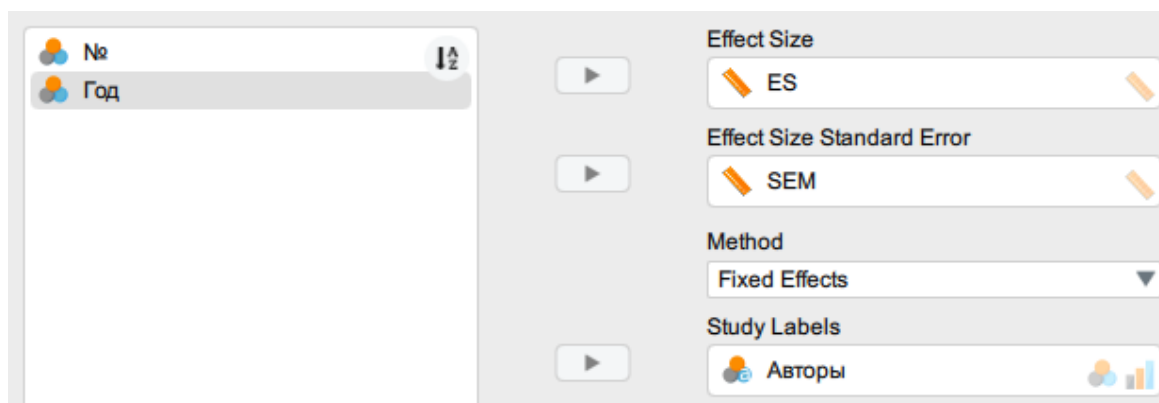
2. Запустим мета-анализ в JASP. Щелкнем крестик в правом верхнем углу JASP и отметим «Meta-Analysis». Это добавит модель мета-анализа на ленту главного меню:



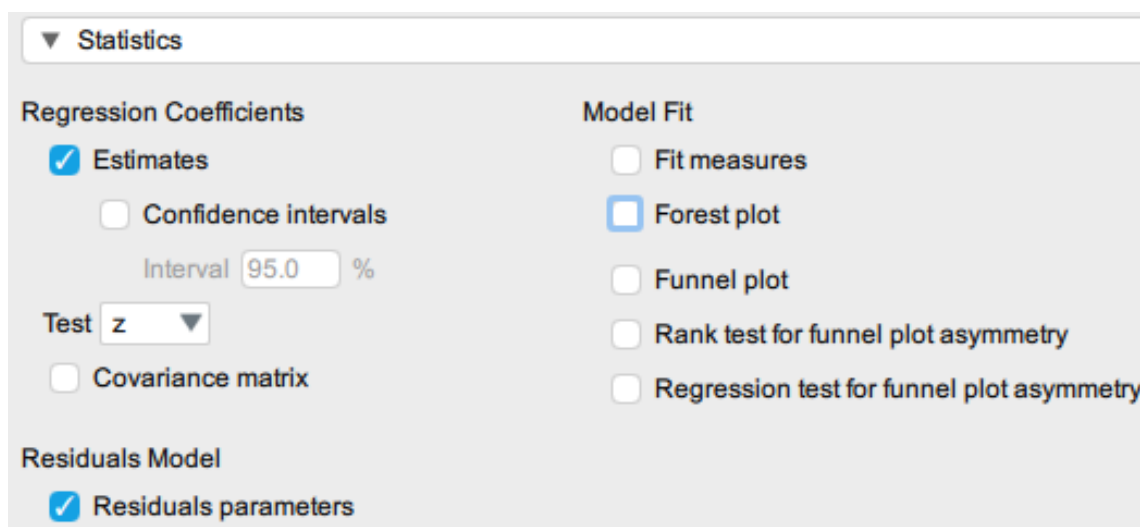
3. Перейдем в раздел «Meta-Analysis» и нажмем «Meta-Analysis» в разделе классического мета-анализа:



4. Перенесем переменные и изменим метод на «Fixed Effects»:



5. Зададим настройки для анализа, как показано на рисунке:



6. Получим следующий результат:

Fixed and Random Effects				
	Q	df	p	
Omnibus test of Model Coefficients	28.286	1	< .001	
Test of Residual Heterogeneity	2.238	12	0.999	
Note. p-values are approximate.				
Note. The model was estimated using Fixed Effects method.				
Coefficients				
	Estimate	Standard Error	z	p
intercept	-0.480	0.090	-5.318	< .001
Note. Wald test.				

