Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра информационных технологий и управления бизнесом

АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ

Методические рекомендации

Витебск
ВГУ имени П.М. Машерова
2024

УДК 004.6:61:57(076.5)
ББК 22.161я73+5с517я73+28с517я73
А64

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 2 от 20.12.2023.

Составители: старший преподаватель кафедры информационных технологий и управления бизнесом ВГУ имени П.М. Машерова **Н.В. Булгакова;** доцент кафедры информационных технологий и управления бизнесом ВГУ имени П.М. Машерова, кандидат биологических наук, доцент **А.А. Чиркина**

Р е ц е н з е н т ы :
заведующий кафедрой прикладного и системного программирования
ВГУ имени П.М. Машерова,
кандидат физико-математических наук, доцент *Е.А. Корчевская*;
доцент кафедры логистики и менеджмента
Витебского филиала Международного университета «МИТСО»,
кандидат физико-математических наук, доцент *А.М. Воронов*

А64 **Анализ медико-биологических данных :** методические рекомендации / сост.: Н.В. Булгакова, А.А. Чиркина. – Витебск : ВГУ имени П.М. Машерова, 2024. – 39 с.

Методические рекомендации разработаны для изучения учебной дисциплины «Анализ медико-биологических данных» и ориентированы на поддержку лабораторных занятий. Содержат краткие теоретические сведения и указания к выполнению лабораторных работ обучающимися. Предназначены для студентов специальности 1-40 05 01-07 Информационные системы и технологии (в здравоохранении) факультета математики и информационных технологий.

УДК 004.6:61:57(076.5)
ББК 22.161я73+5с517я73+28с517я73

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Применение информационных технологий для анализа медико-биологических данных является актуальным научным направлением. На основе проведенного анализа данных выявляются закономерности, выполняются прогнозы, оптимизируется лечение, разрабатываются новые методики и технологии в здравоохранении. Машинное обучение и искусственный интеллект в этой сфере позволяют улучшить диагностику заболеваний, проводить исследования более эффективно.

Статистический анализ медико-биологических данных включает в себя несколько основных аспектов:

1. Дескриптивный анализ: описательная статистика позволяет получить характеристики и систематизировать данные, выявить ключевые закономерности и паттерны, а также подготовить данные для дальнейшего анализа.
2. Инференциальный анализ: на основе статистических методов делается вывод о генерализации результатов на всю совокупность изучаемых объектов (например, пациентов, клеток и т.д.). Этот подход включает в себя использование различных статистических методов, таких как доверительные интервалы, тесты гипотез, анализ дисперсии и регрессионный анализ, для того чтобы делать заключения о связях и различиях между группами, а также для определения степени уверенности в этих выводах.
3. Корреляционный анализ: позволяет определить степень взаимосвязи между различными переменными, что может быть важно для выявления факторов риска и предикторов заболеваний. Данный вид анализа может быть полезен для понимания связи между биологическими показателями, заболеваниями, факторами риска и т.д. В медицинских и биологических исследованиях корреляционный анализ помогает выявлять связи между различными переменными, что может указывать на возможные факторы риска или причины заболеваний, а также помогает предсказывать тенденции и результаты.
4. Регрессионный анализ: дает возможность предсказывать значения одной переменной на основе других переменных, что полезно для прогнозирования и моделирования процессов в медицинской и биологической сферах. Этот метод статистического анализа используется для изучения связи между одной или несколькими независимыми переменными и зависимой переменной, которая может быть биологическим показателем, заболеванием, физиологическим параметром и т.д. Регрессионный анализ может помочь определить, как независимые переменные влияют на зависимую переменную и позволяет делать предсказания на основе этих влияний.

Указанные аспекты статистического анализа помогают извлечь ценные знания из медико-биологических данных и принять обоснованные решения, а также являются основополагающими для доказательной медицины.

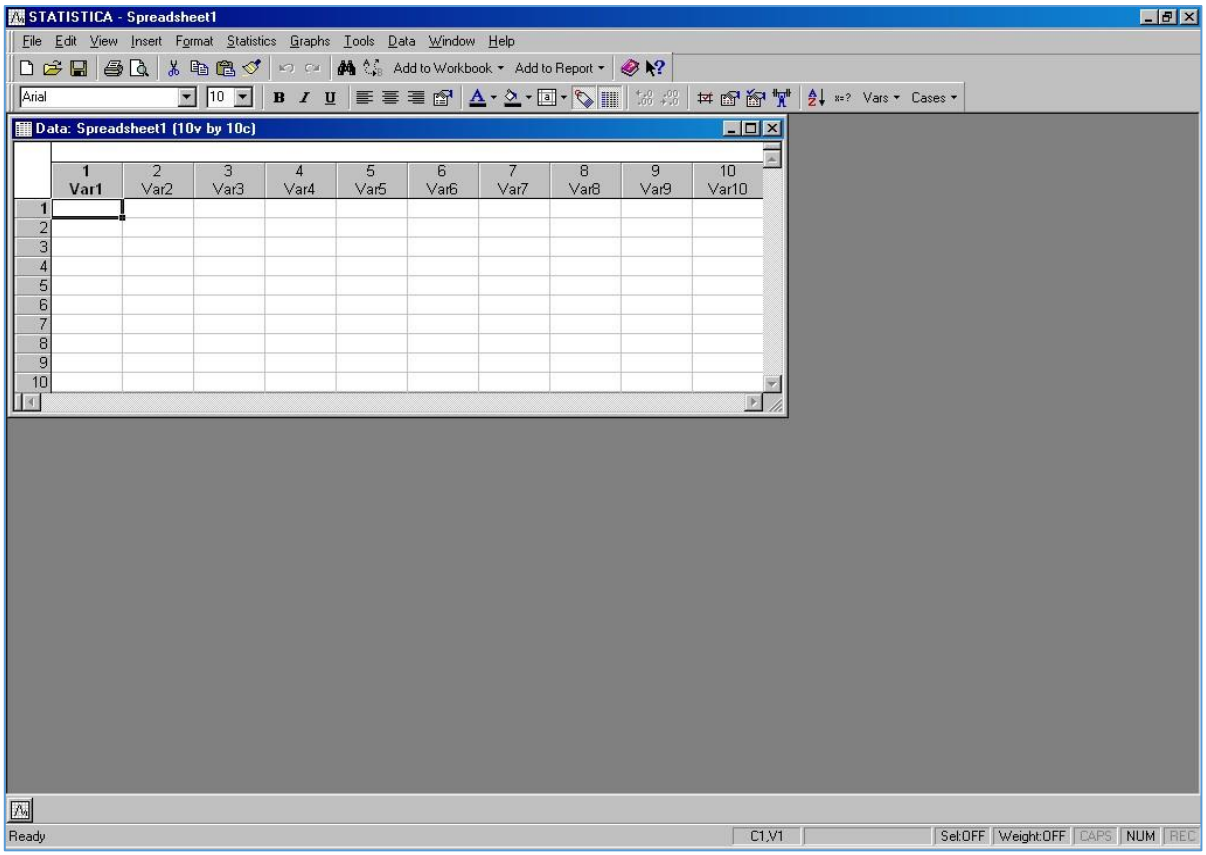
Материалы, вошедшие в методические рекомендации «Анализ медико-биологических данных», предназначены для выполнения лабораторных работ студентами студентов специальности 1-40 05 01-07 Информационные системы и технологии (в здравоохранении). Авторы пытались решить две задачи:

1. Определить базовые понятия теории вероятностей и статистики, необходимые для анализа медико-биологических данных.
2. Рассмотреть этапы проведения анализа на конкретных примерах с помощью пакета STATISTICA таким образом, чтобы студент, используя эти методические рекомендации и определив цель исследования, всегда мог самостоятельно получить нужный результат.

В издании рассмотрены далеко не все используемые сегодня на практике статистические методы. Мы ограничиваемся следующими темами: «Знакомство с пакетом STATISTICA», «Предварительный анализ данных. Представление данных в исследованиях», «Универиантный анализ. Решение задач дескриптивной статистики в пакете STATISTICA», «Сравнение с пороговым значением. Использование интервальной оценки параметров генеральной совокупности», «Бивариантный анализ. Взаимосвязь двух переменных». Именно они прежде всего необходимы при анализе медико-биологических данных. Все файлы с необходимой для работы исходной информацией размещены на LMS Moodle: https://newsdo.vsu.by/mod/assign/view.php?id=667530 в материалах курса «Анализ медико-биологических данных».

ТЕМА 1. ЗНАКОМСТВО С ПАКЕТОМ STATISTICA

Пакет *STATISTICA* состоит из статистических модулей для анализа и обработки данных, которые в свою очередь состоят из статистических процедур. Управление программой максимально приближено к семейству Windows-приложений. Выбор необходимого модуля осуществляется с помощью меню **Statistics**, либо с помощью кнопки в левом нижнем углу окна:



Основным является рабочее окно, в котором вводятся исходные данные и выводятся результаты их статистической обработки в табличном или графическом виде.

Ввод данных осуществляется в табличном виде.

Набор данных в пакете STATISTICA – это прямоугольная таблица, столбцам которой соответствуют обрабатываемые *переменные* (**Variables**), а строкам отвечают *наблюдения* (**Cases**) значений переменных. Для создания нового набора данных нужно, прежде всего, завести файл с *трафаретом* таблицы нужных размеров.

Сделать это можно так: по меню *File –New Data…* через раскрывшееся диалоговое окно нужно завести новый файл с расширением *.sta*. В строке для заголовка можно дать комментарий к содержимому набора данных (для входа в строку заголовка достаточно дважды кликнуть на ней левой кнопкой мыши). В результате открытия нового файла в окне пакета появляется

начальный трафарет создаваемого набора данных с исходными размерами в 10 переменных на 10 наблюдений. Реально нужное количество переменных и наблюдений выставляется после этого у трафарета по меню инструментальных кнопок *Vars* и *Cases*. Как наблюдениям, так и переменным в трафарете создаваемого набора данных можно дать содержательные названия по меню *Cases – Case Name Manager*.

Названные действия по определению переменных могут быть проделаны из основного окна с трафаретом набора данных по меню *Vars –All Specs*. В результате появляется окно со списком установленных по умолчанию атрибутов переменных, которые можно поправить и дополнить с клавиатуры. При этом особенно тщательно нужно определить формат каждой переменной. По умолчанию он есть числовой с размерами “8.3” (т.е. с фиксированной точкой, где под все значащие цифры, знак числа и десятичную точку отведено 8 символов, 3 из которых предназначены для дробной части). Сменить и детализировать формат отдельной переменной можно в диалоговом окне, которое раскрывается, если дважды кликнуть левой кнопкой мыши на нужной переменной в трафарете. Это же окно раскрывается и по меню *Vars – Specs*.

Что касается имён переменных, то их лучше всегда давать содержательными (а не абстрактными VAR1, VAR2 и т.д.). Также немаловажным является то, что в электронной таблице не допускаются пропущенные (незаполненные) строки, т.к. программа будет расценивать их как ячейки, заполненные нулями, соответственно при выполнении какой-либо процедуры будут получены искаженные результаты. Поэтому кроме имени (**Name**) для каждой переменной надо указать так называемый *код пропущенного значения* (**MD Code**). По умолчанию этот код есть “–9999”, и он отмечает в памяти для процедур обработки пакета, что, на самом деле, на его месте (в определенной клетке трафарета) реального значения нет. А изображается пропущенное значение на экране в наборе данных пробелом. Из обязательных атрибутов переменной надо указать тип и формат её значений. Тип определяет, будет ли переменная числовой, текстовой, датой, временем и проч., а формат (**Format**) описывает размеры значений переменной. Значениям переменной можно также дать развернутый содержательный комментарий (**Long Name**).

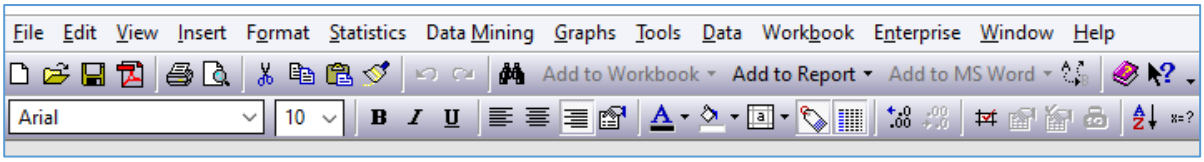
Файлы проекта

При работе с пакетом *STATISTICA* образуется пять различных типов документов: рабочая книга (*Workbook*), рабочий лист – мультимедийная таблица (*Spreadsheet*), отчет (*Report*), графическую область (*Graph)* и макрокоманда (*Macros*) для языка *STATISTICA* Visual Basic. Рабочая книга представляет собой упорядоченный вывод данных, объединяя в себе рабочие листы и графики. Каждый документ представляет собой таблицу. Файл рабочей книги имеет расширение **.stw*. Рабочие листы пакета *STATISTICA* предназначены для ввода данных в числовой или текстовой форме имеет расширение **.sta*. Форматом рабочего листа является двумерная таблица с неограниченным

количеством наблюдений (строк) и переменных (столбцов), каждый их которых содержит неограниченное количество символов. Отчеты *STATISTICA* позволяют организовать вывод данных в текстовом формате, более удобном для вывода документов на печать. По умолчанию файл отчета имеет расширение **.str*, но существует возможность преобразования отчета в стандартный файл формата RTF. Вся графическая обработка данных сохраняется в отдельные файлы с расширением **.stg*. При этом поддерживается внедрение графических объектов из других программ. Макрокоманда представляют собой программный код, написанный на языке Visual Basic. Каждый из описанных компонентов проекта отображается в отдельном окне и имеет свою пиктограмму в дереве проекта на панели экрана слева.

Главное меню

В верхней части рабочего окна пакета (как и в любом Windows-приложение) расположено главное меню:



Как видим из рисунка, пункты из панели главного меню: File (Файл), Edit (Правка), View (Вид), Insert (Вставка), Format (Формат), Tools (Сервис), Windows (Окно) и Help (Справка) по своей функциональной принадлежности являются стандартными для Windows-приложений. Специфическими пунктами меню можно считать следующие:

Statistics (Вычисления) – данный пункт меню содержит огромное количество методов статистической обработки данных начиная от расчета описательных статистик (максимум, минимум, средняя и т.д.) до сложнейших многомерных статистических алгоритмов.

Graphs (Графики) – в данном пункте доступны огромное количество разнообразных графиков и диаграмм, как двухмерных так и трехмерных.

Date (Данные) – в данном меню доступны алгоритмы направленные на преобразование имеющихся данных (стандартизация, ранжирование и т.д.).

Создание таблицы исходных данных

Создать файл данных в пакете STATISTICA можно двумя способами:

1) импортировать готовые данные из других программ (баз данных, математических и статистических пакетов прикладных программ);

2) Ввести необходимую информацию с клавиатуры.

Для создания файла данных первым способом введем в табличном редакторе Excel следующие данные:

	A	B
1	**Month**	**Data**
2	Январь	215
3	Февраль	352
4	Март	215
5	Апрель	254
6	Май	222
7	Июнь	212

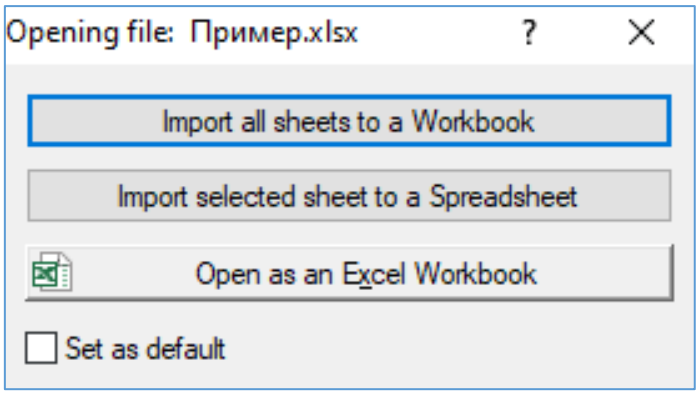
Далее необходимо сохранить файл и закрыть табличный редактор Excel.

Для импорта файла с данными в пакет STATISTICA необходимо пройти следующие шаги:

Шаг 1. В главном меню пакета выберем *File* → *Open* (Файл Открыть).

Шаг 2. В появившемся окне необходимо выбрать тип файла, в данном случае файл электронной таблицы Excel (т.е. необходимо выбрать расширение *.xlsx) и имя искомого файла, далее нажать кнопку Открыть.

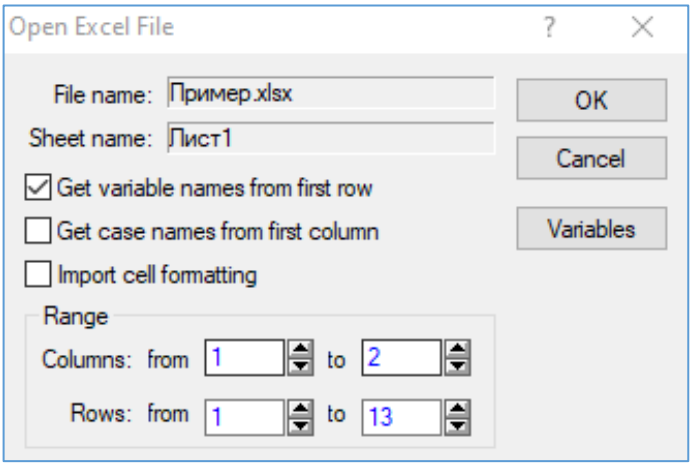
Шаг 3. В открывшемся диалоговом окне будет предложено импортировать отдельную страницу или все страницы рабочей книги.



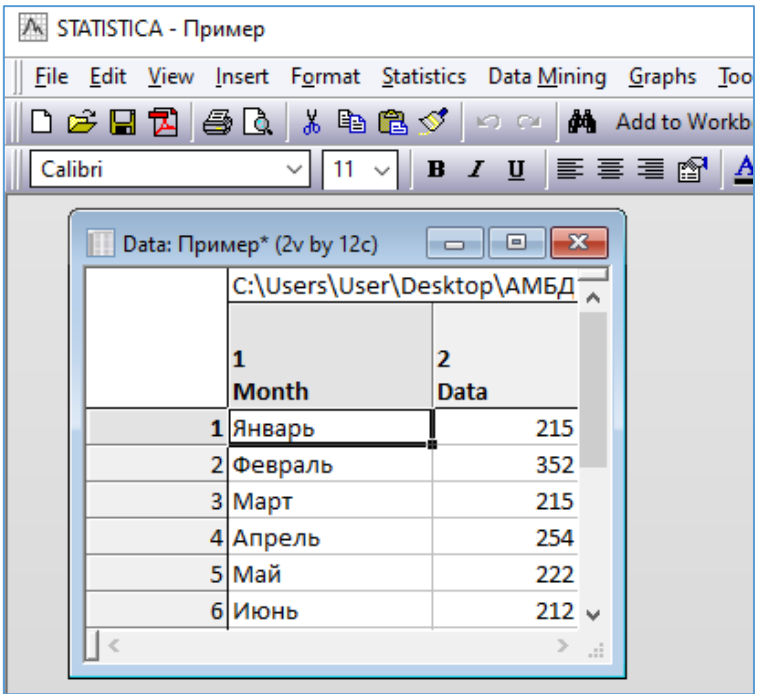
где: Import all sheets to a Workbook – Импорт всех страниц рабочей книги Import selected sheets to a Spreadsheet – Импорт выбранных страниц рабочей книги. В нашем случае выберем второй вариант и перейдем к следующему шагу.

Шаг 4. В появившемся диалоговом окне Select Sheet to Import (Выбор импортируемой страницы) выберем необходимую страницу и нажмем клавишу ОК.

Шаг 5. В следующем окне будет предложено указать размер таблицы, а также предоставлена возможность оставить имеющиеся имена импортируемых переменных и имена записей.



Установить диапазон импортируемых данных – Range; имена наблюдений из первого столбца – Get cases names from first column; имена переменных из первой строки – Get variable names from first row; импорт формата ячеек – Import cell formatting. Результат:



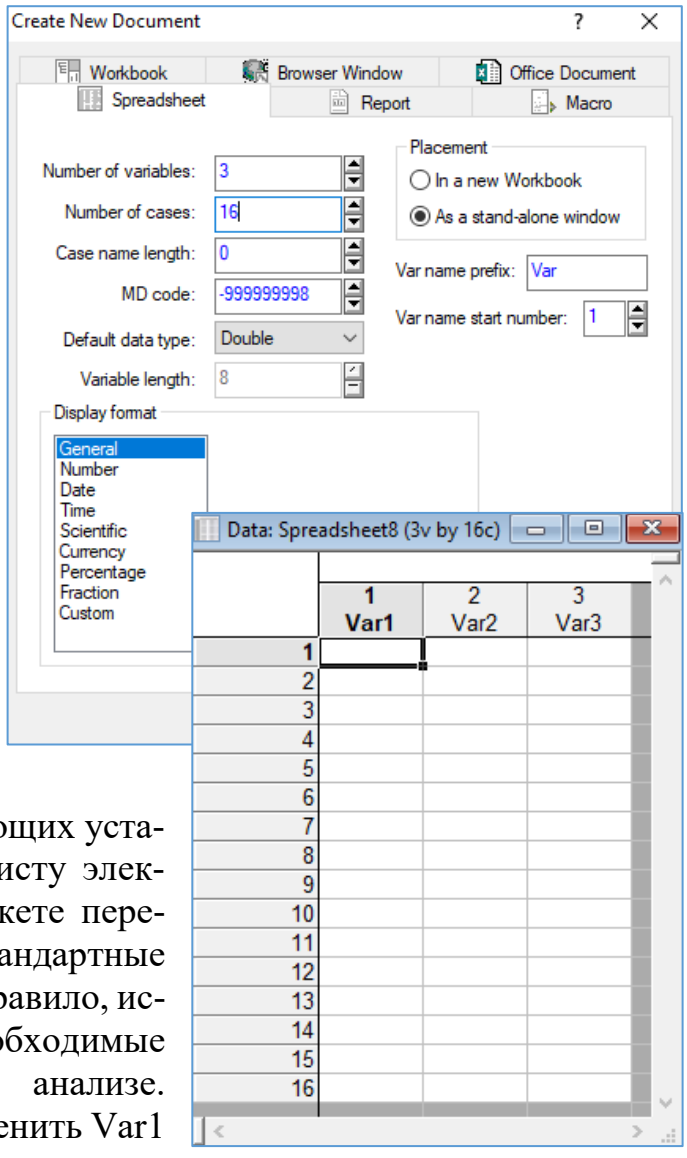
Ручной ввод информации

Шаг 1. После запуска программы STATISTICA в главном меню нужно выбрать *File New* (Файл Новый). В появившемся окне *Create New Document* (Создание нового документа), необходимо ввести количество переменных *Number of variables* и объем совокупности *Number of cases*.

В данном случае исходная матрица данных 3×16, где:

In a new Workbook – Создание новой рабочей книги,

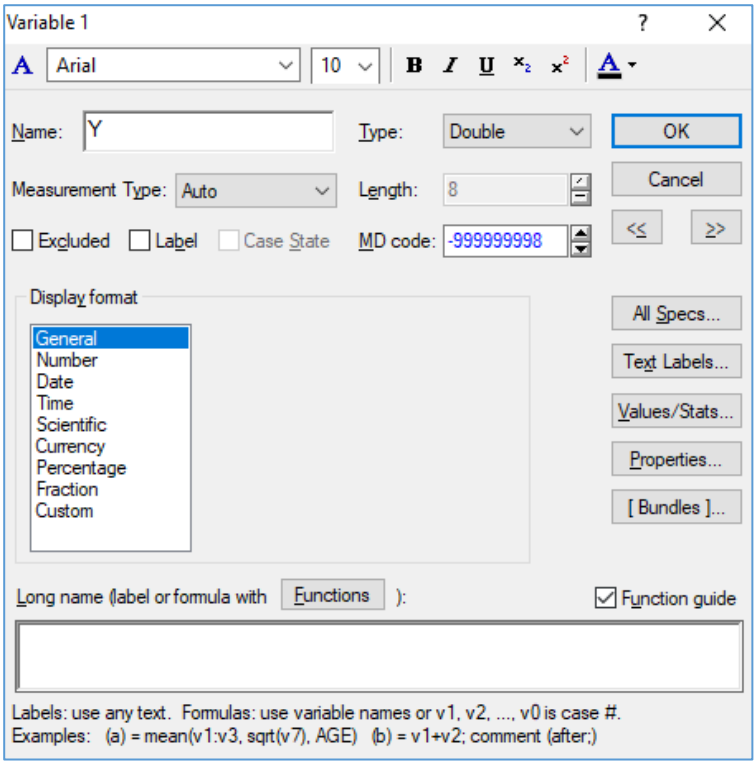
As a stand-alone window – Создание листа рабочей книги.



Шаг 2. После соответствующих установок переходим к рабочему листу электронной таблицы. В данном пакете переменные (столбцы) имеют стандартные имена Var1, Var2, Var3, …, как правило, исследователь заменяет их на необходимые обозначения, используемые в анализе. В нашем случае необходимо заменить Var1

на Y, Var2 на X1, Var3 на X2. Для этого, необходимо дважды щелкнуть по заглавию переменной (Var1) или выбрать в главном меню Data Variable Specs… (Данные – Спецификация переменных).

Шаг 3. В появившемся окне можно изменить шрифт, его размер и т.д. В группе опций Display format (Выводимый формат) можно задать формат данных содержащихся в соответствующим столбце.



В поле Long name (label or formula with Functions): (Длинное имя (вставка функции) можно внести дополнительную информацию по переменной или преобразовать (создать новую) переменную, введя формулу.

В данном случае ограничимся внесением в поле Name: (Имя:) вместо обозначения Var1 букву Y. Далее вносим данные в поле таблицы.

Форматирование файла данных

Часто при создании рабочей таблицы (рабочей книги) возникает необходимость добавления или удаления строк (столбцов). Для этого в главном меню выберем Insert – Add Cases (Вставка – Добавить значение/строку). В появившемся окне укажем, сколько строк необходимо ввести (How many). Также необходимо указать после какой строки произвести вставку нового значения (Insert after cases).

Аналогичным образом производиться добавление новой переменной Insert – Add Variables (Вставка Добавить переменную).

Сохранение файла данных

Для того чтобы сохранить созданный файл данных, необходимо выбрать в строке главного меню пакета команду File Save As. (Файл Сохранить как), после этого откроется диалоговое окно Save As (Сохранить как).