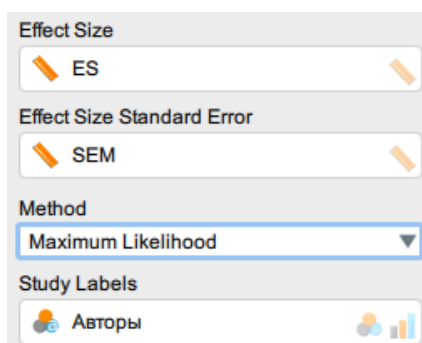
Значимость ES показана в таблице «Fixed and Random Effects» («Фиксированные и случайные эффекты»).

Омнибусный тест коэффициентов модели проверяет нулевую гипотезу о том, что все коэффициенты во второй таблице равны нулю, т. е. вмешательства не имеют существенного эффекта. В данном случае это существенно, и нулевую гипотезу можно отвергнуть. Таким образом, вмешательство здесь имеет значительный эффект.

Тест на остаточную гетерогенность проверяет нулевую гипотезу о том, что все величины эффекта во всех исследованиях равны. Если эти величины незначительны, следует использовать модель с фиксированными эффектами. Если они существенны, то более подходящей будет модель случайных эффектов. В данном случае эти величины незначимы.

Таблица коэффициентов «Coefficients» дает оценку (критерий Вальда) общего размера эффекта комбинированного исследования и его значимости. Это подтверждает результат омнибусного теста в таблице фиксированных и случайных эффектов.

7. Используем для ознакомления модель случайных эффектов, которая не подходит для этого набора данных. Вернемся к параметрам и изменим «Method» на «Maximum Likelihood»:



Effect Size

ES

Effect Size Standard Error

SEM

Method

Maximum Likelihood

Study Labels

Авторы

В результате будет добавлена еще одна таблица, показывающая различные оценки остаточной однородности:

Residual Heterogeneity Estimates ▼	
Estimate	
τ^2	0.000
τ	0.000
I^2 (%)	0.000
H^2	1.000

И τ^2 , и I^2 не демонстрируют чрезмерную дисперсию (гетерогенность) между исследованиями. Это еще раз подтверждает, что модель случайных эффектов не подходит для этих данных.

8. Вернемся к параметрам статистики и при подборе модели отметим «Forest Plot», «Funnel Plot» и «Rank Test for Funnel Plot Asymmetry»:

▼ Statistics

Regression Coefficients

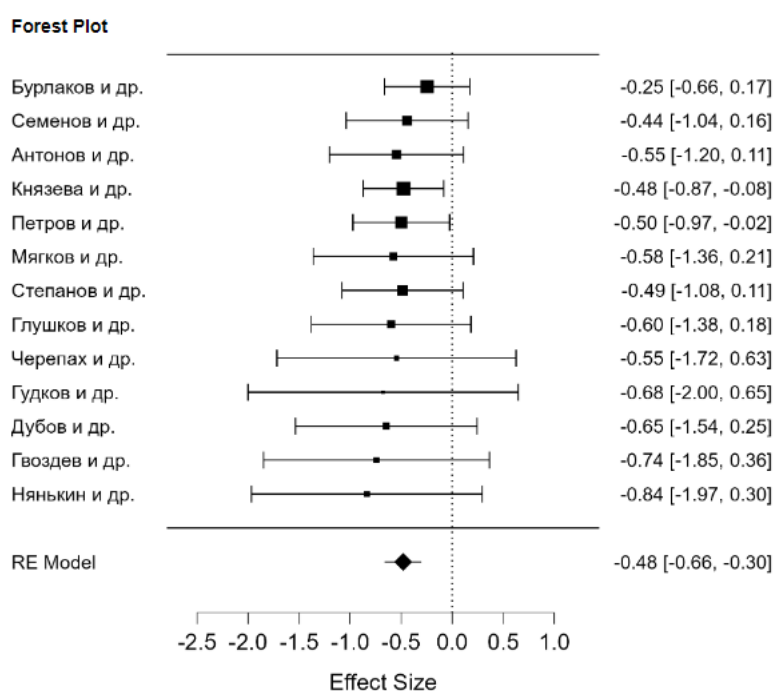
☒ Estimates
☐ Confidence intervals
Interval %
Test
☐ Covariance matrix

Model Fit

☐ Fit measures
☒ Forest plot
☒ Funnel plot
☒ Rank test for funnel plot asymmetry
☐ Regression test for funnel plot asymmetry

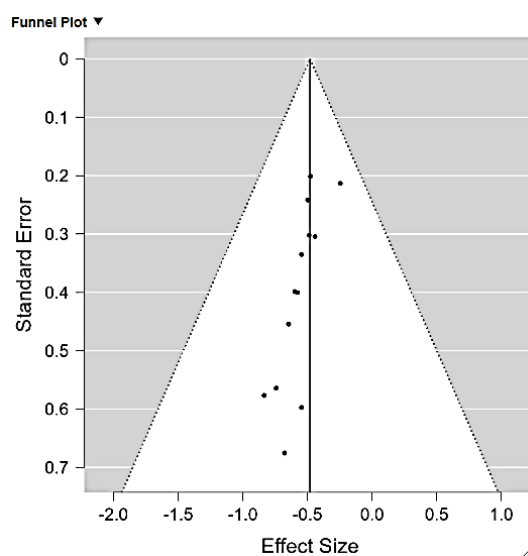
Residuals Model
☒ Residuals parameters

9. Получим следующие графики:



На этом графике показаны взвешенные размеры эффекта (размер квадратов отражает вес каждого исследования) и CI, использованные для определения комбинированного ES (ромб). Легко видеть, что общий эффект криотерапии оказывает значительное влияние на показатели DOMS ($ES = -0,48$), снижая болевую симптоматику.

Воронкообразный график показывает, что наблюдаемые размеры эффектов не симметрично распределены вокруг вертикальной оси (на основе общей оценки размера эффекта, в данном случае $-0,48$), но лежат в пределах доверительного треугольника 95 %:



Асимметрия часто считается показателем предвзятости публикации.

График ниже сопровождается «Тестом ранговой корреляции» для выявления асимметрии воронкообразного графика, которая в данном случае значительна ($p = 0,004$), что указывает на необходимость подбора дополнительных литературных источников, используя методы «слепого включения»:

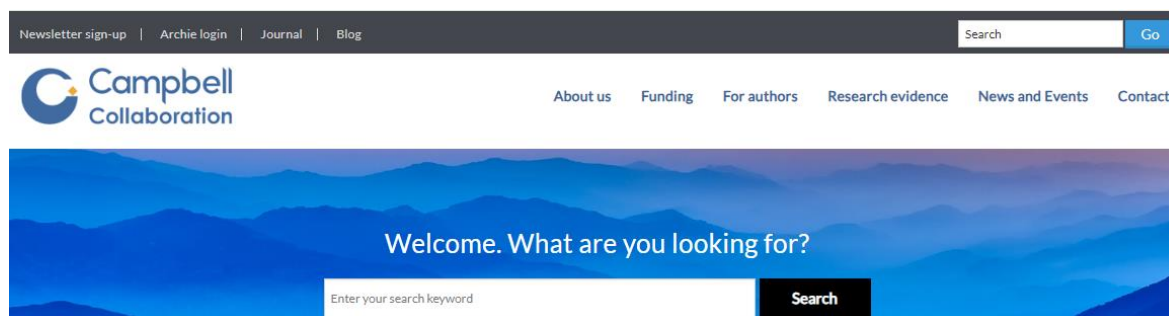
Rank correlation test for Funnel plot asymmetry