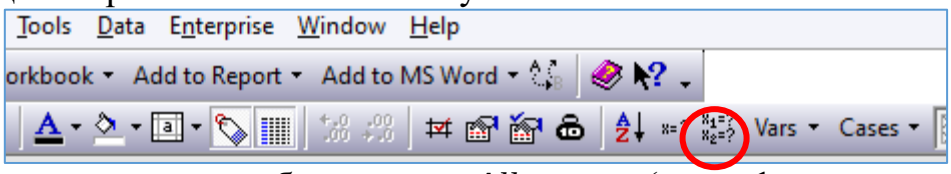
Стоит упомянуть о том, что файлы, получаемые при анализе данных в пакете STATISTICA можно разделить на два типа:

1. Рабочая таблица (книга). В связи с этим рекомендуется все таблицы сохранять под именем Таблица X.x.sta. (X – номер выполняемой работы, x – номер таблицы в данной работе).

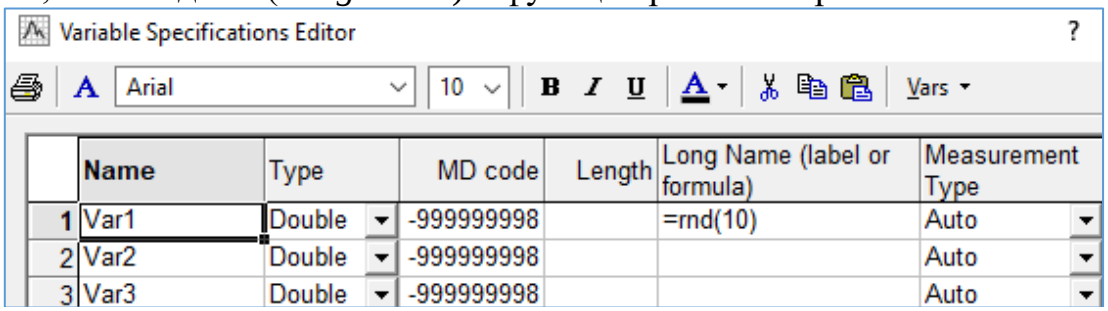
2. Результаты расчетов и рисунки (графики) выводятся в Workbook (Рабочая книга). В связи с этим данные этого типа рекомендуется сохранять под именем Итоги X.x.stw (X – номер выполняемой работы, x – номер итоговой таблицы или рисунка в данной работе).

Задание

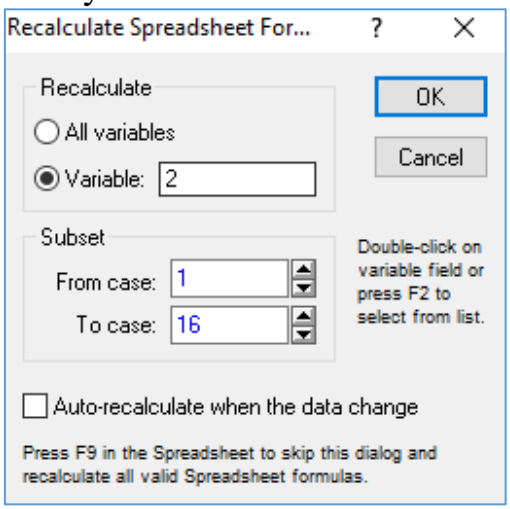
Создать таблицу для 3 выборок по 10 наблюдений. Заполнить таблицу следующим образом: нажать клавишу *Vars*



и в появившемся окне выбрать пункт *All specs*... (спецификация всех). В результате этих действий окно-таблица, в первом столбце которой находятся названия переменных (*var1, var2,..., var20*), во втором – тип этих переменных (по умолчанию – *Double*), в третьем – код переменной, а в четвертом длина, в последнем (*Long Name*) – функция расчета переменных.



Выделить первую клетку пятого столбца и ввести формулу = *rnd(10)*. Это означает, что будут сгенерированы случайные числа, равномерно распределенные на отрезке [0, 10]. Если нужно сделать перерасчет по формуле, можно использовать кнопку  :



Аналогичным образом заполнить второй столбец случайными числами в диапазоне [0, 20].

Отформатировать полученные данные. Изменить имена столбцов: выделить столбец, в контекстном меню выбрать *Variable Spec*, в открывшемся

окне задать новое имя столбца в поле *Name*, формат данных в поле *Display Format* – числовой (*Number*) и количество цифр после десятичной запятой – *Decimal places* –3). Скорректировать размеры таблицы: в меню *Format* выбрать *Variables*, далее *AutoFit* (*Автоподбор*). Сохранить полученную таблицу.

ТЕМА 2. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ. ПРЕДСТАВЛЕНИЕ ДАННЫХ В ИССЛЕДОВАНИЯХ

Генерация случайных числе с заданным законом распределения

Одной из основных составляющих любого анализа данных является описательная статистика (дескриптивная статистика). Её главной задачей является предоставление сжатой и концентрированной характеристики изучаемого явления в числовом и графическом виде.

Формирование выборки

Ознакомимся с некоторыми внутренними функциями пакета STATISTICA. Генератор случайных чисел, распределенных равномерно на отрезке [0;1], запускается формулой =***rnd(1)***. Случайные числа, распределенные равномерно на отрезке [0;2], можно сгенерировать с помощью оператора ***=rnd(2)***. Оператор =***rnd(b-a)+a*** генерирует числа, распределенные равномерно на отрезке *[a,b]*.

Примеры функций, позволяющие сформировать выборки чисел, распределенных по разным законам распределения.

1. Нормальное распределение. **= *VNormal (rnd(1); 2; 3)*** *для N (2;3)*
2. Экспоненциальное распределение **= *VExpon (rnd(1); 2)*** *для E (0.5) со средним 1/2 = 0.5*
3. Распределение Коши **=*VCauchy (rnd(1); 0; 1)*** *для C (0; 1)*
4. Логарифмическое распределение **= *VLognorm (rnd(1); 0,5; 0,5)*** *для Lgn (0,5; 0,5)*
5. Распределение *Хи-квадрат* **= *VChi2 (rnd(1); 8)*** *для* χ *2*

Задание 1. Сформировать выборку по заданному закону распределения из 100 элементов (R – равномерный закон распределения, N – нормальный закон распределения, E – экспоненциальный закон распределения).

Вариант	**Закон**	**Объем**	**Вариант**	**Закон**	**Объем**
1	R [0; 2]	50	**7**	R [3; 6]	70
2	N(2;0.5)	60	**8**	N(l;4)	60
3	E(3)	70	**9**	E(1)	70
4	R [1,3]	80	**10**	R[0;3]	80
5	N(0; 1)	50	**11**	N (0; 4)	50
6	E (2)	60	**12**	E(5)	60

С помощью меню Graphs отобразить сгенерированные данные графически. Отформатировать полученный график.

Задание 2. Построить случайную последовательность, состоящую из чисел 1, 2, 3, для которой вероятность появления единицы равна 0,25, вероятность появления двойки – 0,25, вероятность появления тройки –0,5. Подсказка: используйте для формирования последовательности выражение =(c<=0,25)*1+(c>0,25 and c<=0,5)*2+(c>0,5)*3, где c – случайное число из диапазона [0,1]. Выполнить задание по варианту:

Вариант 1.

X	1	2	3	4	5
p(x)	0,6	0,1	0,05	0,1	0,15

Вариант 2.

X	-2	-1	0	1	2
p(x)	0,02	0,28	0,5	0,1	0,1

Вариант 3.

X	-1	0	1	2	3
p(x)	0,35	0,2	0,05	0,1	0,3

Вариант 4.

X	-2	-1	1	2	3
p(x)	0,5	0,1	0,1	0,1	0,2

Вариант 5.

X	-3	-2	-1	0	1
p(x)	0,2	0,2	0,1	0,3	0,2

Вариант 6.

X	-1	-2	1	2	3
p(x)	0,25	0,1	0,35	0,15	0,15

Вариант 7.

X	1	2	3	4	5
p(x)	0,3	0,25	0,1	0,05	0,3

Вариант 8.

X	-2	0	2	4	6
p(x)	0,2	0,2	0,1	0,4	0,1

Вариант 9.

X	-3	-1	1	2	3
p(x)	0,35	0,2	0,15	0,1	0,2

Вариант 10.

X	-2	-1	0	1	2
p(x)	0,35	0,2	0,2	0,15	0,1

Вариант 11.

X	0	2	3	4	5
p(x)	0,1	0,1	0,1	0,5	0,2

Вариант 12.

X	2	3	4	5	6
p(x)	0,3	0,2	0,1	0,3	0,1

Задание 3. Создать файл с результатами воздействия лекарственного препарата на кровяное давление. Исходные данные содержатся в таблице (имена столбцов задайте самостоятельно):

Номер пациента	Систолическое давление до	Систолическое давление после	Разность	Диастолическое давление до	Диастолическое давление после	Разность
1	210	201	-9	130	125	-5
2	169	165	-4	122	121	-1
3	187	166	-21	124	121	-3
4	160	157	-3	104	106	2
5	167	147	-20	112	101	-11
6	176	145	-31	101	85	-16
7	185	168	-17	121	98	-23
8	206	180	-26	124	105	-19
9	173	147	-26	115	103	-12
10	146	136	-10	102	98	-4

Построить линейные графики и коробочные графики (ящик с усами) для систолического и диастолического давления (до и после).

Построить двухмерное и трехмерное корреляционное поле для разностей (для трехмерного поля в качестве третьей переменной можно взять номер пациента).

Для построения двухмерного корреляционного поля в главном меню нужно выбрать *Graphs 2D Graphs Scatterplots* (Графики – Двухмерные графики – Точечная диаграмма), в появившемся окне необходимо указать переменные.

Для построения трехмерного корреляционного поля в главном меню нужно выбрать *Graphs 3D XYZ Graphs Scatterplots* (Графики – Трехмерные графики – Точечная диаграмма).

Для корректировки (вращения) полученного графика необходимо выбрать *View Rotete* (Вид – Вращение).

Задание 4. Создать файл с качественными данными (данные задать самостоятельно). Построить столбиковую диаграмму, круговую диаграмму.

Вопросы для самоконтроля

1. Для чего используется предварительный анализ данных в контексте исследования?
2. Какие типы данных можно анализировать при помощи пакета STATISTICA?
3. Какие методы визуализации данных предлагает пакет STATISTICA для представления распределения значений переменных?
4. Каким образом можно провести анализ описательной статистики (среднее, медиана, квартили, стандартное отклонение) с использованием STATISTICA?
5. Какие методы предварительного анализа данных позволяют исследовать связи между переменными (например, корреляционный анализ)?
6. Какие возможности предоставляет пакет STATISTICA для проведения регрессионного анализа данных?

ТЕМА 3. УНИВАРИАНТНЫЙ АНАЛИЗ. РЕШЕНИЕ ЗАДАЧ ДЕСКРИПТИВНОЙ СТАТИСТИКИ В ПАКЕТЕ STATISTICA

Одной из основных составляющих любого анализа данных является описательная статистика (дескриптивная статистика). Её главной задачей является предоставление сжатой и концентрированной характеристики изучаемого явления в числовом и графическом виде.

Показатели описательной статистики можно разбить на несколько групп:

- показатели положения, описывающие положение экспериментальных данных на числовой оси – максимальный и минимальный элементы выборки, среднее значение, медиана, мода и др.;
- показатели разброса, описывающие степень разброса данных относительно центральной тенденции. К ним относятся: выборочная дисперсия, разность между минимальным и максимальным элементами (размах, интервал выборки) и др.;
- показатели асимметрии: положение медианы относительно среднего и др.;
- графические представления результатов – гистограмма, частотная диаграмма и др.

Данные показатели используются для наглядного представления и анализа результатов всей исследовательской выборки, экспериментальной и контрольной группы.

Мода (Mo) – это наиболее частое значение в выборке, или среднее значение класса с наибольшей частотой. Мода как центральная тенденция

используется чаще всего для того, чтобы дать общее представление о распределении. В некоторых случаях у распределения могут быть две моды, в таком случае это свидетельствует о бимодальном распределении, что указывает на наличие двух относительно самостоятельных групп.

Медиана (Me) соответствует центральному значению в последовательном ряду всех полученных значений или среднему значению наиболее часто встречающихся значений выборки. Медиана вместе с квартилями используется для представления дискретных переменных или количественных непрерывных переменных с ненормальным распределением.

Среднее арифметическое (M) — это показатель центральной тенденции, полученный делением суммы всех значений данных на число этих данных. Среднее арифметическое используется для представления количественных переменных с нормальным распределением. Среднее значение, как мера центральной тенденции в описательной статистике количественных данных, имеет одно из двух представлений. Первое в виде «M±S», или в зарубежной традиции M (S), где M – среднее, а S – стандартное отклонение (Standard Deviation, равное корню квадратному из дисперсии). Стандартное отклонение предназначено для описания выборок с нормальным распределением и не приспособлено для распределений, отличных от нормального. При нормальном распределении в диапазон M±S укладывается порядка 70% всех значений признака.

Второе представление результатов – в виде «M ± m», где m – стандартная ошибка среднего (Standard Error of Mean), определяемая следующим образом: $m = s/\sqrt{n}$. Однако, подобная форма представления данных в медицине является малоинформативной. В действительности, в биологии (соответственно, и в медицине) определяется не точное значение, а диапазон, в который укладывается большинство значений исследуемого признака, т.е. ширина распределения. Поэтому оптимальным описанием ширины распределения в медицинских исследованиях в настоящее время принимается представление 95% доверительного интервала с указанием нижней (5%) и верхней (95%) границы.

Доверительный интервал представляет собой диапазон значений, который с определённой исследователем вероятностью (чаще всего в медицине это α=0,05 или 95%) включает в себя настоящее популяционное значение.