	Kendall's τ	p
Rank test	-0.590	0.004

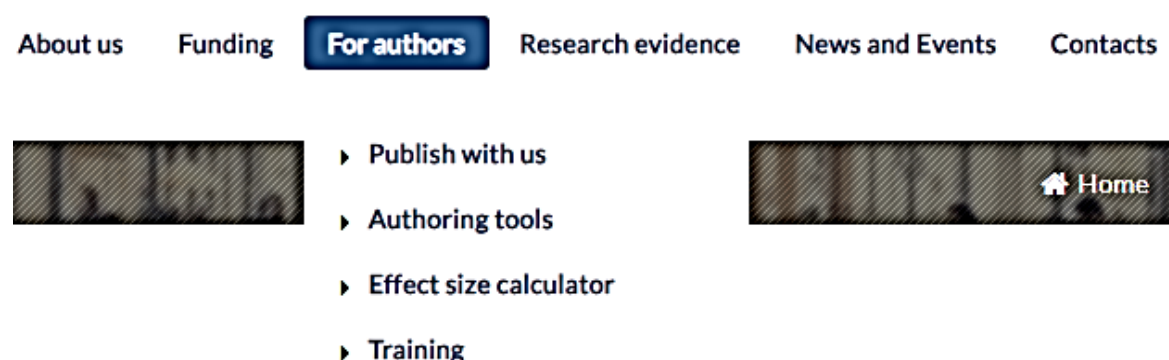
Далее рассмотрим, как получить доступ к «Калькулятору величины эффекта мета-анализа» Дэвида Уилсона² и как его использовать для расчета ES и SE.

² По ссылке: <https://campbellcollaboration.org/>

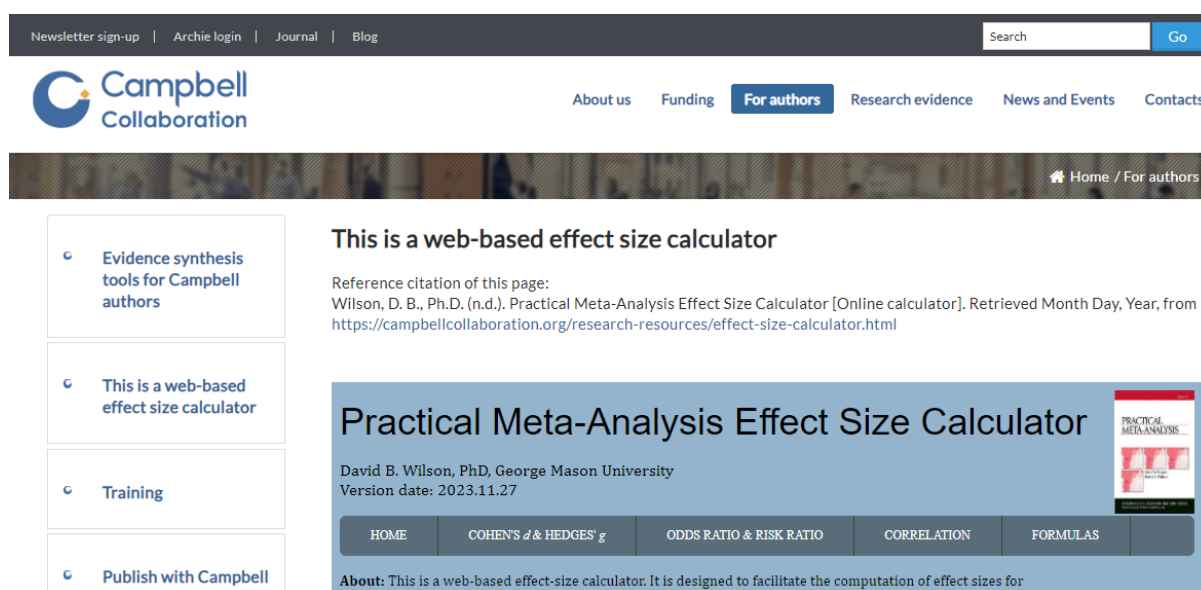
1. Перейдя по ссылке <https://campbellcollaboration.org/>, мы попадем на сайт:



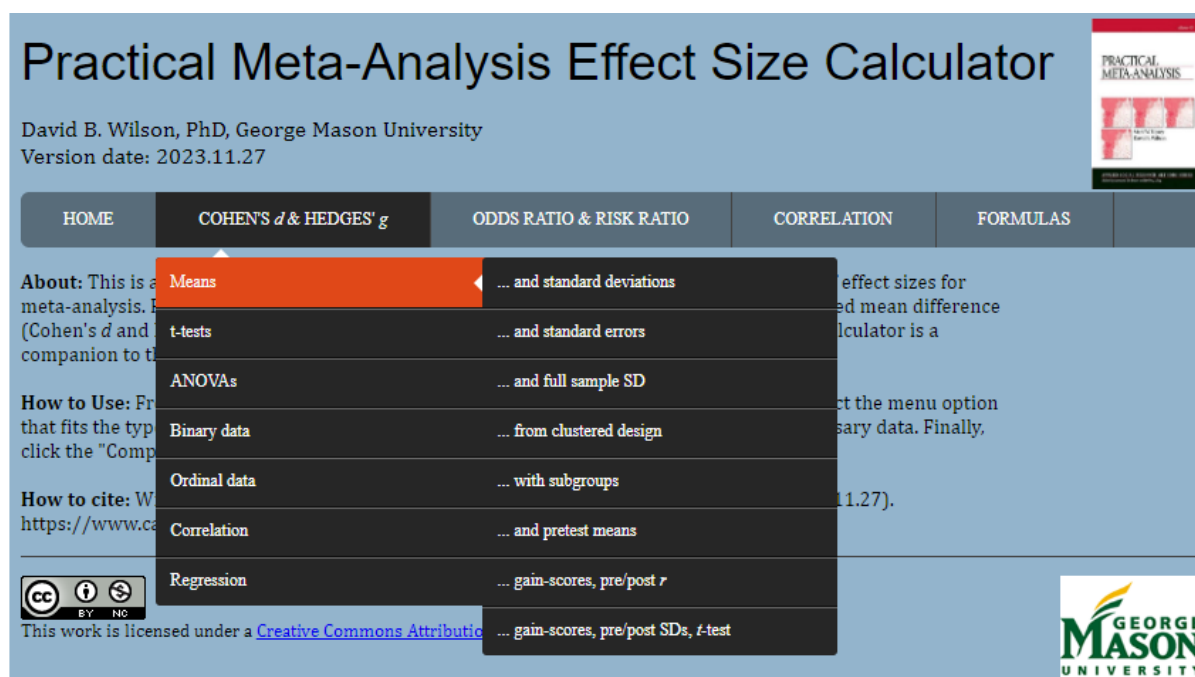
2. Выберем раздел «For authors». В выпадающем списке отметим «Effect size calculator»:



3. Дождемся окончательной загрузки страницы. Откроется следующее окно:



4. Перейдем в раздел «COHERNS d & HEDGES g », выберем «Means» и в выпадающем списке отметим первый пункт «Standard deviation»:



5. Откроется следующий калькулятор:

Предположим, у нас есть следующие данные:

№	Groups	Mean, %	SD	N	Годы
1	Treatment	40	10	30	2000
	Control	100	20	30	
2	Treatment	60	17	45	2003
	Control	98	20	45	
3	Treatment	70	16	100	2005
	Control	99	25	100	
4	Treatment	50	19	300	2007

	Control	98	10	300	
5	Treatment	30	12	150	2009
	Control	95	22	150	
6	Treatment	55	10	1500	2011
	Control	98	23	1500	
7	Treatment	70	20	350	2013
	Control	98	22	350	
8	Treatment	39	14	2000	2015
	Control	97	17	2000	
9	Treatment	65	13	400	2017
	Control	97	15	400	
10	Treatment	57	13	60	2019
	Control	100	17	60	
11	Treatment	38	14	800	2021
	Control	98	18	800	
12	Treatment	41	8	250	2023
	Control	98	14	250	

Оценивалось влияние экспериментального антибиотика на выживаемость у мышей, больных бактериальной пневмонией. Соответственно, имеются две группы: контрольная, которая не получала антибактериальную терапию, и экспериментальная группа, которой вводили антибиотик. Среднее значение представлено в процентах умерших мышей. Под номерами представлены исследования разных ученых в разные годы. Необходимо найти стандартную ошибку (SE) и размер эффекта (ES).

Найдем SEM для первой группы исследуемых. Для этого введем данные в онлайн-калькулятор:

	Mean	SD	N
Treatment	40	10	30
Control	100	20	30
Calculate Reset			
Cohen's d =	-3.7947		Hedges' g = -3.7455
95% C.I. =	-4.6415 -2.9479		95% C.I. = -4.5813 -2.9096
v =	0.1867		v = 0.1818
se =	0.432		se = 0.4264

В данном случае мы получили: $ES = -3,7947$, а $SE = 0,432$.