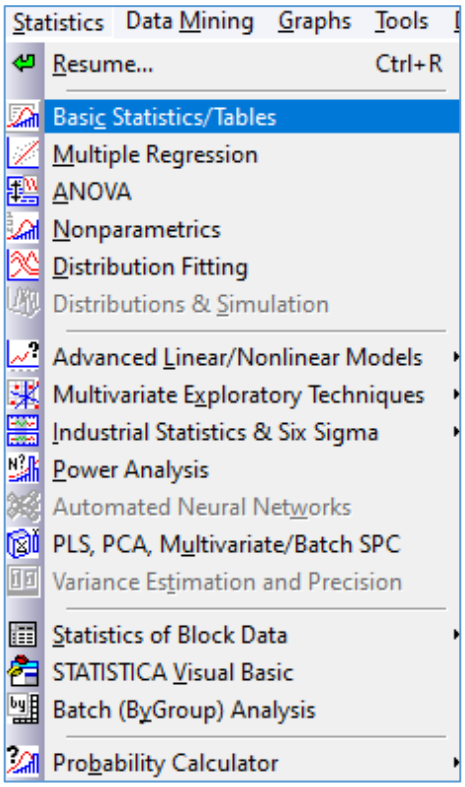
Наиболее адекватная непараметрическая характеристика ширины – это квантили. Квантили представляют собой частоту попадания значений переменной в определённые интервалы. Чаще всего используется разделение на 10 (по 10%) или на 4 интервала (25%, 50%, 75%). При разделении на четыре квантиля (именуемых квартилями) для предоставления оценки центральной тенденции, ширины и асимметрии распределения результатов достаточно трёх чисел: нижний квартиль (25%), 50% квартиль, который соответствует медиане, и верхний квартиль (75%).

Для качественных данных единственной корректной характеристикой будет являться число объектов с данным конкретным значением критерия.

Представляются подобные данные в виде гистограммы или количества объектов с данным конкретным значением критерия относительно общего количества объектов. Проценты, как относительное долевое выражение числа объектов от общего числа объектов равного 100, указываются при объёме выборки более 20.

Расчет описательных статистик производится при помощи модуля *Statistics / Basic Statistics / Tables*. В этом модуле объединены наиболее часто использующиеся на начальном этапе обработки данных процедуры.

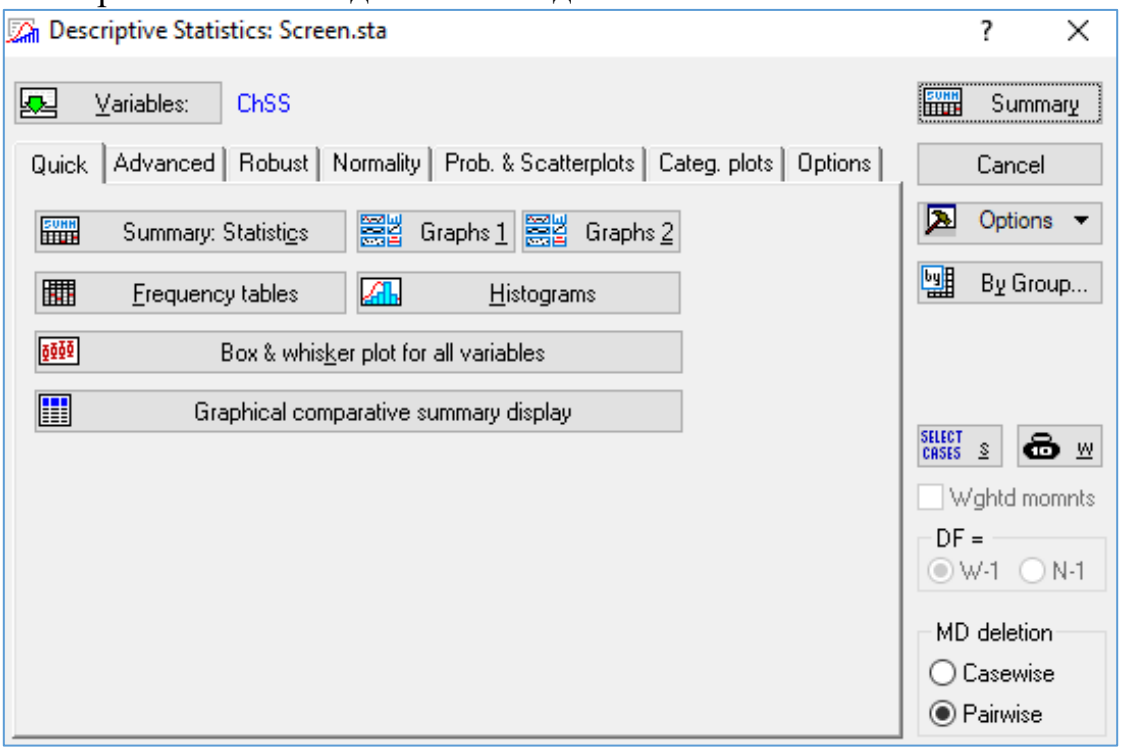
В стартовой панели модуля приводится перечень статистических процедур этого модуля (см. рисунок):



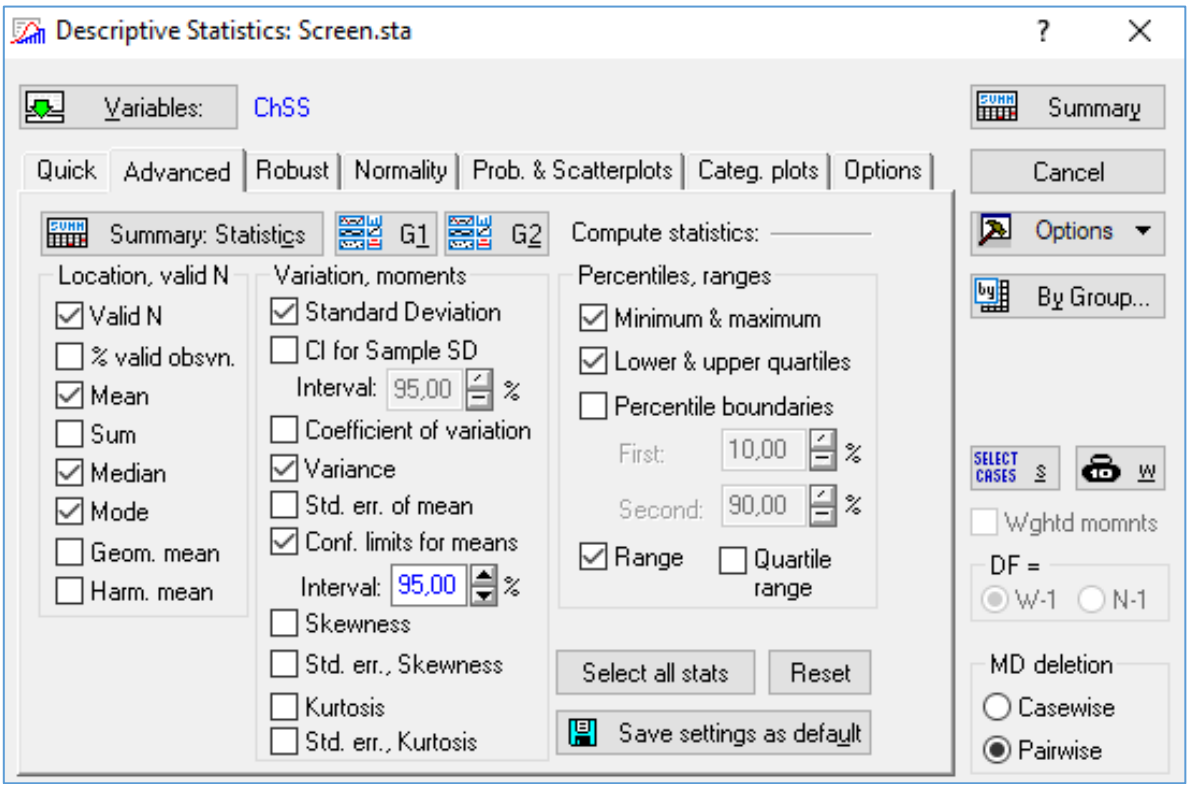
1. Процедура Descriptive statistics (Описательные статистики)

Рассмотрим возможности этой процедуры на примере. В группе – из 60 пациентов проводилось скрининговое обследование, включающее определение частоты сердечных сокращений (ЧСС), исходные данные находятся в файле *screen.sta*. Выполним первичную обработку статистических данных.

После открытия файла *screen.sta* и выбора процедуры *Descriptive statistics* на экране появится одноименное диалоговое окно:



В этом окне при помощи кнопки *Variables* следует выбрать переменные для анализа (в данном случае это переменная *ChSS*).



Далее на вкладке *Advanced* нужно указать, какие показатели требуется вычислить:

Valid N – объем выборки;

Mean – среднее значение;

Sum – сумма;

Median – медиана;

Standard Deviation – стандартное отклонение;

Variance – дисперсия;

Standard error of mean – стандартная ошибка среднего;

95% confidence limits of mean – 95%-ый доверительный интервал для среднего;

Minimum, maximum – минимальное и максимальное значения;

Lower, upper quartiles – нижний и верхний квартили;

Верхний квартиль – это такое значение случайной величины, больше которого по величине 25% случаев выборки. Нижний квартиль – это такое значение случайной величины, меньше которого по величине 25% случаев выборки.

Range – размах (расстояние между наибольшим (maximum) и наименьшим (minimum) значениями признака);

Quartile range – интерквартильная широта (расстояние между нижним и верхним квартилями);

Skewness – асимметрия;

Асимметрия характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричной кривой коэффициент асимметрии равен нулю. Если правая ветвь кривой, начиная от вершины, больше левой (правосторонняя асимметрия), то коэффициент асимметрии больше нуля. Если левая ветвь кривой больше правой (левосторонняя асимметрия), то коэффициент асимметрии меньше нуля. Асимметрия менее 0,5 считается малой.

Standard *error of Skewness* – стандартная ошибка асимметрии;

Kurtosis – эксцесс;

Эксцесс характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутости кривой. В кривой нормального распределения эксцесс равен нулю. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью, т.е. является более крутой по сравнению с нормальной, а случаи более густо группируются вокруг среднего. При отрицательном эксцессе кривая является более плосковершинной, т.е. более пологой по сравнению с нормальным распределением. Отрицательным пределом величины эксцесса является число – 2, положительного предела нет.

Standard error of Kurtosis – стандартная ошибка эксцесса.

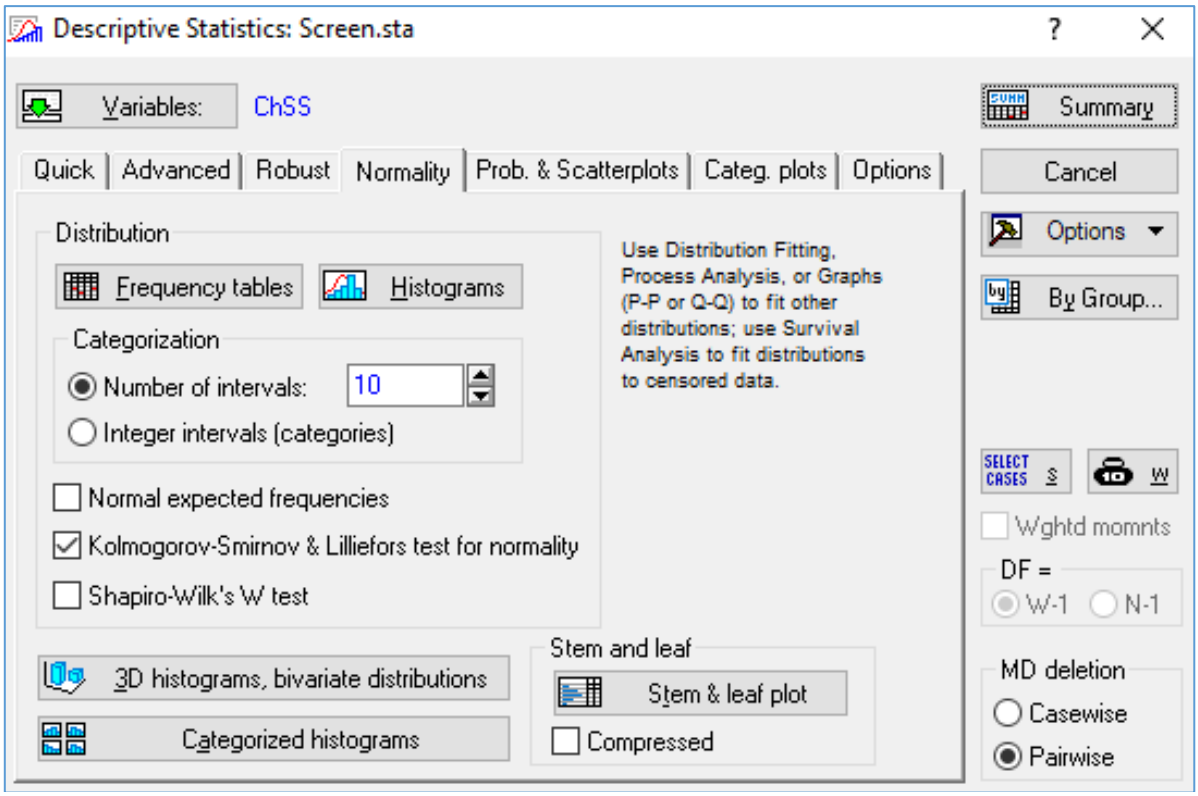
Далее, нажав на кнопку *Summary* можно получить таблицу с требуемыми показателями.

Variable	Descriptive Statistics (Screen.sta)								
	Valid N	Mean	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range	Variance
ChSS	60	67,28333	64,00000	58,00000	84,00000	62,00000	72,00000	26,00000	49,73192

2. Построение таблицы, графиков частот и гистограммы

На первом этапе обработки данных часто возникает необходимость в их группировке. Группировка позволяет представить первичные данные в компактном виде, выявить закономерности варьирования изучаемого признака. Количество классов можно приблизительно наметить следующим образом: при количестве наблюдений 25-40 – 5-6 классов, при количестве наблюдений 40-60 – 6-8 классов, 60-100 – 7-10, 100-200 наблюдений – 8-12, более 200 наблюдений – 10-15 классов.

Для построения гистограмм и таблиц частот используется вкладка *Normality* окна *Descriptive statistics*. Число классов (интервалов) группировки данных устанавливается при помощи счетчика переключателя *Number of intervals*. Если сделать активным переключатель *Integer intervals (categories)*, то классы (интервалы) группировки будут представлять собой целые числа.

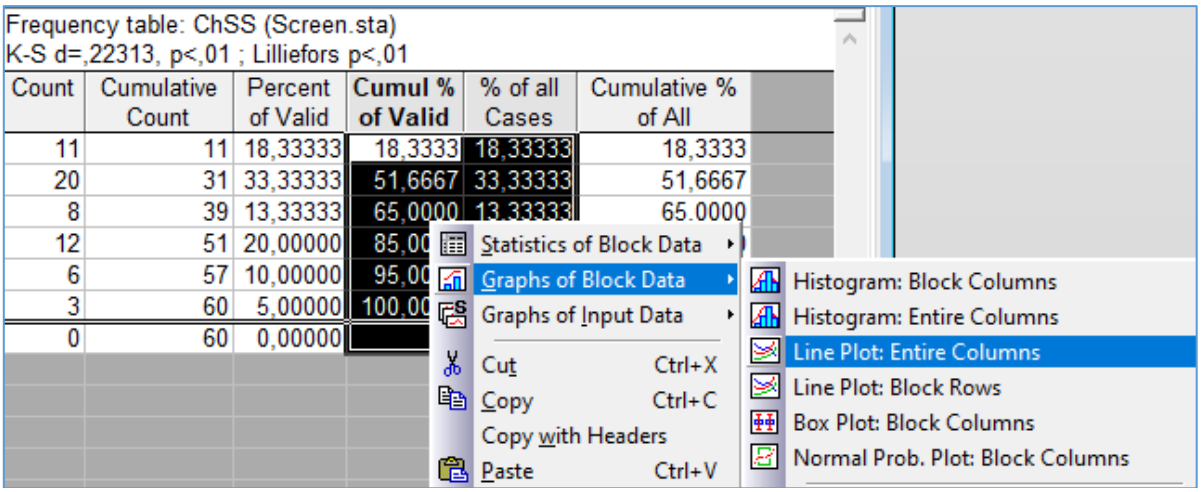


После нажатия кнопки *Frequency tables* (частотная таблица) получаем таблицу следующего вида:

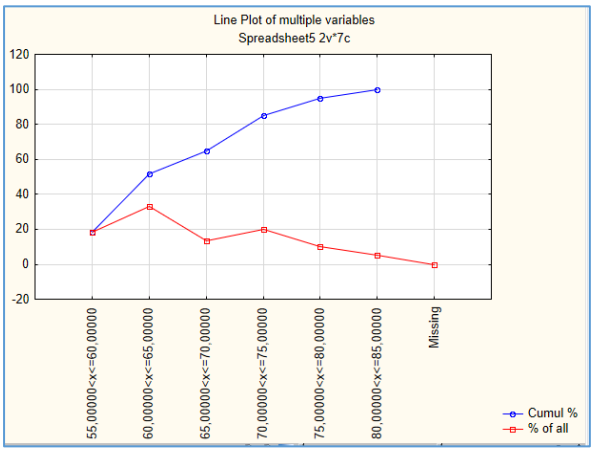
Category	Frequency table: ChSS (Screen.sta) K-S d=,22313, p<,01 ; Lilliefors p<,01					
	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
55,00000<x<=60,00000	11	11	18,33333	18,3333	18,33333	18,3333
60,00000<x<=65,00000	20	31	33,33333	51,6667	33,33333	51,6667
65,00000<x<=70,00000	8	39	13,33333	65,0000	13,33333	65,0000
70,00000<x<=75,00000	12	51	20,00000	85,0000	20,00000	85,0000
75,00000<x<=80,00000	6	57	10,00000	95,0000	10,00000	95,0000
80,00000<x<=85,00000	3	60	5,00000	100,0000	5,00000	100,0000
Missing	0	60	0,00000		0,00000	100,0000

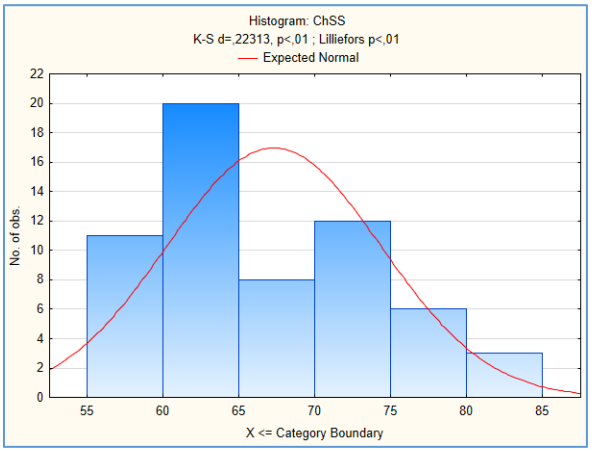
В первом столбце таблицы заданы интервалы для переменной *ChSS*, причем последняя строка содержит пропущенные значения. Второй столбец содержит число попаданий переменной в интервалы, третий столбец – кумулятивное число попаданий, четвертый и шестой столбцы – частоты в процентах соответственно для имеющихся в наличии наблюдений и для всех наблюдений, пятый и седьмой столбцы – кумулятивные частоты в процентах, соответственно для имеющихся в наличии и для всех наблюдений.

Для построения графиков частот и кумулятивных частот нужно выделить четвертый и пятый столбцы и в контекстном меню выбрать пункт Graphs of Block Data / Line Plot Columns.



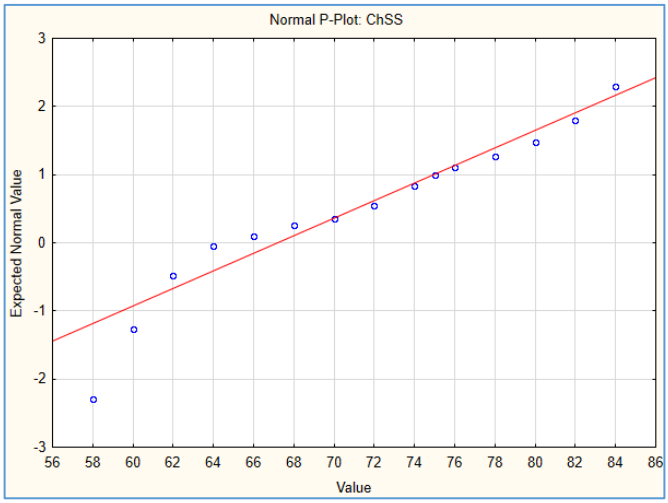
Для построения частотной гистограммы требуется в окне описательной статистики нажать *Histograms*. Частоты в классах исследуемой совокупности будут изображены в виде столбцов, а сплошной линией будет показана нормальная функция распределения.





На гистограмме показана кривая плотности нормального распределения, а также критерий Колмогорова-Смирнова (d). Статистика Колмогорова-Смирнова оказалась равной 0,223. Чем меньше величина этой статистики, тем ближе распределение случайной величины к нормальному.

О нормальности распределения можно судить по графику на нормальной вероятностной бумаге. Его легко построить при помощи опции *Normal probability plots* вкладки *Prob.& Scatterplots*.



Чем ближе распределение к нормальному виду, тем лучше значения ложатся на прямую линию. На данном графике отображается зависимость реальных

частот значения признака от ожидаемых, «нормальных». Если между наблюдаемым и ожидаемым распределениями нет никакой разницы, точки на этом графике выстроятся строго вдоль прямой.

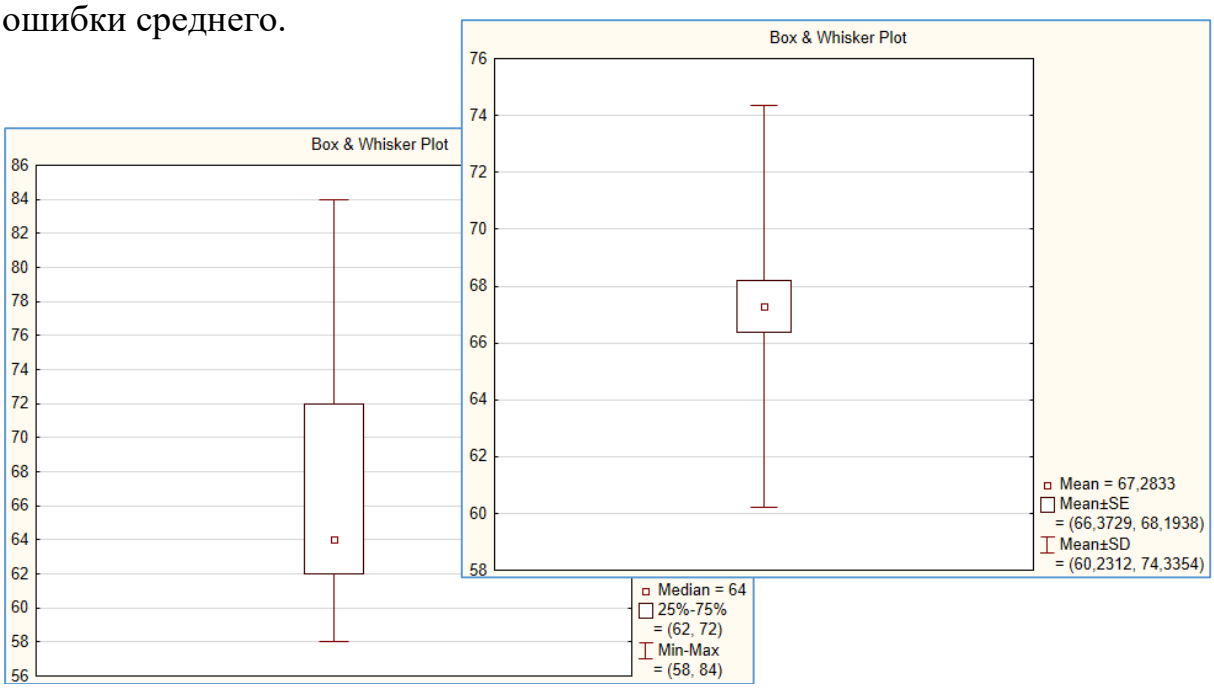
Для визуализации описательных статистик можно построить статистические графики типа «коробок». Это легко можно сделать при помощи кнопки *Box & Whisker plot for all variable* вкладки *Quick*, предварительно указав во вкладке *Options* нужные статистики:

Median/Quart./Range – Медиана / Квартили / Размах;

Mean/SE/SD – Среднее / Ошибка среднего / Стандартное отклонение

Mean/SD/1.96SD – Среднее / Стандартное отклонение / Интервал 1,96* стандартного отклонения;

*Mean/SE/1.96*SE* – Среднее / Ошибка среднего / Интервал 1,96* ошибки среднего.



3. Определение вида распределения случайной величины

В основе выборочных данных лежит некоторое распределение. Его идентификация по выборочным значениям дает возможность более точного анализа и установления некоторых характеристик выборочной совокупности. При использовании описательной статистики важно учитывать тип данных и параметры распределения, характеризующиеся показателями асимметрии и гистограммой распределения. Наиболее часто употребляемыми критериями для проверки гипотезы о законе распределения являются критерий Пирсона, критерий χ^2 и критерий Колмогорова-Смирнова: при отличии распределения признака в изучаемой выборке от нормального распределения со статистической значимостью $p<0,05$ распределение признака в выборке признаётся отличающимся от нормального, и наоборот.

Основными типами распределений признаков являются: дискретные (для дискретных признаков – биномиальное, распределение Пуассона,

распределение Бернулли) и непрерывные (для непрерывных признаков – нормальное (гауссово, или распределение Гаусса), логнормальное, постоянное, экспоненциальное, хи-квадрат χ^2). В соответствии с типом распределения применяется два принципа статистической обработки: параметрический и непараметрический. Параметрический принцип включает все методы анализа нормально распределенных количественных признаков. Непараметрический принцип используется во всех остальных случаях – для анализа количественных признаков независимо от вида их распределения и для анализа качественных признаков.

Непараметрические методы считаются менее мощными по сравнению с параметрическими, т.е. иногда они не позволяют выявить статистические закономерности, которые могут быть выявлены с помощью параметрических методов. В то же время непараметрические методы более надежны в случаях, когда есть сомнения в том, что анализируемый признак имеет нормальное распределение. Для нормально распределенных признаков параметрические и непараметрические методы дают близкие результаты.