Аналогичным образом производится расчет для всех остальных значений, данные загружаются в JASP, проводится мета-анализ и делаются выводы.

Контрольные вопросы

1. Что такое мета-анализ?
2. В каких случаях применяется мета-анализ?
3. Какие преимущества у мета-анализа перед традиционными статистическими методами?
4. Как оценивается качество исследований при проведении мета-анализа?
5. Каковы основные этапы проведения мета-анализа?
6. Что такое фиксированное и случайное моделирование в мета-анализе?
7. Что такое «публикационная предвзятость» и как она влияет на результаты мета-анализа?
8. Как оценить гетерогенность исследований в мета-анализе?
9. Какие существуют ограничения у мета-анализа?
10. Какие существуют методы визуализации данных в мета-анализе?
11. Какие существуют программные продукты для проведения мета-анализа?
12. Какие существуют области применения мета-анализа?

Рекомендуемая литература

1. Жученко Ю.М., Ковалев А.А., Игнатенко В.А. Основы статистики. Лабораторный практикум. Гомель: ГомГМУ, 2016. 176 с.
2. Кричевец А.Н., Корнеев А.А., Рассказова Е.И. Основы статистики для психологов. Москва: Акрополь, 2019. 286 с.
3. Платонов А.Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы. Москва: РАМН, 2000. 52 с.
4. Работа пользователя в Microsoft Excel 2010 / Т.В. Зудилова [и др.]. Санкт-Петербург: НИУ ИТМО, 2012. 87 с.
5. Стариченко Б.Е. Обработка и представление данных педагогических исследований с помощью компьютера. Екатеринбург: УрГПУ, 2004. 218 с.
6. Glantz, Stanton A. Primer of Biostatistics. 7th ed. New York: McGraw Hill, 2012. 312 p.
7. Goss-Sampson M. Statistical analysis in JASP: A guide for students. 2019. 168 p. DOI: 10.6084/m9.figshare.9980744
8. Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry. Learning Statistics with JASP: A Tutorial for Psychology Students and Other Beginners. 2019. 418 p. DOI: 10.24384/hgc3-7p15.
9. Howitt D. Cramer D. Introduction to Statistics in Psychology. L.: Financial Times, 2008.
10. Navarro D and Foxcroft D. learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.75). 2022. 504 p. DOI: 10.24384/hgc3-7p15

Оглавление

Предисловие.....	3
1. ОСВОЕНИЕ ПРОГРАММНОЙ СРЕДЫ MS EXCEL И JASP (ВВОДНАЯ ЧАСТЬ).....	4
1.1. Первоначальное знакомство с MS Excel.....	4
1.2. Методы оформления таблиц.....	10
1.3. Визуализация данных.....	12
1.4. Методы обработки данных.....	15
1.5. Создание связей между таблицами и консолидация данных	18
1.6. Знакомство с программой для статистического анализа данных JASP.....	21
2. БИОСТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В MS EXCEL И JASP (ОСНОВНАЯ ЧАСТЬ).....	44
2.1. Полигон частот.....	44
2.2. Описательная статистика.....	67
2.3. Дисперсионный анализ.....	90
2.4. Критерий Стьюдента.....	113
2.5. Корреляционный и регрессионный анализ.....	128
2.6. Криволинейная регрессия.....	145
2.7. Логистическая регрессия.....	157
2.8. Анализ качественных признаков.....	171
2.9. Непараметрические критерии для анализа количественных признаков.....	187
2.10. Кластерный и дискриминантный анализы.....	217
2.11. Факторный анализ.....	243
2.12. Мета-анализ.....	263
Рекомендуемая литература.....	279

Учебное издание

Родькин Станислав Владимирович

БИОСТАТИСТИКА

Редактор Г.В. Владимирова
Компьютерная обработка: О.И. Пушкина

В печать 20.11.2024
Формат 60×84/16. Объем 17,6 усл. п. л.
Тираж 100 экз. Заказ № 1213. Цена свободная

Издательский центр ДГТУ
Адрес университета и полиграфического предприятия:
344003, г. Ростов-на-Дону, пл. Гагарина, 1