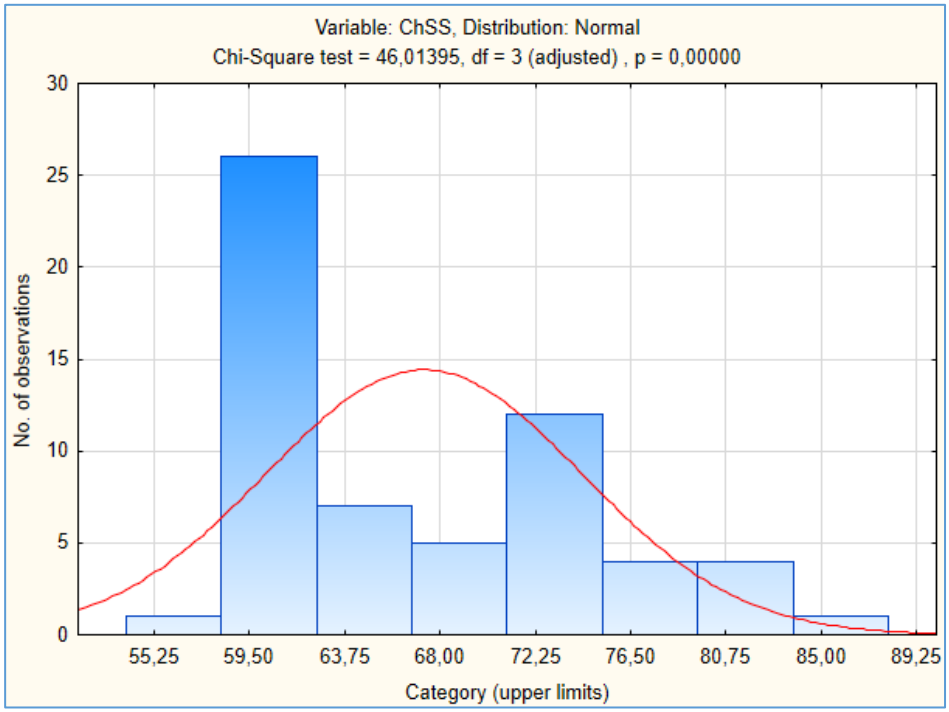
Наиболее полной характеристикой случайной величины, полученной при измерении какого-либо свойства или параметра, является дифференциальная или интегральная функция распределения, устанавливающая зависимость между значением случайной величины и вероятностью появления данного значения.

Для определения вида распределения в исследуемой совокупности используется критерий Пирсона (хи-квадрат). Необходимо выбрать *Statistica / Distribution Fitting*, в появившемся окне выбрать переменную и параметры (количество категорий). Затем поочередно выбирать в окне *Distribution* различные виды распределений (нормальное, равномерное, экспоненциальное, гамма-распределение и т.д.) и строить частотные гистограммы. Проверим выборку на соответствие нормальному закону распределения.

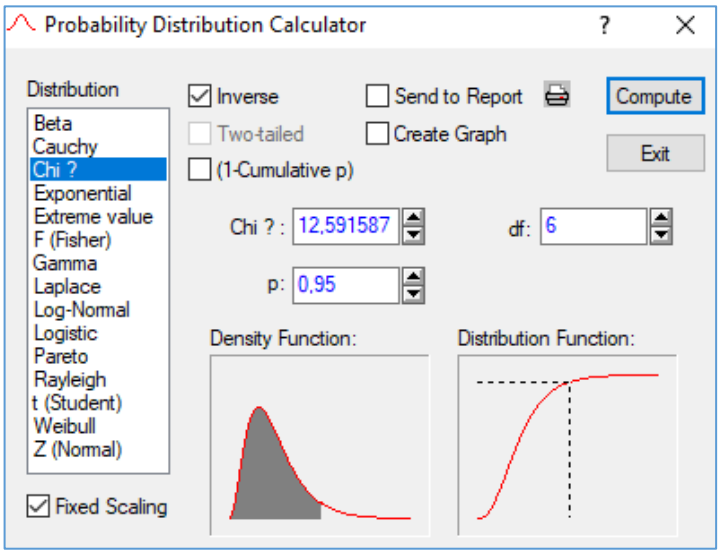
Полученный рисунок показывает, что распределение значений исследуемого параметра отличается от «нормального» (столбики гистограммы не формируют колоколообразную кривую). Это заключение основано на визуальном анализе, однако оно имеет и более строгое подтверждение. В верхней части гистограммы представлены результаты теста хи-квадрат. Данный тест проверяет гипотезу о том, что наблюдаемое распределение не отличается от теоретически ожидаемого, «нормального». Если вероятность ошибки при отклонении этой гипотезы оказалась намного больше 0,05 ($p>0,05$), то гипотеза верна. Иными словами, распределение значений, составляющих данную выборку, статистически не отличается от «нормального».

Также можно воспользоваться расчетным значением хи-квадрат. Подходящим можно считать то распределение, для которого критерий хи-квадрат, представленный на графике вверху, не превышает критического значения, которое определяется по таблице с учетом указанных на рисунке

значений числа степеней свободы *df*. Для расчета критических значений данного и других коэффициентов (Стьюдента, Фишера и т.д.) можно воспользоваться процедурой Вероятностный калькулятор *Statistica / Probability Calculator / Distributions*.



Критическое значение хи-квадрат для 6 степени свободы и уровня значимости 0,95 составляет 12,6, следовательно. Так как расчетное значение хи-квадрат, равное 46,01 значительно больше критического, то можно сделать вывод, что распределение ЧСС в данной выборке не является нормальным.



Однако необходимо отметить, что применение теста хи-квадрат достаточно часто приводит к ошибочному выводу о «нормальности» распределения (мощность данного теста относительно невысока).

В связи с этим лучше воспользоваться другими тестами, которые можно найти в модуле *Basic Statistics / Descriptive Statistics*.

Открыть закладку *Normality* и выбрать опции *Kolmogorov-Smirnov and Lilliefors test for normality* (Тест Колмогорова-Смирнова и Лиллифорса) или *Shapiro-Wilk's W test* (W-тест Шапиро-Уилка). Эти тесты также проверяют гипотезу об отсутствии различий между наблюдаемым и теоретически

ожидаемым, «нормальным» распределением. Наибольшей мощностью, особенно при небольших выборках (n<50), обладает тест Шапиро-Уилка (*Shapiro-Wilk's W test*).

Задание для самостоятельного выполнения

Имеются данные обследования двух групп пациентов: экспериментальной (группа 1) и контрольной (группа 2) – файл MedStat.sta. Расшифровка переменных (n – номер группы):

Вид исследования	Переменная	Показатель	Номер варианта
Физикальное обследование на этапе скрининга	ntemp	Температура	1
	nChSS	Частота сердечных сокращений	2
	nChD	Частота дыхания	
	nADv	Артериальное давление верхнее	
	nADn	Артериальное давление нижнее	
Общий анализ крови	nGem	Гемоглобин	3
	nLejkKr	Лейкоциты	4
	nLimf	Лимфоциты	5
	nMono	Моноциты	
	nNejtrP	Нейтрофилы палочкоядерные	
	nNejtrS	Нейтрофилы сегментоядерные	
	nSOE	СОЭ	
	nTromb	Тромбоциты	6
Биохимические показатели крови	nGluk	Глюкоза	7
	nKalij	Калий	8
	nNatr	Натрий	
	nKreat	Креатинин	9
	nObBel	Общий белок	10
	nObBilir	Общий билирубин	11
	nObXol	Общий холестерин	12

По имеющимся данным требуется:

1) изучить основные характеристики выборки для экспериментальной и контрольной групп пациентов по заданному показателю;

2) построить диаграмму размаха;

3) рассчитать частоты, построить полигон частот и график эмпирической функции;

4) построить гистограмму, сделать вывод о нормальности частотного распределения в группах;

5) используя полученные значения коэффициентов асимметрии и эксцесса для каждой выборки, проанализировать их для контрольной

и исследуемой группы и сравнить полученные выводы с видом частотных распределений на гистограмме:

- в какую сторону смещено частотное распределение выборки (влево, вправо) относительно нормального;
- является ли вершина частотного распределения выборки более острой или пологой по сравнению со стандартным частотным распределением;

6) подобрать теоретическое распределение, хорошо сглаживающие исходные данные;

7) обосновать выбор методов дальнейшей статистической обработки данных (параметрический, непараметрический).

Вопросы для самоконтроля

1. Что понимают под генеральной совокупностью?
2. Для каких целей формируется выборка?
3. Что означает понятие репрезентативная выборка?
4. Какие характеристики позволяют описать выборку?
5. Объясните сущность понятия доверительного интервала среднего, рассчитанного с вероятностью 95%.
6. Когда для описания центральной тенденции выборки следует применять среднее, медиану, моду?
7. Какой показатель описательной статистики характеризует симметричность частотного распределения выборки?
8. Какой показатель описательной статистики характеризует островершинность частотного распределения выборки?
9. Перечислите основные критерии нормальности частотного распределения выборки.
10. Для каких целей определяют нормальность частотного распределения выборки?
11. Какие методы статистического анализа (параметрические или непараметрические) являются наиболее точными?

ТЕМА 4. СРАВНЕНИЕ С ПОРОГОВЫМ ЗНАЧЕНИЕМ. ИСПОЛЬЗОВАНИЕ ИНТЕРВАЛЬНОЙ ОЦЕНКИ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Уровень значимости – максимальное значение вероятности события, при котором событие считается практически невозможным. Как правило, рассматривают уровни значимости α=0,05 (обычные требования надежности) и α=0,01 (повышенные требования надежности). Величину P=1-α называют доверительной вероятностью или уровнем надежности. Эта величина используется, чтобы судить о статистической значимости принятого

решения. Уровень значимости находится в обратной зависимости от надежности результата. Более высокая статистическая значимость соответствует более низкому уровню доверия к полученному значению.

Доверительный интервал – это вычисленный на основе выборки интервал значений признака, который с заданной доверительной вероятностью содержит неизвестный параметр генеральной совокупности. Если мы увеличиваем количество измерений, то оценка параметра становится более точной и доверительный интервал уменьшается (если в измерении нет систематических ошибок).

Таким образом, интервал, в который с заданной вероятностью попадает истинное значение исследуемого признака, называется доверительным интервалом, а вероятность того, что истинное значение оцениваемой величины находится внутри этого интервала – доверительной вероятностью, или надежностью.

Доверительные интервалы используются для сравнения переменной исследования с пороговым значением, в частности, для сравнения выборки с генеральной совокупностью. Пусть известно среднее значение изучаемого признака генеральной совокупности. Необходимо сопоставить с ним среднее значение выборки. Одним из способов решения данной задачи является оценка изучаемого признака с помощью 95% доверительного интервала для среднего значения (в случае нормального закона распределения изучаемого признака) или медианы (если вид распределения не является нормальным или неизвестен). Если рассчитанный доверительный интервал для среднего не включает среднее значение генеральной совокупности, то с заданной доверительной вероятностью можно утверждать, что выборка статистически отличается от генеральной совокупности.

Случай нормального распределения признака

Постановка задачи: известно среднее значение изучаемого признака в популяции. Необходимо сопоставить с ним среднее значение изучаемой группы.

Пример. Пусть имеются данные по росту детей в возрасте 10 лет (в см): 122, 136, 139, 134, 128, 130, 145, 128, 129, 132, 133, 124, 131, 134, 142. Из предыдущих исследований известно, что данные такого типа распределены нормально и что средний рост детей такого возраста составляет 140 см. Определить, не отстают ли в росте дети исследуемой группы.

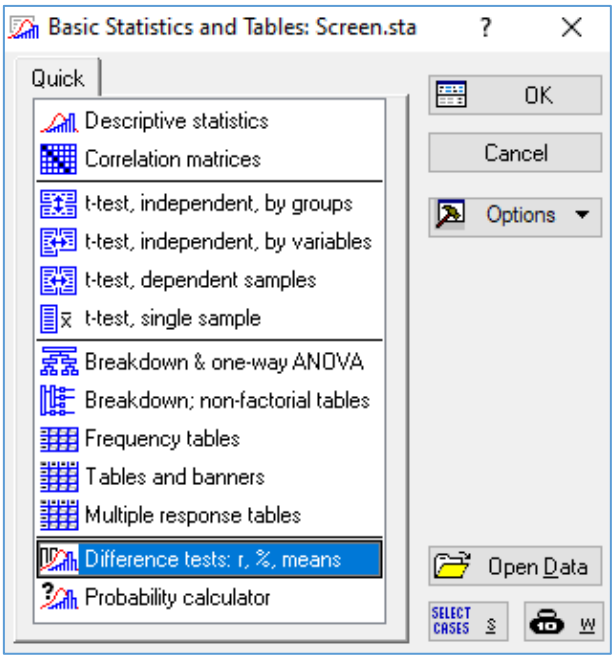
Способ 1 – оценка с помощью 95% доверительного интервала для среднего значения. Для этого необходимо определить среднее значение и доверительный интервал для среднего выборки; рассчитать значения верхней и нижней границ доверительного интервала. Среднее значение по выборке и границы доверительного интервала можно получить с помощью модуля дескриптивной статистики:

Variable	Descriptive Statistics (Spreadsheet2)					
	Valid N	Mean	Confidence -95,000%	Confidence 95,000%	Median	Std.Dev.
Var1	15	132,4667	128,9965	135,9368	132,0000	6,266312

Интерпретация результатов: если рассчитанный доверительный интервал не содержит популяционное среднее, то с заданной доверительной вероятностью можно утверждать, что выборка статистически значимо отличается от генеральной совокупности.

В данном случае, так как известное среднее генеральной совокупности, равное 140 см, не входит в доверительный интервал от 130 до 136 см, можно утверждать с доверительной вероятностью 95%, что в исследуемой группе дети отстают по росту от сверстников.

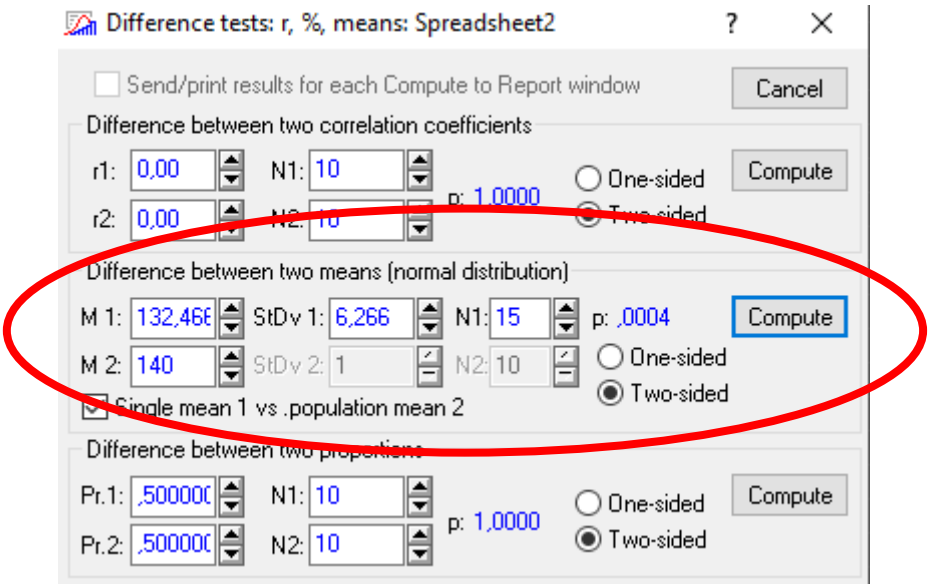
Способ 2 – применение t-критерия Стьюдента для проверки гипотезы об отсутствии различий средних. Для этого можно воспользоваться модулем *Basic Statistics / Difference tests*:



Далее в блок для сравнения двух средних значений нужно внести значения среднего, среднеквадратичного отклонения и число наблюдений в соответствующие окна, выполнить вычисления и оценить полученное значение p.

Если вычисленное значение p>0,05, то нулевая гипотеза об отсутствии различий средних не отклоняется. В противном случае следует отклонить нулевую гипотезу и принять альтернативную гипотезу о том, что различия выборочного и популяционного среднего значения существуют.

В примере значение p=0,0004, что много меньше заданного уровня значимости 0,05, следовательно, выборочное среднее отличается от среднего генеральной совокупности.



Случай любого распределения признака

Если выборка имеет распределение, отличное от нормального, то используется вычисление 95% доверительного интервала для медианы. Для этого нужно выполнить следующее:

1) по таблице с учетом объёма выборки и необходимой доверительной вероятности найти ранги, значения которых соответствуют границам доверительного интервала;
2) ранжировать значения изучаемого признака в порядке возрастания;
3) установить, которые из значений изучаемого признака соответствуют табличным рангам. Эти значения признака и являются границами доверительного интервала.

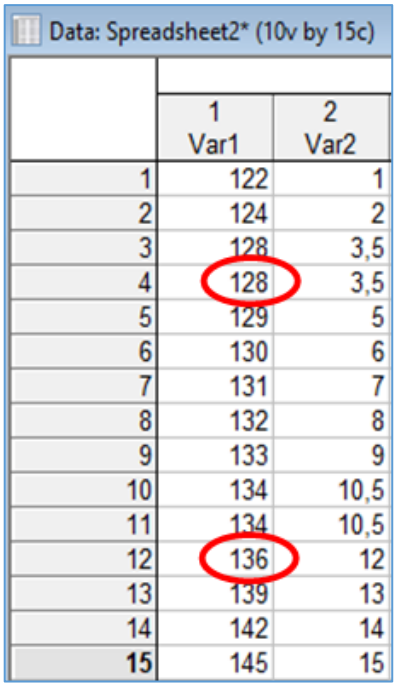
Для заданной выборки медиана равна 132 см. Для того, чтобы получить ранги, нужно выделить диапазон данных, отсортировать его с помощью пункта меню *Data / Sort* и найти для каждого элемента выборки ранг с помощью *Data / Rank*. Следует отметить, что программа Statistica уничтожает исходные данные, проставляя на их место проранжированный ряд. Поэтому желательно сделать копию данных в другом столбце и искать ранги в нем.

Ранги, с помощью которых вычисляется доверительный интервал для медианы распределения, приведены в таблице, фрагмент которой показан ниже.

Объем выборки	Уровень значимости (доверительная вероятность)		
	α=0,1 (90%)	α=0,05 (95%)	α=0,01 (99%)
10	2; 9	2; 9	1; 10
11	3; 9	2; 10	1; 11
12	3; 10	3; 10	2; 11
13	4; 10	3; 11	2; 12

14	4; 11	3; 12	2; 13
15	4; 12	4; 12	3; 13
16	5; 12	4; 13	3; 14
17	5; 13	5; 13	3; 15
18	6; 13	5; 14	4; 15
19	6; 14	5; 15	4; 16
20	6; 15	6; 15	4; 17

По таблице определяем, что для заданного количества наблюдений, равного пятнадцати и уровня значимости α=0,05 доверительный интервал определяется 4 и 12 рангами. Этим рангам соответствуют значения наблюдений 128 и 136 см.


Data: Spreadsheet2* (10v by 15c)

	1 Var1	2 Var2
1	122	1
2	124	2
3	128	3,5
4	128	3,5
5	129	5
6	130	6
7	131	7
8	132	8
9	133	9
10	134	10,5
11	134	10,5
12	136	12
13	139	13
14	142	14
15	145	15

Верхнюю и нижнюю границы двустороннего доверительного интервала для медианы также можно определить с помощью двух порядковых статистик, соответствующих рангам $[x_{[k]}, x_{[n-k+1]}]$, где целочисленное k находят с использованием приближенной формулы:

$$y = \left\{0{,}5\left(n + 1 - u\sqrt{n - 0{,}5}\right)\right\}.$$

Тогда k – целая часть y, а u – квантиль стандартного нормального распределения (для доверительной вероятности p=0,95 u=1,96).

Таким образом, доверительный интервал для медианы находится в диапазоне от 128 до 136 см. Известное среднее генеральной совокупности 140 см не входит в доверительный интервал для данной выборки, следовательно, можно утверждать с доверительной вероятностью 95%, что в исследуемой группе дети отстают по росту от сверстников.

Задание для самостоятельного выполнения

Будем считать, что известны средние значения по генеральной популяции для каждого показателя базы данных из лабораторной работы №3. Определить, различаются ли средние показатели исследуемой и контрольной групп от среднего по генеральной популяции.

Вид исследования	Переменная	Показатель	Генеральное среднее
Физикальное обследование на этапе скрининга	ntemp	Температура	36,35
	nChSS	Частота сердечных сокращений	70

Общий анализ крови	nGem	Гемоглобин	137
	nLejkKr	Лейкоциты	5,0
	nLimf	Лимфоциты	31,5
	nTromb	Тромбоциты	265
Биохимические показатели крови	nGluk	Глюкоза	4,5
	nKalij	Калий	4,32
	nKreat	Креатинин	73
	nObBel	Общий белок	72
	nObBilir	Общий билирубин	13
	nObXol	Общий холестерин	4,58

Проверить данное утверждение в предположении, что:

1) выборочные данные распределены нормально;
2) выборочные данные распределены по закону, отличающемуся от нормального.

ТЕМА 5. БИВАРИАНТНЫЙ АНАЛИЗ. ВЗАИМОСВЯЗЬ ДВУХ ПЕРЕМЕННЫХ

Выявление и измерение связи между признаками, характеризующими изучаемые явления или процессы, является важнейшей частью исследования. Различают **функциональную и корреляционную** связи. Функциональная связь между явлениями присуща неживой природе. В биологических и медицинских исследованиях) чаще приходится иметь дело со связью между явлениями, когда одной и той же величине одного признака соответствует ряд варьирующих (разных, но близких по величине) значений другого признака, что обусловлено чрезвычайным многообразием взаимодействия различных явлений живой природы. Такого рода связь носит название ***корреляционной*** (correlation—соответствие, соотносительность). В то время как функциональная связь имеет место в каждом отдельном наблюдении, корреляционная связь проявляется только при многочисленном сопоставлении признаков. Характеристика тесноты (силы) связи между ними, выраженная одним числом, называется **коэффициентом корреляции**, **r**. Коэффициент корреляции может принимать значения от –1 до +1. Знак коэффициента корреляции показывает направление связи (прямая или обратная), а абсолютная величина — тесноту связи.

Коэффициент корреляции Пирсона (*r*) представляет собой меру *линейной зависимости* двух переменных. Если возвести его в квадрат, то полученное значение **коэффициента детерминации (*r2*)** представляет 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**.

Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции.

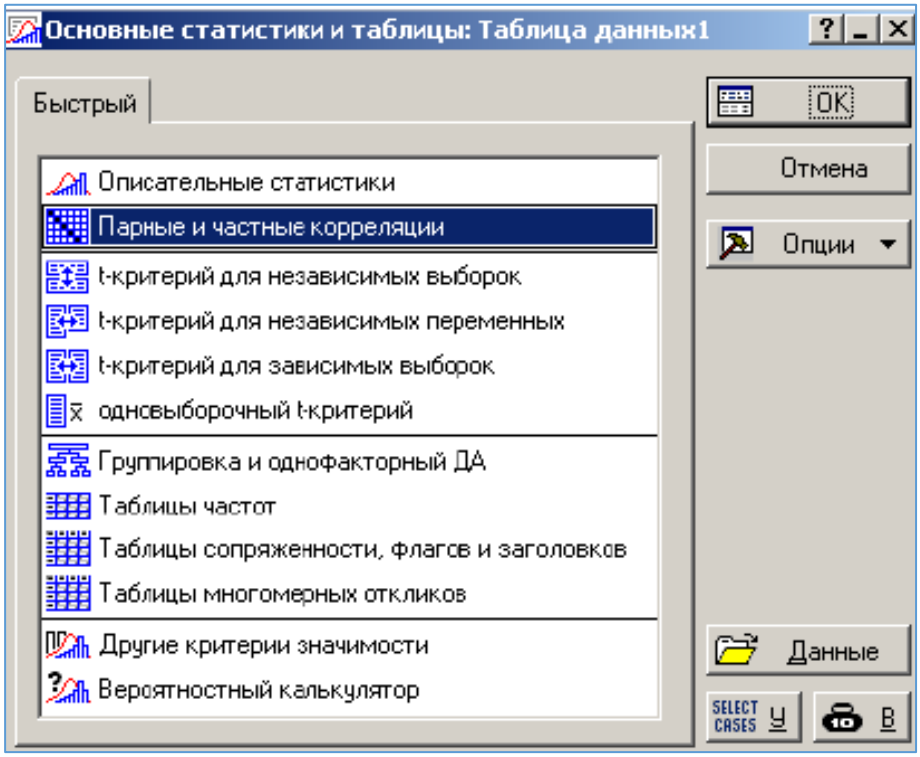
Таким образом, для ответа на вопрос, имеется ли какая-либо статистическая связь между исследуемыми переменными и какова теснота связи, необходимо:

- выбрать с учетом специфики анализируемых переменных показатель статистической связи (парный, множественный или ранговый коэффициент корреляции);
- определить его числовое значение по имеющимся данным;
- проверить полученное значение на статистическую значимость.

Парная корреляция

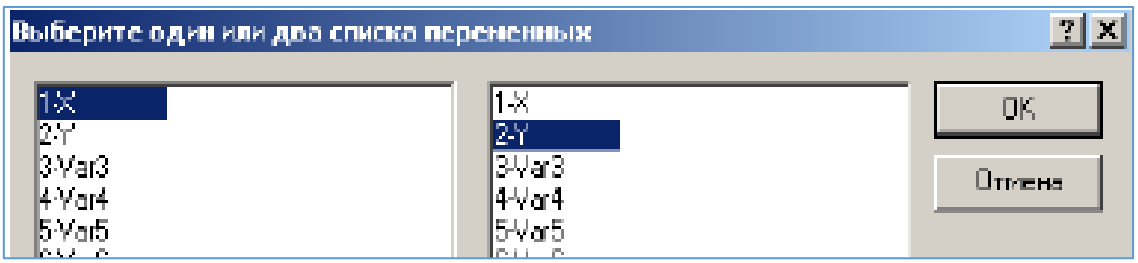
Пример 1. Исходные данные:

Исследуем предполагаемую связь между X и Y.

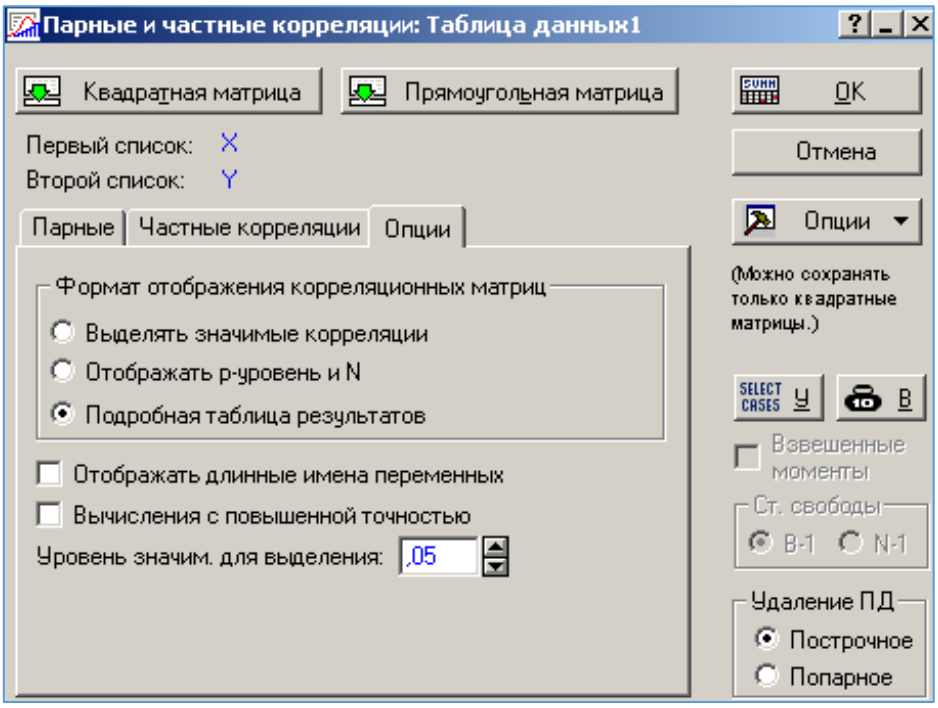


	X	Y
1	3,4	14,3
2	3,6	14,9
3	4,5	17,3
4	4,8	17,3
5	4,9	17,4
6	5,2	17,5
7	5,4	17,6
8	5,7	17,6
9	6,2	17,6
10	6,7	17,8
11	7,1	18
12	7,5	18
13	7,7	18,1
14	7,8	18,1
15	7,9	18,6
16	8	19,7

Переменные для анализа:



Желательно указать подробную таблицу результатов:



Результаты расчета корреляции:

Корреляции (Таблица данных1)
Отмеченные корреляции значимы на уровне p <,05000
(Построчное удаление ПД)

Пер.X и Пер.Y	Среднее	Стд.откл	r(X,Y)	r2	t	p	N	Св.член завис.Y	Наклон завис.Y	Св.член завис.X	Наклон завис.X
X	6,02500	1,554134									
Y	17,48750	1,277432	0,847730	0,718647	5,979923	0,000034	16	13,28929	0,696798	-12,0108	1,031355

На рисунке представлена следующая информация:

- среднее;
- стандартное отклонение;
- значение коэффициента корреляции r;
- значение коэффициента детерминации r2;
- t – критерий;
- p – уровень значимости;
- число коррелируемых пар;
- 13,289 – свободный член уравнения регрессии.
- 0,696 – коэффициент при независимой переменной уравнения регрессии.

В этом примере r = 0,847. Это очень высокое значение (подсвечено красным цветом), показывающее, что построенная регрессия объясняет более 90% разброса значений переменной X относительно среднего.

Множественная корреляция

Пример 2. Для 34 работников предприятия были проанализированы следующие показатели (Prakt5.sta):

X1 – квалификация (разряд);
X2 – трудовой стаж;
X3 – оценка работника с точки зрения руководителя (по 5-бальной шкале);
X4 – количество нарушений трудовой дисциплины в течение года;
X5 – личное отношение к работе (1 – не нравится, 2 – нравится);
X6 – семейное положение (1 – холост, 2 – женат);
X7 – количество пропущенных по болезни дней;
Y1 – средняя заработная плата;
Y2 – выработка (в % к средней по цеху);
Y3 – процент брака.

Факторы X1 – X7 составляют первую группу (группу входных факторов), а факторы Y1 – Y3 вторую группу выходных параметров.

Построим корреляционную матрицу для данных. Для проведения корреляционного анализа необходимо выбрать *Статистика / Основная статистика/Таблицы*, раздел *Correlation matrices*.

Далее необходимо определить, какие переменные будут находиться в строках и столбцах (все, кроме номера). Для представления в корреляционной матрице значений уровня значимости *p* для всех коэффициентов корреляции необходимо во вкладке Options выбрать позицию *Display r, p-levels*. После нажатия на *Summary* будет рассчитана матрица. Красным цветом будут выделены те парные коэффициенты корреляции, для которых уровень значимости не превышает 0,05. (Уровень значимости характеризует вероятность событий, условно принимаемых за невероятные, т.е. чем ниже его величина, тем достоверней результат.)

Для построения графического отображения корреляционной взаимосвязи любых двух переменных необходимо выбрать переменные и нажать *2D scatterplots*. На рисунке будут представлены: функция, отражающая взаимосвязь переменных, ее графическое отображение, значение парного коэффициента корреляции, корреляционное поле точек и границы доверительного интервала.

Анализ парных ранговых связей

Под ранговой корреляцией понимается статистическая связь между ординальными (порядковыми) переменными. Методы ранговой корреляции широко используются при проведении различных экспертных исследований. Основной задачей ранговой корреляции является ответ на вопрос: есть ли связь между упорядочением (ранжировкой) анализируемых объектов по свойству X1 и упорядочением тех же объектов по свойству X2? Степень взаимосвязи оценивают, как правило, по коэффициентам Спирмена или Кендала.

Пример 3. Десять предприятий были проранжированы вначале по степени прогрессивности их оргструктур (форма собственности и т.п.), а затем – по эффективности их функционирования. В результате были получены следующие ряды рангов: X1 = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) и X2 = (2, 3, 1, 4, 6, 5, 9, 7, 8, 10).

Создать файл с данными, затем выбрать *Статистика / Непараметрические данные*, далее *Correlations (Spearman ...),* задать переменные, для которых необходимо оценить тесноту взаимосвязи и во вкладке *Advanced* определить требуемый коэффициент. В рассматриваемом примере корреляция по Спирмену дает коэффициент = 0,915 с высоким уровнем статистической значимости.

Анализ множественных ранговых связей

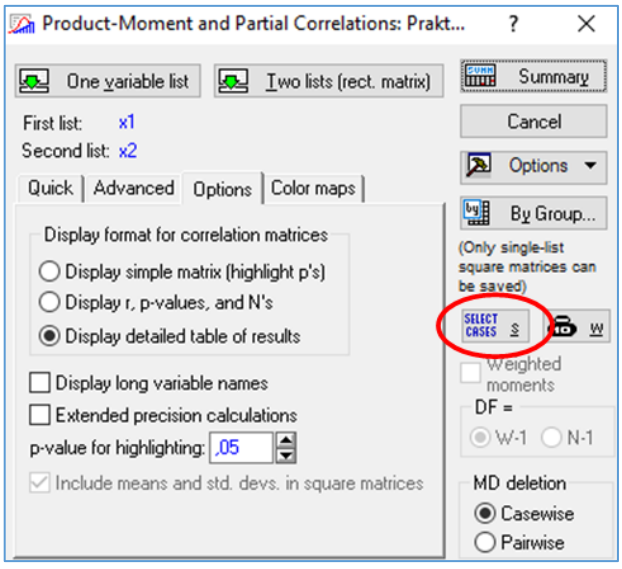
Для оценки тесноты связи между несколькими переменными (>2) используется коэффициент конкордации. Для расчета коэффициента конкордации необходимо выбрать *Статистика / Непараметрические данные / Comparing multiple dep.samples (variables).* Далее, необходимо выбрать переменные (все) и нажать *Summary: Friedman ANOVA & Kendall's concordance*. Значение коэффициента конкордации будет представлено в верхней части итоговой таблицы.

Пример 4. В качестве примера рассчитайте коэффициент конкордации из второго задания.

Задание для самостоятельного выполнения

Имеется база по лабораторным исследованиям крови для 3500 пациентов с различными диагнозами. База содержит следующие сведения: пол, возраст, вес, ИМТ, диагноз, липидный спектр крови, индекс атерогенности, глюкозу.

Необходимо произвести корреляционный анализ данных для выделенных групп больных.

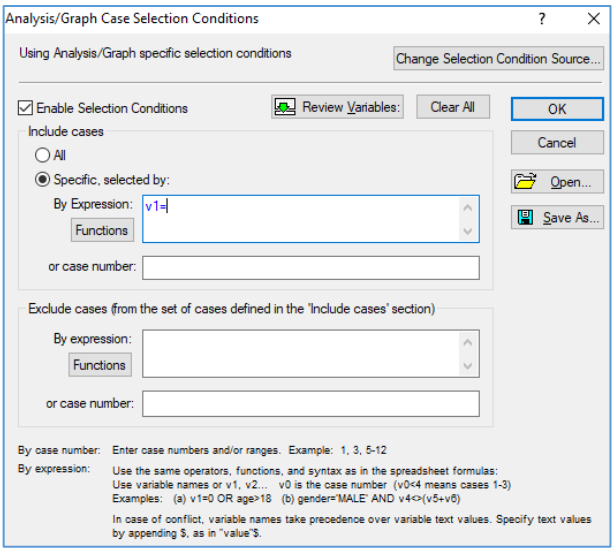


Исходя из величины коэффициента корреляции, сделать выводы о силе зависимости между переменными и ее направлении. Сделать выводы о значимости корреляции. Для наиболее значимых зависимостей построить графики рассеяния.

Для того, чтобы включать в анализ не всю выборку, а определенную ее часть, можно использовать кнопку **Select Cases**:

В появившемся окне можно задать необходимые условия для отбора случаев, для построения сложных условий можно использовать логические «AND» и «OR».

Замечание: перед проведением анализа определить вид распределения и использовать соответствующий метод корреляционного анализа.



Задания по вариантам:

Вариант	*Условие формирования выборки*
1.	Артериальная гипертензия (АГ)
2.	Мужчины, Ишемическая болезнь сердца (ИБС)
3.	Женщины, Ишемическая болезнь сердца (ИБС)
4.	Мужчины, Гипертоническая болезнь (ГБ)
5.	Женщины, Гипертоническая болезнь (ГБ)
6.	Обследование или Практически здоров (Пр.здоров)
7.	Бронхиальная астма (Бр.астма) или хронический бронхит (Хр.бронхит)
8.	Нейроциркуляторная дистония (НЦД), вегетососудистая дистония (ВСД)
9.	Язвенная болезнь (ЯБ), Гастрит, Панкреатит
10.	Мужчины, Повышенный вес (ИМТ>25)
11.	Женщины, Повышенный вес (ИМТ>25)
12.	Пожилой возраст (>60 лет), Ишемическая болезнь сердца (ИБС)

Вопросы для самоконтроля

1. Какие цели преследуются при изучении зависимости между переменными?
2. Какие виды связей между переменными вы знаете?
3. Что означает функциональная зависимость?
4. Что означает корреляционная связь? Приведите примеры.
5. Что означает коэффициент корреляции Пирсона? Спирмена?
6. Приведите примеры графиков зависимостей между переменными с разными коэффициентами корреляции
7. Что означает уровень значимости корреляции?

СПИСОК ИСТОЧНИКОВ

1. Гараничева, С. Л. Основы статистики : учеб.-метод. пособие / С. Л. Гараничева, В. А. Таллер, Е. Г. Машеро. – Витебск, ВГМУ, 2019. – 163 с.

2. Гельман В.Я., Тихомирова А.А. Статистический анализ медико-биологических данных в MS Excel : Учебно-методическое пособие. – СПб.: Санкт-Петербургский государственный педиатрический медицинский университет, 2016. – 54 с.

3. Жижин, К.С. Медицинская статистика: учебное пособие / К.С. Жижин – Ростов н/Д: Феникс, 2007, – 160 с.

4. Макарова, Н.В. Статистический анализ медико-биологических данных с использованием пакетов статистических программ Statistica, SPSS, NCSS, SYSTAT : методическое пособие / Н.В. Макарова ; Всерос. центр экстрен. и радиац. медицины им. А.М. Никифорова МЧС России – СПб.: Политехника-сервис, 2012. – 178 с.

5. Медик, В.А Математическая статистика в медицине: учеб пособие / В.А. Медик, М.С. Токмачев. – М.: Финансы и статистика, 2007. – 800 с.

6. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. – М. : ГЭОТАР-Медиа, 2016. – 528 с.

7. Трухачёва Н.В. Математическая статистика в медико-биологических исследованиях с применением пакета Statistica. – М.: ГЭОТАР-Медиа, 2013. – 384 с.

8. Шеламова, М. А. Статистический анализ медико-биологических данных с использованием программы Excel : учеб.-метод. пособие / М. А. Шеламова, Н. И. Инсарова, В. Г. Лещенко. – Минск : БГМУ, 2010. – 96 с.

Учебное издание

АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ

Методические рекомендации

Составители:
БУЛГАКОВА Наталья Валентиновна
ЧИРКИНА Анна Александровна

Технический редактор *Г.В. Разбоева*
Компьютерный дизайн *Л.В. Рудницкая*

Подписано в печать 06.06.2024. Формат 60x84 $^{1}/_{16}$. Бумага офсетная.
Усл. печ. л. 2,27. Уч.-изд. л. 1,84. Тираж 9 экз. Заказ 81.

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий
№ 1/255 от 31.03.2014.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».
210038, г. Витебск, Московский проспект, 33.